CHAPTER 1

INTRODUCTION

1.1 Urinary system

The urinary system is responsible for keeping the body balanced by eliminating waste and excess fluids. It consists of the kidneys, ureters, bladder, and urethra, which all work together to produce and remove urine. The urinary system also referred to as the renal system or urinary tract coordinates a wide range of important processes that are necessary to preserve the body's delicate internal balance. This complex system is essential for maintaining stable blood pressure, controlling fluid and electrolyte balance, promoting the production of red blood cells, and supporting bone health in addition to its well-known function in the removal of metabolic wastes. The system comprises four principal components

1.1.1 Kidneys

These are paired organs responsible for filtering the blood. Each kidney contains millions of nephrons the functional units that remove toxins, excess water, and waste metabolites, thereby producing urine.

1.1.2 Ureters

These slender tubes conduct the urine from the kidneys to the bladder, ensuring that waste products are efficiently transported.

Fig 1.1 Urinary System

1.1.3 Bladder

A muscular sac, the bladder stores urine until it reaches a volume that triggers the urge to void. Its elasticity and muscular contractions are critical for effective urine expulsion.

1.1.4 Urethra

This is the final conduit through which urine exits the body. In both men and women, its structure and length vary, influencing the risk of urinary infections.

This intricate system is not only essential for waste elimination but also for regulating blood pressure, acid-base balance, and overall metabolic stability. The integration of renal, endocrine, and nervous system functions underscores the complexity of urinary physiology.

1.2 Urology

The field of medicine known as urology treats conditions affecting the kidneys, ureters, bladder, and urethra in both men and women. It also addresses the male reproductive organs (prostate, scrotum, testes, penis, etc.). Urologic health is crucial because everyone can have health issues in these sections of their body. One surgical specialization is urology. In addition to performing surgeries, urologists are medical professionals with expertise in internal medicine, pediatrics, gynecology, and other fields. This is due to the fact that urologists deal with a variety of clinical issues. UTIs, Nephrotic syndrome, and kidney stones are among the serious conditions that can impact the urinary system. Among these is urinary incontinence, which is defined as the uncontrollably occurring flow of pee. In order to diagnose urinary tract infections and choose the best antibiotic treatment based on the organism's susceptibility, a urine culture is conducted to detect the presence of bacteria or other microorganisms in the urine. Female Urology - Focuses on pelvic floor disorders, urinary incontinence, and recurrent urinary tract infections in women. Improving the urological health of women is a major goal of female urology, particularly for older women and those with complicated gynecological or obstetric histories. Gynecology, physiotherapy, and behavioral therapy are just a few of the interdisciplinary treatments that are frequently incorporated into its emphasis on tailored care. Male Infertility and Andrology - This field focuses on erectile dysfunction, fertility problems, and male reproductive health.

Urologic conditions are very common and can have a major effect on one's quality of life. Millions of people worldwide suffer from ailments like kidney stones, incontinence, urinary tract infections (UTIs), and cancers. Urological research and public health initiatives are heavily focused on urinary tract infections, which are among the most prevalent infectious disorders, especially in women.

1.3 Urinary Tract Infection

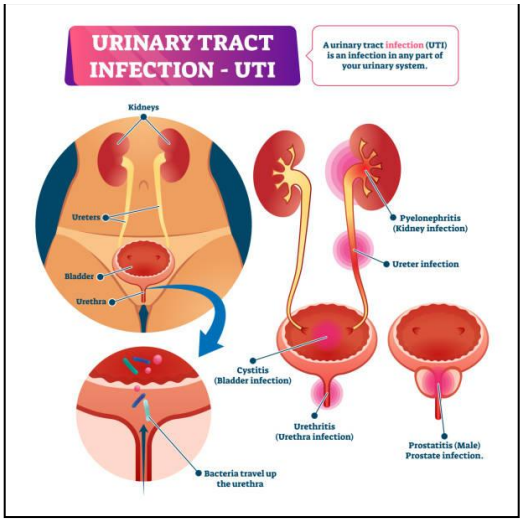


Fig 1.2 Urinary Tract Infection

Urinary tract infections (UTIs) can affect any part of the urinary system, although most typically affect the lower urinary tract, which includes the bladder and urethra. UTIs are roughly four times more widespread in women compared to men. Vulnerability to these infections may be increased by certain diseases, such as diabetes, spinal cord injury, or urinary catheter use. Since UTIs can be successfully identified by a urine test and treated with antibiotics, early medical intervention is crucial. Although bladder infections, also referred to as cystitis, are the most frequently observed kind of UTI, they can occur in several sections of the urinary system. More severe consequences may arise from bladder infections that spread to the kidneys or upper urinary system if treatment is not received. Urine is first produced in the kidneys and subsequently sent to the bladder through the ureters. Urine is held in the bladder until the urethra allows it to be released. The urethral aperture is found at the tip of the penis in males and in front of the vagina in females. Preventing infections and other conditions requires maintaining the health of the urinary tract. UTIs and urinary incontinence (UI) are common urinary tract disorders that have a major negative influence on women's health. For men, women, and children, the National Kidney and Urologic Diseases Information Clearinghouse offer extensive information on urinary tract health. Antibiotics are frequently used by medical professionals to treat urinary tract infections. Additionally, there are things you may do to reduce the likelihood of developing a UTI in the first place. A frequently occurring and reoccurring infection, urinary tract infections (UTIs) can be minor or potentially fatal. Bacteria entering the bladder and causing infection is the main cause of UTIs. Nearly 50% of women will get at least one UTI in their lives, making them especially susceptible to these illnesses. In addition to being uncomfortable, untreated UTIs can be harmful to one's health. The growing problem of antibiotic resistance highlights the necessity for further research into more efficient treatment strategies and the preservation of beneficial bacteria, even though antibiotics are still the standard of care.

1.4 The Formation of Urinary Tract Infection

Urinary tract infections (UTIs), the most common bacterial infectious disease in the general population, are associated with a high rate of morbidity and financial expenses. Urine disposal from the body and waste removal from the bloodstream are the main tasks of the urinary tract. Urinary tract infections (UTIs) are caused by the invasion and growth of harmful microorganisms, most frequently bacteria, within the urinary tract. The kidneys, ureters, bladder, and urethra are all parts of the urinary system. Although the majority of UTIs affect the lower tract, which includes the bladder and urethra, the infection can affect any region of this system.

The urinary system is the most prevalent site of hospital infection, accounting for nearly 40% of nosocomial infections (an estimated 600,000 patients yearly), according to acute care hospitals. Urinary tract infections (UTIs) are among the most common medical diseases that physicians treat. Urinary tract infections (UTIs) are among the most common bacterial illnesses in the general population that are being identified in hospitals. In those without anatomical or functional issues, UTIs typically resolve on their own, despite their propensity to recur.

In addition to being transmitted from person to person, uropathogens can also be transmitted by food or water. Their distinctive characteristics, such as the production of adhesins, siderophores, and toxins, enable them to colonize and infiltrate the urinary system. Although they are typically self-limiting, antibiotic-treated UTIs resolve symptoms more quickly and are more likely to completely cure bacteriuria. They have a detrimental effect on the gut and vaginal microbiota, nevertheless, and also favor commensal bacteria and resistant uropathogens.

Given that uropathogens are becoming increasingly resistant to the available antibiotics, now may be a good moment to investigate alternative treatments for UTIs. In urology, urolithiasis, UTIs, and benign prostatic hyperplasia (BPH) are three of the most common non-cancerous conditions. However, there is still a shortage of comprehensive and up-to-date epidemiological data. The microbiological etiology of urinary tract infections is believed to be well-established and rather consistent. An elevated risk of urinary tract infections (UTIs) and other infectious diseases is associated with excess body weight. The management of neurological lower urinary tract dysfunction has been well documented, however it is typically thought of in connection with common conditions such as spinal cord injury or multiple sclerosis.

1.5 Symptoms

UTIs do not always present noticeable symptoms. However, when symptoms do occur, they may include

- A strong, persistent urge to urinate.
- A burning sensation during urination.
- Frequent urination, often passing small amounts of urine and Cloudy urine.
- Urine that appears red, bright pink, or cola-colored, indicating the presence of blood.
- Strong-smelling urine, Pelvic pain, especially in women, centered on the pelvis and pubic bone.

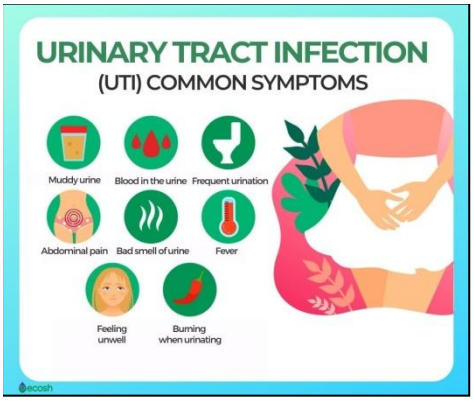


Fig 1.3 Symptoms

In older adults, UTIs may be overlooked or mistaken for other conditions.

1.5.1 Types of Urinary Tract Infections and Associated Symptoms

The UTI may present distinct symptoms, depending on the part of the urinary tract affected

Kidneys (Acute Pyelonephritis)

Symptoms of Urinary Tract Infection in Kidney are Back or side pain, High fever, Shaking and chills, Nausea and Vomiting.

Bladder (Cystitis)

Symptoms of Urinary Tract Infection in Bladder are Pelvic pressure, Lower abdominal discomfort, Frequent, painful urination and Blood in urine.

Urethra (Urethritis)

Symptoms of Urinary Tract Infection in Urethra are Burning sensation during urination and Discharge. Proper identification of the type of UTI is crucial for targeted treatment and effective management.

1.6 Current Situation of Urinary Tract Infections

Urinary tract infections (UTIs) are among the most serious bacterial disorders, with an estimated 150 million cases occurring annually worldwide. In the United States, UTIs are thought to be the cause of 10.5 million office visits, 3 million emergency department visits, and 400,000 hospitalizations annually, with an estimated cost of more than $4.8 billion. UTIs are more common in women, then in older males and male infants. The lifetime risk of UTIs for women is far higher than 60%, according to epidemiologic data from a study done in the United States. Approximately one in three women will experience at least one UTI that was diagnosed by a doctor and will necessitate the use of antibiotics. Persistent UTIs are common, usually occurring more than once.

UTI recurrences can be defined as the second or third UTI within six months of an index UTI, the second or third UTI within twelve months of an index UTI, or the second or third UTI within six or twelve months of an index UTI. One study found that 24% of college women who experienced their first UTI experienced another one within six months, while another found that women who experienced two or more UTIs were 2–5 times more likely to experience another one within a year than those who had only experienced one. A survey of self-reported UTIs revealed that women experience the majority of UTIs. Women who had at least two UTI episodes were more likely to have one, according to a nationwide US study of self-reported UTIs. Nevertheless, despite the high incidence of UTIs and recurrences and the substantial burden they place on patients and the healthcare system, there is a dearth of current data on the epidemiology of rUTIs in the US.

While several earlier studies examined risk factors for rUTIs, most of them only evaluated a small sample size and many of them focused on specific populations, such as children, young women, and the elderly. Furthermore, the COVID-19 pandemic has likely worsened the trend of increasing outpatient UTI incidence, particularly in virtual care settings.

1.7 Complication of Urinary Tract Infections

Clinically, UTIs can be classed as difficult or uncomplicated. Uncomplicated UTIs are classified into two types: upper (pyelonephritis) and lower (cystitis). uUTIs typically affect healthy individuals. Klebsiella pneumoniae and other pathogens are the next leading cause of uUTIs, accounting for around 75% of infections. Indwelling catheterization, immune suppression, renal failure/transplantation, and urinary obstruction are all risk factors for complex UTIs that impair the urinary tract or host defense. In the United States, indwelling catheters cause 1 million complex UTIs each year, accounting for 70-80% of all cases.

Several kinds of viral illnesses affect the urinary tract, from the urethra to the kidneys. According to studies, 50-60% of women will encounter at least one UTI in their lifetime, making it one of the most common infections. Enterobacterales, a category of bacteria typically found in the gastrointestinal (GI) tract, include Escherichia coli, Klebsiella pneumoniae, and Proteus mirabilis, which cause this disorder. Although it is relatively uncommon, bloodstream bacteria that travel to the kidneys or bladder can cause UTIs. Female sex, recent sexual intercourse, diabetes, and structural or functional urological problems all increase the risk of urinary tract infections. The diagnosis and therapy are determined by the severity of the disease as well as the patient's features.

One type of UTI is acute uncomplicated cystitis, sometimes known as "simple cystitis." It is a bladder infection that arises in an immune-competent host with normal urinary tract anatomy. The characteristic symptoms include suprapubic discomfort, frequency, urgency, or dysuria without upper urinary tract involvement (e.g., flank pain or costovertebral angle tenderness) or systemic disease (e.g., fever, rigors, or vomiting). The term "complicated UTI" has historically been used to refer to UTIs that do not meet the previously specified criteria for uncomplicated cystitis, as well as UTIs that affect people with severe immune suppression or significant anatomical abnormalities. There is no single treatment for complicated UTIs because they encompass a wide spectrum of conditions. Instead of utilizing the binary of simple versus complicated UTIs, consider the different syndromes that fall under the umbrella of "complicated" UTIs, such as pyelonephritis, prostatitis, or catheter-associated UTIs.

1.8 Types of Uropathogens

1.8.1 Escherichia coli (E. coli)

E. coli is a gram-negative bacillus that is the most common cause of uncomplicated urinary tract infections (uUTIs), as well as bacteremia, pneumonia, and stomach infections. Early detection and successful treatment are critical given the considerable burden E. coli infections place on healthcare systems

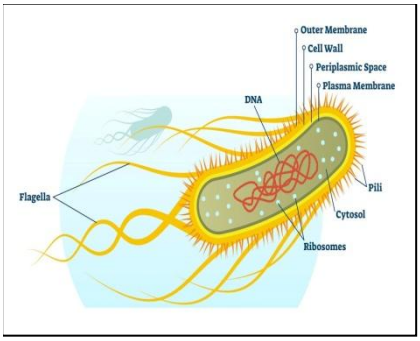


Fig 1.4 E.coli

1.8.2 Enterobacterales

Enterobacterales is a diverse order of approximately 250 bacterial species, which includes some of the most prevalent human pathogens that cause bloodstream infections, gastroenteritis, and urinary tract infections. Taxonomic modifications, such as the reclassification of Enterobacter aerogenes into the Klebsiella genus, have made reporting and treatment more complicated.

1.8.3 Klebsiella pneumoniae

Klebsiella pneumoniae, first discovered by Carl Friedlander in 1882, is a gram-negative, encapsulated, non-motile bacterium that can be found in the environment and the human gastrointestinal tract. It is a leading cause of hospital-acquired pneumonia and is frequently associated with diabetes and alcohol use disorders. In the United States, it accounts for 3–8% of all nosocomial bacterial infections.

1.8.4 Proteus mirabilis

Proteus mirabilis is a gram-negative facultative anaerobe distinguished by its swarming motility, which allows it to adhere to surfaces such as catheters and other medical devices. This trait contributes to its significance in complex UTIs. To make an accurate diagnosis, the patient's symptoms must be considered in addition to the culture results.

1.9 Current Situation of UTI

1.9.1 UTI in India

Urinary tract infections (UTIs) are among the most frequent disorders affecting people of all ages, from newborns to the elderly. Every year, roughly 150 million people worldwide are diagnosed with urinary tract infections. Women in India are substantially more likely to have UTIs, with a lifetime risk of 60% compared to 13% for men. Women's heightened sensitivity is linked to variables such as a shorter urethra, the absence of prostatic secretions, pregnancy, and a higher chance of fecal contamination.

Clinically, UTIs are classified as simple or complex. Uncomplicated UTIs usually occur in healthy people with no neurological or anatomical abnormalities of the urinary system and manifest as cystitis or pyelonephritis.. In contrast, complicated UTIs occur as a result of factors that compromise the urinary tract, such as urinary obstruction, neurological disorders causing urine retention, renal failure, renal transplantation, pregnancy, and the presence of foreign bodies such as calculi, indwelling catheters, or other drainage devices.

The practical treatment of UTIs without urine culture or susceptibility testing has become common. However, this method might lead to antibiotic abuse and resistance. Antibiotic susceptibility patterns for UTI-causing bacteria vary by area and across time. To provide effective treatment techniques, UTI pathogens and resistance trends must be monitored on a regular basis.

1.9.2 UTI in Tamil Nadu

Women in Tamil Nadu between the ages of 15 and 44 are most sensitive to UTIs. UTIs are one of the most prevalent bacterial illnesses treated in primary care. Left untreated, UTIs can pose serious public health risks, resulting in consequences that could have been avoided with quick diagnosis and care. Early detection and treatment of UTIs are critical for reducing patient suffering, shortening hospital stays, and lowering economic costs. Continuous surveillance and tailored antibiotic medication are critical for reducing the spread of resistant strains and improving patient outcomes.

1.10 Reproductive Tract Infections (RTIs)

Females of reproductive age are especially vulnerable to reproductive tract infections (RTIs), which are prevalent in developing countries. RTIs can cause serious health consequences including as infertility, chronic pelvic pain, and poor pregnancy outcomes. The high prevalence of RTIs in rural girls is mostly due to stigma, a lack of awareness, poor hygiene practices, and restricted access to healthcare services. RTIs are a group of infections that can affect different parts of the female and male reproductive systems, including the uterus, fallopian tubes, ovaries, cervix, vagina, and testes. Bacteria, viruses, fungi, or parasites can cause these infections, which can result in a wide range of symptoms and health problems. Common bacteria responsible for RTIs include Chlamydia trachomatis and Neisseria gonorrhoeae.

Women may hesitate to seek medical assistance because of social stigma and misconceptions about reproductive health. Furthermore, a lack of information regarding RTI prevention and management exacerbates the situation. Early identification and treatment are critical in avoiding long-term consequences.

To address these concerns, health education campaigns should be launched through a variety of venues. Community-based programs, awareness campaigns, and improved healthcare infrastructure can all help reduce the frequency of RTIs and enhance women's reproductive health. Regular monitoring, accessible healthcare services, and promoting personal hygiene can all assist to reduce the development of these illnesses.

Sexually transmitted infections (STIs) and other reproductive tract infections (RTIs) contribute significantly to the burden of disease and negatively affect people's reproductive health. Around the world, they inflict pain on both men and women, but the effects on women are significantly more severe and pervasive than on males. There is a dearth of precise statistics on the prevalence of STIs in India, particularly in the general population. In India, the disease is thought to impact 6% of the population, potentially affecting 30 million of the 340 million people worldwide. Additionally, according to estimates, roughly 40% of women experience RTIs or STIs at any given time, but only 1% of them finish treating both partners.

1.11 Urinary Tract Infections during Pregnancy

Urinary tract infections (UTIs) are among the most common infections encountered by pregnant women, accounting for up to 10% of pregnancies. In fact, UTIs are considered the second most common disease during pregnancy, after anemia. If left untreated, these infections endanger both maternal and fetal health.

Fig 1.5 Pregnancy

Urinary tract infections (UTIs) during pregnancy are a frequent but potentially significant condition that can endanger both maternal and fetal health. Pregnancy causes anatomical and physiological changes that increase susceptibility to UTIs, making it critical to understand the specific dynamics of these infections in this group. Untreated UTIs can cause consequences like as pyelonephritis, preterm birth, low birth weight, and even maternal sepsis, emphasizing the significance of early detection and treatment.

Symptoms such as dysuria, frequency, urgency, and suprapubic pain in pregnant women may also be caused by physiological changes associated with pregnancy, such as increased urine frequency due to uterine strain on the bladder.

1.11.1 Types of UTIs in Pregnancy

There are three distinct clinical forms of pregnancy-related UTIs

1. **Asymptomatic Bacteriuria (ASB):** This condition involves the presence of bacteria in the urinary tract without noticeable symptoms. Group B streptococci are commonly detected in ASB.
2. **Cystitis:** An infection of the bladder, often accompanied by symptoms such as frequent urination, urgency, and discomfort.
3. **Pyelonephritis:** A severe kidney infection that can lead to high fever, back pain, and systemic complications.

1.11.2 Risks and Complications

All forms of UTIs during pregnancy require immediate attention and treatment to prevent severe consequences. Untreated infections may lead to

- Preterm labor and low birth weight
- Maternal sepsis
- Increased risk of perinatal mortality
- Pyelonephritis, which can cause significant morbidity in mothers

Pregnant women with low socioeconomic status and a history of prior UTIs are at higher risk of developing these infections. The most frequent causative agent of both symptomatic and asymptomatic infections is Escherichia coli.

1.11.3 Diagnosis and Management

The gold standard for identifying UTIs during pregnancy is quantitative urine culture. Screening for asymptomatic bacteriuria has become common practice in obstetrics, as therapy has been demonstrated to minimize the risk of pyelonephritis during pregnancy. Antibiotic therapy is critical, however local resistance patterns must be examined in light of developing antibiotic resistance.

1.11.4 Impact on Newborns

According studies, babies of moms with untreated UTIs are more likely to be underweight and shorter in length at birth. Furthermore, there is a significant link between UTIs and manner of delivery, with more cesarean sections found in infected moms.

Urinary tract infections during pregnancy must be detected early, treated properly, and closely monitored to avoid problems. Systematic screening, timely intervention, and tailored treatment strategies are critical for protecting maternal and fetal health. Improved public health initiatives, increased awareness, and better access to prenatal care can significantly lower the prevalence of UTIs in pregnant women.

1.12 Economic Impact of Urinary Tract Infections (UTIs)

Urinary tract infections (UTIs) are among the most common bacterial illnesses, dramatically increasing healthcare costs and patient morbidity. The economic burden of UTIs goes beyond direct medical costs to include lost productivity, long-term health issues, and societal consequences.

The past studies show how retrospective study will emphasize the importance of nosocomial UTI surveillance for the development of effective therapeutic strategies appropriate to the local flora, as well as raise awareness and highlight the inherent economic impact and the need for infection prevention strategies.

1.12.1 Incidence and Prevalence

According to the 1997 National Ambulatory Medical Care Survey and National Hospital Ambulatory Medical Care Survey, UTIs resulted in around 7 million office visits, 1 million emergency department visits, and 100,000 hospitalizations in the United States per year. However, actual incidence rates are difficult to calculate due to a lack of mandated reporting and varied diagnostic techniques.

1.12.2 High-Risk Populations

Women are disproportionately affected by UTIs, with nearly 50% experiencing at least one UTI in their lifetime. By age 24, one in three women requires antibiotic treatment for a UTI. Other high-risk populations include:

- Infants and pregnant women
- The elderly
- Patients with diabetes, multiple sclerosis, or spinal cord injuries
- Individuals with indwelling catheters
- Patients with immunodeficiency disorders or underlying urologic abnormalities

1.12.3 Economic Burden

The financial implications of UTIs are substantial. Community-acquired UTIs alone are projected to cost $1.6 billion annually. The costs arise from Medical consultations and diagnostic tests, Antibiotic treatments, Hospitalizations and emergency care for severe cases and lost productivity due to illness and recovery time.

1.12.4 Health Consequences and Associated Costs

In non-pregnant adults, acute uncomplicated UTIs are generally benign. However, in specific populations, UTIs can result in severe health complications

- **Pediatric patients:** Increased risk of end-stage renal disease and reduced renal function.
- **Pregnant women:** Higher risks of fetal death, preterm birth, and pyelonephritis.
- **Elderly populations:** UTIs are the second most common infection, representing 25% of all infections in non institutionalized elderly individuals.

UTIs are a major economic and public health concern, especially among high-risk populations. The financial cost emphasizes the importance of effective prevention efforts, early diagnosis, and proper treatment to reduce both medical and economic effects.

Urinary tract infections (UTIs) can cause major health problems if left untreated, even though they are usually treatable. Urinary tract infections are one of the most common causes of hospitalization. This constellation of illnesses poses a significant cost to the global healthcare system due to their prevalence in community-based settings as well as those that arise in hospitals.

Recurrent infections, which are characterized by several episodes in a brief period of time, are prevalent, particularly in women. Bladder infections have the potential to spread to the kidneys, resulting in irreversible harm and diminished renal function if left untreated. UTIs during pregnancy can increase the risk of low birth weight and premature delivery, which is dangerous for both the mother and the unborn child.

1.13 Diagnosis and Treatment of Urinary Tract Infections (UTIs)

1.13.1 Diagnostic Procedures

Accurate diagnosis of UTIs is essential for effective treatment. Several tests and procedures are commonly used

- **Urine Analysis:** A urine sample is examined in a lab to check for white blood cells, red blood cells, or bacteria. To prevent contamination, patients are typically instructed to clean the genital area with an antiseptic pad and collect the midstream urine.
- **Urine Culture:** In cases where further analysis is needed, a urine culture is conducted to identify the bacteria responsible for the infection and determine the most effective antibiotics.
- **Imaging Studies:** Recurrent UTIs may indicate structural abnormalities in the urinary tract. Ultrasounds, CT scans, or MRIs, sometimes using contrast dye, help visualize these structures.
- **Cystoscopy:** For persistent UTIs, a cystoscopy may be performed. This procedure involves inserting a thin tube with a camera through the urethra into the bladder to inspect the urinary tract.

1.13.2 Treatment Options

Antibiotics are the primary treatment for UTIs, with the choice of medication and duration depending on the infection's severity and the patient's health.

Treatment for Simple Infections (not recommended without professional advice)

Common antibiotics for uncomplicated UTIs include

- Trimethoprim and sulfamethoxazole (Bactrim, Bactrim DS)
- Fosfomycin (Monurol)
- Nitrofurantoin (Macrodantin, Macrobid, Furadantin)
- Cephalexin, Ceftriaxone

Fluoroquinolones, such as ciprofloxacin and levofloxacin, are typically reserved for more serious cases due to severe negative effects. Symptoms usually resolve after a few days, but antibiotics should be used for the recommended length to guarantee total eradication.

1.13.3 Treatment for Frequent Infections

For individuals with recurrent UTIs, additional interventions may be recommended

- Long-term, low-dose antibiotics taken for six months or more.
- Self-diagnosis and treatment in consultation with a healthcare provider.
- A single dose of antibiotics post-intercourse if UTIs are linked to sexual activity.
- Vaginal estrogen therapy for postmenopausal women.

1.13.4 Treatment for Severe Infections

Severe UTIs, particularly those involving the kidneys, may require hospitalization and intravenous (IV) antibiotics.

Lifestyle and Home Remedies

In addition to medical treatment, several lifestyle changes and home remedies can help manage UTI symptoms:

- **Hydration:** Drinking plenty of water dilutes urine and helps flush out bacteria.
- **Avoid Irritants:** Coffee, alcohol, and caffeinated soft drinks should be avoided, as they can irritate the bladder.
- **Heat Therapy:** Applying a warm heating pad to the abdomen can relieve discomfort.

1.13.5 Alternative Therapies

Cranberry products, such as juice or pills, have been demonstrated to reduce bacterial adherence to the urinary system, which may help prevent UTIs. While research is ongoing, cranberry products are generally safe for most people, while they may induce stomach distress or interact with blood thinners such as warfarin.

1.14 Treatment and Antibiotic Resistance

In England, urinary tract infections (UTIs) are one of the most common causes for women to seek primary care, with nearly half of all women experiencing one at some point in their lives. Treatment typically consists of empirical antibiotic therapy without urine culture, with first-line alternatives including nitrofurantoin, trimethoprim, pivmecillinam, and fosfomycin. Antibiotic resistance, on the other hand, is a major public health concern that requires careful consideration of local resistance patterns in order to assure both effective and cost-effective treatment. Trimethoprim is the most cost-effective alternative when resistance rates are less than 30%, but as resistance increases to 50%, a single 3 g dose of fosfomycin becomes more inexpensive and effective. Despite standards, medication practices vary greatly, with nitrofurantoin use increasing because to its efficacy against resistant strains. Optimizing antibiotic selection necessitates balancing therapeutic efficacy, cost-effectiveness, and the need to prevent the spread of multidrug-resistant infections through informed prescribing.

1.15 Complications

When diagnosed and treated promptly, lower urinary tract infections rarely result in serious complications. However, untreated UTIs can lead to severe health concerns, including:

- **Recurrent Infections:** Experiencing two or more UTIs within six months or three or more within a year is classified as recurrent infection, with women being particularly susceptible.
- **Kidney Damage:** Untreated bladder infections can ascend to the kidneys, potentially causing permanent damage.
- **Pregnancy Complications:** UTIs during pregnancy can result in low birth weight or preterm delivery.
- **Urethral Stricture in Men:** Repeated urethral infections may lead to urethral narrowing, causing urinary difficulties.
- **Sepsis:** In severe cases, infection can spread into the bloodstream, resulting in sepsis, a life-threatening medical emergency.

1.16 Overview of Urinary Tract Infection

UTI is one the most common infection that can cause severe Physical, Mental, economic and social stress to a person.

1.16.1 Physical stress

The physical pain and burning sensation will happen every time we go to do the natures call. After a period of time we can't handle the pain which at least 8 time in a day. You will always feel pressure in your lower belly. With Back or side pain you can't constantly do work properly. Blood in urine will increase blood loss this will the body very weak and will cause high fever.

1.16.2 Mental stress

If we are in schools, college, office, meetings, traveling, social engagements and in festival time we can't properly study or works and we can't spend good time with our family and friends. If you always feel pressure in your lower belly your daily mental stress will increase.

1.16.3 Economical stress

For uncomplicated UTI - The average national ER visit cost for the treatment of a UTI is $2,215 for patients with insurance, and $2,474 for patients without insurance. However, some hospitals are publishing prices up to 10 to 20 times higher, according to Goodbill's price analysis of more than 2,500 hospitals across the country.

For complicated UTI - In India, the cost of treating a complicated urinary tract infection (UTI) can range from ₹2,000 to ₹10,000 or more depending on the severity of the infection, the required tests, the type of antibiotics needed (which may be more expensive for severe cases), and the healthcare facility you choose, with potential additional costs for hospitalization if necessary. The mean cost per case was €5700, with considerable variation between countries (largest value €7740 in Turkey; lowest value €4028 in Israel), mainly due to differences in length of hospital stay. Factors associated with higher costs per patient were: type of admission, infection source, infection severity, the Charlson comorbidity index and presence of MDR.the currency conversion rate for 1 EUR stands at INR 91.3321.

According to recent studies, the average monthly income for a lower middle-class individual in India is around Rs 33,000, which translates to an annual income of roughly Rs 3.96 lakh.

Key points about the lower middle class in India

- Average monthly income: Rs 33,000
- Average monthly expenses: Around Rs 19,000
- Cities studied: Delhi - NCR, Mumbai, Kolkata, Chennai, Bengaluru, Hyderabad, Ahmedabad, Pune, Lucknow, Jaipur.

For lower middle class families in India the uncomplicated UTI is very expensive and they can't afford for complicated UTI cases.

1.16.4 Social stress

UTI are sometimes misunderstood as STD's and the awareness of UTI is very less in rural areas in India. Because we can't say that we have infection in the places of reproduction. In urban areas it is very easy to get affected by UTI because association with western toilet. Frequent and painful urination you will seek for toilet very often. While traveling in it very hard to find clean and hygienic toilets.

1.16.5 Awareness

Many women are unaware of the urinary tract infection which highly infected by women. Hormonal fluctuations throughout adolescence increase the colonization of the vagina by nephritogenic bacterial strains, which can spread to the periurethral region and result in UTIs. It is linked to sadness, social isolation, low self-esteem, and a lower quality of life. Adolescent UTI has been linked to a number of variables, including inadequate hydration, infrequent voiding, and poor menstrual and sexual hygiene. A young woman is at risk for UTIs due to inadequate hydration, unclean bathrooms, and poor menstrual and sexual hygiene. We must teach our woman the importance of proper hygiene and hydration in all public spaces, including schools, colleges and working places. Schools and all public places should have basic and clean sanitation facilities.

1.17 Objective

1.17.1 Primary Objective

• Identify and quantify risk factors for Urinary Tract Infections (UTIs) in the study population, including demographic, health-related, lifestyle, behavioral, and reproductive health variables.

• Evaluate machine learning models (Logistic Regression, Random Forest, XGBoost, and Neural Networks) for accurately predicting UTI occurrence.

1.17.2 Secondary Objectives

• To assess the prevalence of UTI in the study population and subgroups (e.g., age ranges, commode using groups).

• Evaluate the relationship between commode preference (Indian vs. Western) and UTI risk, taking into account potential confounding factors like age.

• Evaluate how lifestyle factors such as coffee, tea, soft drink, etc., consumption affect the risk of having UTIs.

• Analyze the impact of bladder behaviors and management measures (e.g., fluid restriction and pad protection) on UTI occurrence.

• Compare the predictive ability of machine learning algorithms for identifying individuals at high risk of UTI.

• Use feature selection approaches (e.g., Chi-square test) to identify key predictor factors for UTI and evaluate their impact on model accuracy.

• Provide evidence-based advice for UTI prevention, early detection, and care based on prediction models and specific risk categories.

CHAPTER 2

REVIEW OF LITERATURE

2.1 Neurogenic Lower Urinary Tract Dysfunction (NLUTD) in Uncommon Neurological Diseases

Welk et al. (2022) conducted a comprehensive review of lower urinary tract dysfunction in patients with uncommon neurological diseases, filling a significant gap in the existing literature, which focuses on more common conditions such as spinal cord injury or multiple sclerosis. The study looked at a variety of neurological problems, including degenerative and traumatic brain disorders, autoimmune diseases, genetic ailments, and peripheral neuropathies. This study made an important addition by investigating the underlying disease processes and their impact on lower urinary tract function, which was supported by disease-specific clinical and urodynamic data. The review highlighted the complexities of addressing neurogenic lower urinary tract dysfunction, emphasizing the importance of tailored care plans that take into account the patient's neurological condition, functional level, prognosis, and personal preferences. The authors emphasized the importance of thorough patient assessment and proper urodynamic investigations in guiding management decisions, as well as raising awareness among specialists about these lesser-known neurological conditions and their urological implications.

2.2 Epidemiological Factors and Risk Assessment in Urinary Tract Infections (UTIs)

Taramian et al. (2024) investigated the relationship between body mass index (BMI) and urinary tract infections in the PERSIAN Guilan Cohort Study, finding a strong link between obesity and an increased risk of UTIs. This study supported broader epidemiological findings that obesity can impair immune function, enhance bacterial colonization, and affect urinary tract structure, potentially increasing vulnerability to UTIs. The study's findings provide vital insights into the public health implications of rising obesity rates, as well as the need of taking BMI into account when assessing and managing UTI risk.

Jadoun (2020) did a comparison study on the prevalence of urinary tract infections among residential girls who used Indian versus Western-style toilets, and found that Western toilet users had a greater incidence of UTIs. Various factors have been recognized as

contributing to this increasing occurrence, including prolonged urine retention caused by avoiding unclean toilets, decreased water intake, and the cleanliness of Western toilet seats. The study stressed the need of educational initiatives targeted at increasing awareness of UTI risk factors and encouraging healthy behaviors.

Moore et al. (2008) investigated the relationship between sexual intercourse and the risk of symptomatic urinary tract infections in postmenopausal women, finding that recent sexual intercourse was strongly associated with incident UTIs in this demographic, implying that sexual activity remains a significant risk factor for UTIs after menopause and emphasizing the need for targeted prevention strategies. The PLUS research consortium (2018) investigated the relationship between occupation and lower urinary tract symptoms in women, using occupational patterns as indirect measures of infrequent voiding behaviors. The findings suggest that certain occupational environments may limit toilet access and contribute to voiding dysfunction, and they call for future research to standardize the assessment of voiding frequency and toilet access across occupations.

2.3 Pathogenesis and Microbial Etiology of UTIs

Foxman (2010) presented a thorough summary of the epidemiology of urinary tract infections, finding them as one of the most prevalent bacterial illnesses acquired both in the community and in hospitals. The study underlined the recurrent nature of UTIs, particularly in those who have no anatomical or functional abnormalities, as well as the specialized properties of uropathogens that aid in colonization and invasion of the bladder. The study looked into the transmission channels of uropathogens, concluding that they spread through person-to-person contact and maybe through contaminated food or water. While antibiotics remain the predominant treatment strategy, the study cautioned against unexpected consequences such as antibiotic-resistant uropathogen selection and alterations to the gut and vaginal microbiome.

Zhu et al. (2021) conducted a global epidemiological analysis of urinary tract infections, urolithiasis, and benign prostatic hyperplasia across 203 countries and territories from 1990 to 2019, revealing that the disease burden for UTIs has increased over the last three decades, while urolithiasis and BPH have decreased.

Mohapatra et al. (2022) investigated the prevalence and antimicrobial resistance patterns of uropathogens in community settings across various regions of India, highlighting the growing incidence of antimicrobial resistance among uropathogens, with a significant percentage of extended-spectrum beta-lactamase producers found in Escherichia coli and Klebsiella pneumonia isolates. Ronald (2002) investigated the microbial etiology of UTIs, concluding that Escherichia coli remains the most common pathogen in acute, community-acquired uncomplicated infections, followed by Staphylococcus saprophyticus, with Klebsiella, Enterobacter, Proteus species, and enterococci being less common causes.

The study emphasized the changing nature of UTI pathogens, particularly in light of developing antibiotic resistance, and examined the various etiologies of severe UTIs, which can involve species less commonly observed in healthy people.

2.4 Clinical Management and Preventative Strategies for UTIs

Bono et al. (2023) concentrated on uncomplicated urinary tract infections, identifying them as one of the most prevalent bacterial infections affecting the lower urinary tract in otherwise healthy people with no anatomical abnormalities or substantial concomitant conditions. The study emphasized the necessity of early identification and management to avoid problems and reduce the strain on healthcare systems. Furthermore, the authors recommended for preventative methods and emphasized the importance of interprofessional collaboration in improving patient care.

2.5 Urinary Tract Infections and Associated Pathogens

Urinary tract infections (UTIs) are among the most prevalent bacterial illnesses, affecting individuals of all ages. While Escherichia coli (E. coli) is the most common causative agent, Enterobacteriaceae, Klebsiella pneumoniae, and Proteus mirabilis also play important roles, especially in hospital-acquired infections and cases involving anatomical anomalies such as vesicoureteral reflux. This literature review looks at new insights about the microbial landscape of UTIs and their clinical implications.

2.5.1 Escherichia coli: The Primary Pathogen

E. coli is the most commonly isolated bacterium in UTIs. Mueller and Tainter (2023) stated that E. coli is the major cause of simple cystitis, as well as other extraintestinal infections such pneumonia and bacteremia. Given the tremendous burden that E. coli infections impose on individuals and healthcare systems, the study underscored the necessity of early detection and treatment.

2.5.2 Entero-bacteriaceae and Vesicoureteral Reflux

DíazÁlvarez et al. (2017) investigated the relationship between Enterobacteriaceae and VUR in newborns. Their findings suggested that the first UTI could reveal underlying urinary tract abnormalities, with non-E.coli Enterobacteriaceae being linked to higher grades of VUR. This shows that pathogen type may be a predictor for urinary tract abnormalities, especially in neonates.

2.5.3 Emergence of Enterobacterhormaechei

Doern (2024) wrote about the growing importance of Enterobacterhormaechei, a member of the Enterobacter cloacae complex (ECC), as a nosocomial pathogen. E. hormaechei is increasingly identified from urine samples, indicating a role in hospital-acquired UTIs. The study linked this species' increased virulence to the existence of a high-pathogenicity island on its chromosome, which is frequently identified.

2.5.4 Klebsiellapneumoniae

Ashurst and Dawson (2023) analyzed the history of Klebsiella pneumoniae and discussed its involvement in hospital-acquired illnesses. The bacterium often colonizes mucosal surfaces and is very virulent and resistant to antibiotics, making it a major concern for healthcare-associated UTIs. The study emphasized the importance of rigorous infection control measures in preventing the development of this opportunistic virus.

2.5.5 Proteus mirabilis: Biofilm Formation and Catheter-Associated Infections

Proteus mirabilis has received attention for its unusual ability to swarm across surfaces and create biofilms, complicating therapy, particularly in catheter-associated UTIs. Jamil et al.

(2023) stated that, while culture remains the gold standard for diagnosing Proteus infections, interpreting culture data necessitates a thorough understanding of the patient's clinical presentation. The reviewed literature highlights the complex microbial landscape of UTIs, with each pathogen having distinct properties that influence diagnosis, therapy, and prognosis. While E. coli remains the most prevalent causal agent, new infections such as Enterobacter hormaechei and Klebsiella pneumoniae require more attention due to their nosocomial origins and drug resistance. Furthermore, the link between Enterobacteriaceae and VUR emphasizes the necessity for specific research in neonatal populations. Understanding the varied roles of these pathogens will help to enhance diagnostic techniques and treatment strategies for better patient outcomes.

2.5.7 Pathogens and Prevalence

Escherichia coli (E. coli) is continuously identified as the most common cause of UTIs, especially in cases of uncomplicated cystitis and newborn UTIs (Mueller & Tainter, 2023). Other members of the Enterobacteriaceae family, such as Enterobacter spp., have emerged as prominent infections, particularly in nosocomial settings (Doern, 2024). Enterobacterhormaechei, a species of the Enterobacter cloacae complex, is commonly isolated from clinical samples and may be more virulent due to genetic variables (Doern, 2024).

Klebsiella pneumoniae and Proteus mirabilis are also important contributors to UTIs, with K. pneumoniae being a leading cause of hospital-acquired pneumonia and P. mirabilis exhibiting unique characteristics such as swarming motility, which allows colonization of medical equipment and complicates treatment (Ashurst & Dawson, 2023; Jamil et al., 2023).

2.6 Risk Factors and Special Populations

Non-E. coli Enterobacteriaceae have been suspected of urinary tract anomalies including vesicoureteral reflux (VUR), especially in neonatal UTIs (Álvarez et al., 2017). Furthermore, pregnancy causes distinct physiological changes that enhance susceptibility to UTIs, with untreated infections risking premature birth and low birth weight (Lee et al., 2019; Habak et al., 2024). Complicated UTIs, defined by anatomical anomalies, immunocompromised states, and the presence of medical devices such as catheters, necessitate specific diagnostic and treatment techniques. These infections have a higher chance of treatment failure and recurrence.

2.7 Sociodemographic and Geographic Insights

Sociodemographic factors also influence UTI prevalence and therapy. Women of reproductive age are vulnerable to UTIs and reproductive tract infections (RTIs), according to studies conducted in rural Tamil Nadu, India. This susceptibility is largely owing to a lack of awareness and healthcare access (Muthulakshmi&Gopalakrishnan, 2017; Balakrishnan et al.,2022). In Bangladesh, the frequency of UTIs in pregnant women was significantly high, with half of the infections being asymptomatic, emphasizing the need for low-cost screening measures.

2.8 Antibiotic Resistance and Treatment Challenges

The extensive use of antibiotics has resulted in the creation of resistant uropathogen strains, necessitating ongoing surveillance and updated treatment guidelines (Pardeshi, 2018). The fluctuating antibiograms of pathogens such as E. coli and Enterobacter spp. hamper empirical therapy, especially in low-resource settings with poor diagnostic skills. The existing body of research emphasizes the multiple character of UTIs, which include various bacteria, risk factors, and geographic differences. Key priorities for future research and clinical practice include improving screening and treatment regimens, particularly for high-risk groups such as pregnant women and those with recurring or severe infections, as well as addressing the growing challenge of antibiotic resistance.

2.9 Urinary Tract Infections (UTIs) During Pregnancy

Urinary tract infections (UTIs) during pregnancy are a major concern due to their frequency and associated problems for both mother and child. The research emphasizes the significance of screening, early detection, and adequate care to avoid negative effects.

2.9.1 Prevalence and Risk Factors

Up to 10% of pregnant women experience UTIs, making them the most frequent type of infection (Szweda & Jóźwik, 2016). Several factors contribute to the increased risk of UTIs in pregnant women, including hormonal changes, mechanical pressure on the urinary tract, and a history of past urinary tract infections (Schnarr & Smaill 2008). Furthermore, low socioeconomic status has been recognized as a risk factor (Schnarr and Smaill, 2008).

2.9.2 Clinical Types and Complications

Three clinical kinds of pregnancy-related UTIs exist: asymptomatic bacteriuria (ASB), cystitis, and pyelonephritis. ASB, defined by the presence of bacteria in the urine without symptoms, requires treatment during pregnancy to avoid progression to symptomatic infection and consequences such as pyelonephritis (Szweda&Jóźwik, 2016). Pyelonephritis can cause considerable maternal and fetal morbidity and mortality if not treated promptly (Millar & Cox, 1997). ASB is characterized by the presence of Group B streptococci in the urinary tract, which may require antibiotic prophylaxis (Szweda&Jóźwik, 2016). Furthermore, untreated UTIs have been linked to poor pregnancy outcomes such as low birth weight and premature birth (Millar & Cox, 1997; Amiri et al., 2015).

2.9.3 Screening and Diagnosis

Because of the hazards, routine screening for asymptomatic bacteriuria is considered standard obstetrical care (Schnarr & Smaill, 2008). Quantitative culture is the gold standard for diagnosis, with Escherichia coli being the most prevalent etiologic agent in both symptomatic and asymptomatic illnesses (Schnarr&Smaill, 2008).

2.9.4 Treatment and Management

Prompt treatment of UTIs during pregnancy is critical to avoiding bad maternal and fetal outcomes. However, there is no clear consensus on the best antibiotics to use or the duration of therapy, and rising antibiotic resistance necessitates taking local resistance rates into account when choosing treatment (Schnarr & Smaill, 2008). In some high-risk groups, systematic screening and close monitoring during pregnancy are advised (Mauroy et al., 1996).

2.9.5 Geographic and Environmental Factors

The prevalence of UTIs in pregnant women varies by area, depending on climate and weather circumstances. For example, a research in Dezful City, Iran, discovered a lower incidence of UTI than other places, possibly due to environmental variables. Furthermore, UTIs in this cohort were substantially associated with higher rates of low-birth-weight infants (Amiri et al., 2015). To reduce hazards to both mother and child, UTIs during pregnancy must be carefully screened, diagnosed early, and treated appropriately.

2.10 Economic Impact and Epidemiology of UTIs

Urinary tract infections (UTIs) are among the most common bacterial infections worldwide, with significant economic and healthcare implications.

2.10.1 Incidence and Morbidity

According to Foxman (2002), UTIs cause around 7 million office visits, 1 million emergency department visits, and 100,000 hospitalizations in the United States each year. Nearly half of all women develop a UTI at some point in their lives. Certain populations, such as pregnant women, the elderly, and catheterized patients, are more vulnerable. Pyelonephritis, preterm delivery, and kidney disease are all consequences of urinary tract infections. The estimated annual cost of community-acquired UTIs in the United States is $1.6 billion.

2.10.2 Financial Burden of UTIs

Ciani et al. (2013) establish the idea of "costs of resignation," which describes how patients with recurrent UTIs suffer repeated infections and incur treatment costs as a result of poor first therapies. Similarly, François et al. (2016) demonstrate that needless treatments and non-recommended diagnostic procedures drive up healthcare expenses in France. The study proposes that cost-cutting methods could be adopted through improved diagnosis and treatment strategies.

Iskandar et al. (2021) look at the economic impact of antibiotic-resistant E. coli UTIs in Lebanon. According to the report, treating resistant illnesses causes longer hospital stays and higher medical expenses, emphasizing the importance of national antibiotic resistance action programs. Callan et al. (2014) evaluate the economic burden of UTIs in Ireland from a healthcare standpoint. The study found that primary care visits are the most expensive, underlining the need for more efficient management measures to avoid unnecessary antibiotic use and hospital stays.

Sulaiman et al. (2024) concentrate on kidney transplant recipients, a particularly susceptible population. UTIs in post-transplant patients drive up healthcare expenses due to extended hospitalization and additional medical procedures. However, the study discovered no direct effect of early UTIs on long-term graft function.

UTIs impose a huge economic cost. Recurrent infections, antibiotic resistance, and high-risk populations all help to drive up healthcare expenditures. To limit the need for unneeded medical procedures, cost-effective strategies should prioritize early diagnosis, adequate antibiotic stewardship, and preventative actions.

2.11 Comprehensive Literature Review on Urinary Tract Infections

Urinary tract infections (UTIs) are one of the most prevalent bacterial illnesses in the world, affecting people of all ages and genders, with women having the highest prevalence. The extant research emphasizes the multidimensional nature of UTIs, which include various bacteria, risk factors, regional inequities, economic burden, and developing concerns like as antibiotic resistance. This review brings together relevant findings from diverse studies to provide a thorough overview of the epidemiology, risk factors, socioeconomic issues, and therapeutic problems related with UTIs.

2.11.1 Pathogens and Microbial Landscape

Escherichia coli (E. coli) is regularly identified as the leading cause of UTIs in many populations (Mueller & Tainter, 2023). Other bacteria, such as Klebsiella pneumoniae and Proteus mirabilis, make significant contributions, especially in hospital-acquired illnesses (Ashurst & Dawson, 2023; Jamil et al., 2023). Enterobacter spp., particularly Enterobacter hormaechei, have received attention in nosocomial settings due to their increased virulence and antibiotic resistance (Doern, 2024). Neonatal UTIs have a unique microbiological profile, often involving non-E. coli Enterobacteriaceae, especially in cases of urinary tract anomalies such as vesicoureteral reflux (VUR) (Álvarez et al., 2017). The developing antibiotic resistance patterns of many infections hamper treatment options, demanding ongoing monitoring and customization of empirical therapy.

2.11.2 Risk Factors and Vulnerable Populations

Several variables enhance the likelihood of having UTIs, with some populations being especially vulnerable. Pregnancy, for example, causes physiological changes that make women susceptible to UTIs. Routine screening and early intervention are crucial for pregnant women to prevent issues including preterm birth and low birth weight (Szweda & Jóźwik, 2016; Lee et al.,

2019; Habak et al., 2024). UTIs are the most common bacterial infections in post-kidney transplant patients, compromising graft function and greatly increasing the cost of post-transplant care (Sulaiman et al., 2024). Diabetes, catheter use, immunocompromised states, and anatomical anomalies all lead to increased UTI risk (Sabih & Leslie, 2024).

Sociodemographic characteristics are crucial, particularly in low-resource situations. According to studies from rural Tamil Nadu, India, and Bangladesh, a lack of awareness, healthcare access, and screening methods contribute to greater UTI prevalence (Muthulakshmi&Gopalakrishnan, 2017; Balakrishnan et al., 2022; Lee et al., 2019).

2.11.3 Economic Burden and Healthcare Implications

UTIs place a significant cost burden on healthcare systems globally. According to Foxman (2002), UTIs cause 7 million office visits, 1 million emergency room visits, and 100,000 hospitalizations in the United States each year, costing an estimated $1.6 billion. In Ireland, Callan et al. (2014) found that primary care costs are the primary economic driver of UTI management, but François et al. (2016) stressed the importance of reducing unnecessary diagnostic tests and treatments to lower healthcare costs. Ciani et al. (2013) established the idea of the "costs of resignation," which refers to patients who have recurring UTIs and incur continuous medical expenses as a result of inadequate early response. In Lebanon, Iskandar et al. (2021) highlighted the rising expenses of treating antibiotic-resistant E. coli UTIs, emphasizing the need for national measures to reduce resistance and enhance treatment efficiency.

2.12.4 Antibiotic Resistance and Treatment Challenges

Antibiotic resistance is becoming a major global concern in UTI management. The increasing use of antibiotics has resulted in the rise of resistant strains among common uropathogens, complicating empirical treatment. Iskandar et al. (2021) discovered that treating antibiotic-resistant E. coli infections leads to longer hospital stays and higher medical expenses. To limit the impact of resistance, treatment recommendations must be revised and antibiotic stewardship programs implemented. Furthermore, the lack of a clear consensus on antibiotic selection or treatment duration, particularly for UTIs in pregnancy, emphasizes the significance of adapting therapy to local resistance trends (Schnarr&Smaill, 2008).

2.12.5 Sociodemographic and Geographic Insights

Sociodemographic factors have a substantial impact on UTI prevalence and management. Rural areas in Tamil Nadu, India, have a higher susceptibility among women of reproductive age due to low healthcare access and awareness (Muthulakshmi&Gopalakrishnan, 2017; Balakrishnan et al., 2022). Similarly, in Bangladesh, the high frequency of asymptomatic UTIs among pregnant women highlights the need for low-cost screening approaches (Lee et al., 2019). Weather and climate variables may also influence UTI incidence. For example, Amiri et al. (2015) discovered a reduced prevalence of UTIs in some climates, such as Dezful City, Iran, implying that the environment may have an impact on infection rates.

2.12.6 Clinical Outcomes and Complications

If left untreated, UTIs can cause serious health consequences, especially in high-risk groups. UTIs in pregnant women increase the risk of pyelonephritis, preterm birth, and fetal death (Foxman, 2002). Recurrent UTIs in pediatric patients are linked to reduced renal function and end-stage renal disease. Early UTIs in kidney transplant recipients lead to longer hospital stays and higher healthcare expenses, despite the fact that they had no direct effect on graft function at 6 and 12 months post-transplant (Sulaiman et al., 2024). Complicated UTIs, as defined by anatomical abnormalities, immunocompromised states, and medical devices such as catheters, necessitate specialist diagnostic and therapeutic techniques due to the higher likelihood of treatment failure and recurrence (Sabih & Leslie, 2024). The body of research on UTIs emphasizes the condition's complexity, with a variety of infections, risk factors, and socioeconomic variables contributing to the global burden. Pregnant women, the elderly, diabetic patients, kidney transplant recipients, and catheterized individuals are high-risk categories who require specialized screening and care measures. Economic studies show the significant cost burden that UTIs impose on healthcare systems, which is caused by frequent primary care visits, recurrent infections, and the growing concern of antibiotic resistance. The idea of "costs of resignation" emphasizes the importance of improved diagnosis methods and early intervention in breaking the cycle of recurring infections and lowering healthcare costs.

CHAPTER 3

METHODOLOGY

3.1 Classification of Data in Statistics

Classification of data refers to the systematic organization of raw data into groups or categories based on shared characteristics or attributes. This process transforms unstructured data into a structured format, making it easier to analyze and draw meaningful conclusions. Data can be classified based on location, time, descriptive characteristics, and measurable characteristics.

3.1.1 Classification of Data

For performing statistical analysis, various kinds of data are gathered by the investigator or analyst. The information gathered is usually in raw form which is difficult to analyze. To make the analysis meaningful and easy, the raw data is converted or classified into different categories based on their characteristics. This grouping of data into different categories or classes with similar or homogeneous characteristics is known as the Classification of Data. Each division or class of the gathered data is known as a Class. The different basis of classification of statistical information are Geographical, Chronological, Qualitative (Simple and Manifold), and Quantitative or Numerical.

For example, if an investigator wants to determine the poverty level of a state, he/she can do so by gathering the information of people of that state and then classifying them on the basis of their income, education, etc. According to Conner, “Classification is the process of arranging things (either actually or notionally) in groups or classes according to their resemblances and affinities, and gives expression to the unity of attributes that may exist amongst a diversity of individuals.”

3.1.2 Basis of Classification of Data

The classification of statistical data is done after considering the scope, nature, and purpose of an investigation and is generally done on four bases; viz., geographical location, chronology, qualitative characteristics, and quantitative characteristics.

1. Geographical Classification

The classification of data on the basis of geographical location or region is known as Geographical or Spatial Classification. For example, presenting the population of different states of a country is done on the basis of geographical location or region.

2. Chronological Classification

The classification of data with respect to different time periods is known as Chronological or Temporal Classification. For example, the number of students in a school in different years can be presented on the basis of a time period.

3. Qualitative Classification

The classification of data on the basis of descriptive or qualitative characteristics like region, caste, sex, gender, education, etc., is known as Qualitative Classification. A qualitative classification cannot be quantified and can be of two types; viz., Simple Classification and Manifold Classification.

Simple Classification - When based on only one attribute, the given data is classified into two classes, which is known as Simple Classification. For example, when the population is divided into literate and illiterate, it is a simple classification.

Manifold Classification - When based on more than one attribute, the given data is classified into different classes, and then sub-divided into more sub-classes, which is known as Manifold Classification. For example, when the population is divided into literate and illiterate, then sub-divided into male and female, and further sub-divided into married and unmarried, it is a manifold classification.

4. Quantitative Classification

The classification of data on the basis of the characteristics, such as age, height, weight, income, etc., that can be measured in quantity is known as Quantitative Classification. For example, the weight of students in a class can be classified as quantitative classification.

3.2 Tabulation of Data

Now, to analyze the collected data, it is essential to present it in an easy-to-understand and interpretable way. The different ways the classified data can be presented are textual, tabular, diagrammatic, and graphical. Tabular Presentation or Tabulation is a systematic way of presenting numerical data in rows and columns. The tabular presentation helps the investigator in simplifying the presentation and facilitating analysis. It can bring the related information close to each other such that the investigator can easily make comparisons between them, and also helps in further statistical analysis and interpretation of the data. According to L.R. Connor, "Tabulation involves the orderly and systematic presentation of numerical data in a form designed to elucidate the problem under consideration. "

3.2.1 Advantages of Tabulation of Data

Tabulation arranges data systematically in rows and columns, making it easier to read and understand. Data presented in tables allows for straightforward comparison across different categories or groups. Tabulated data is easier to analyze, as it provides a clear and concise format that highlights key information. Tables condense large amounts of data into a compact format, saving time for those who need to interpret or analyze the data. By presenting data in a structured format, tabulation eliminates confusion and makes it easier to identify trends and patterns. Well-tabulated data provides a clear overview, aiding in quicker and more informed decision-making. Tables are essential for performing various statistical analyses, providing a foundation for calculations and evaluations.

3.2.2 Classification of Data Vs Tabulation of Data

Generally, classification of data and tabular presentation of data are misunderstood as the same; i.e., a device to present and summarize data. However, in technical terms, both concepts are different from each other. The difference between the classification of data and the tabular presentation of data is as follows:

1. Tabulation succeeds classification of data. It means that tabular presentation of data can be done only when it is classified into different classes.

2. Classification of data includes classifying the given set of data into different classes according to their similarities and differences. However, tabular presentation of data includes arranging the classified data into rows and columns with suitable heads and subheads.
3. Classification is a method of statistical analysis. However, tabular presentation of data is a method of presenting data.

3.2.3 Spatial Classification of Data and Tabular Presentation

Spatial Classification means to classify data based on the geographical location, place, or region such as state, district, town, city, country, etc. For example, a number of students from different states at Delhi University. The Tabular presentation of the same can be shown as follows

Sports	Number of Players
Cricket	25
Football	36
Table Tennis	13
Basketball	27

Fig 2.1 Spatial Classification

3.2.4 Temporal Classification of Data and Tabular Presentation

Temporal Classification of data means to classify data based on the time period. It means that time becomes the classifying variable in the case of temporal classification. **For example,** the sale of Laptops by a manufacturer in different years. The tabular presentation of the same can be shown as follows

Year	Sale (Units)
2015	25,000
2016	46,000
2017	70,000
2018	90,000
2019	1,00,000

Fig 2.2 Temporal

3.2.5 Qualitative Classification of Data and Tabular Presentation

Qualitative Classification of data means to classify data based on qualitative characteristics or attributes. **For example,** data of the students of Class XI can be classified on qualitative attributes like male or female, and Commerce or Science. The tabular presentation of the same can be shown as follows

Sex	Number of Students	
	Commerce	Science
Female	32	26
Male	15	30
Total	**47**	**56**

Fig 2.3 Qualitative Classification

3.2.5 Quantitative Classification of Data and Tabular Presentation

Quantitative Classification of data means to classify data based on quantitative characteristics. For example**,** data on the number of players playing different sports in a school. The tabular presentation of the same can be shown as follows

Sports	Number of Players
Cricket	25
Football	36
Table Tennis	13
Basketball	27

Fig 2.4 Quantitative Classification

Tabulation is critical at all levels of data processing, from data analysis to sophisticated approaches such as machine learning and deep learning. Tabulated data is the typical input for many machine learning methods, including logistic regression, decision trees, and random forests. Feature engineering, encoding, and scaling frequently require data presented in tabular form. Even in deep learning, while unstructured data such as images and text are widespread, tabular data is still employed for tasks such as fraud detection, healthcare forecasts, and recommender systems, which employ fully connected neural networks (dense layers) on structured tables.

3.3 Descriptive Statistics

Descriptive statistics are brief informational coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables, kurtosis, and skewness.

3.3.1 Measures of central tendency

According to Professor Bowley, averages are "Statistical constants which enable to comprehend in a single effect the significance of the whole". They give us an idea about the concentration of the values in the central part of the distribution and hence the name measures of central tendency. Plainly speaking, an average of a statistical series is the value of the variable which is representative of the entire distribution. The following are the five measures of central tendency that are in common use.

Arithmetic mean

Arithmetic mean of a set of observations is their sum divided by the number of observations. e.g the arithmetic mean $\bar{x}$ of n observations x_1 x_2 ….x_n. is given by

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

If the frequency distribution .x; If *i=1,2,3,….n*. where is the frequency of the variable *Xi ;*

$$\bar{x} = \frac{f_1x_1+f_2x_2+\cdots+f_nx_n}{f_1+f_2+\cdots+f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{1}{N}\sum_{i=1}^{n} f_i x_i$$

In case of grouped or continuous frequency distribution. *X* is taken as the mid. value of the corresponding class.

Median

Median of a distribution is the value of the variable which divides it into two equal pans. It is the va1ue which exceeds and is exceeded by the same number of observations, *i.e.*, it, is the value such that the number of observations above it is equal to the number of observations below it the median is thus a positional average.

If the number of observations is odd then median is the middle value after the values have been arranged in ascending or descending order of magnitude. In case of even number of observations, there are two middle terms and median is obtained by taking the arithmetic mean of the middle terms.

Mode

In Statistics, the mode is the number that appears most often in a set of data. It's a measure of central tendency that indicates the most popular choice or most common characteristic of a sample.

To find the mode, count how often each number appears and the number that appears the most times is the mode.

Geometric Mean

Geometric mean of a set of *n* observations is the *nth* root of their product thus the geometric mean G of *n* observations $x_i i$ = 1,2,3,…,n is

$$\mathrm{G} = (x_{1,} x_{2\ldots\ldots} x_n)^{1/\mathrm{n}}$$

$$log\mathrm{G} = \frac{1}{n}(logx_1 + logx_2 + \cdots + \mathrm{G}logx_n) = \frac{1}{n}\sum_{i=1}^{n} logx_i$$

$$\mathrm{G} = \text{Antilog}\left[\frac{1}{n}\sum_{i=1}^{n} logx_i\right]$$

Harmonic mean

Harmonic mean of a number of observations is the reciprocal of the arithmetic of the: reciprocals the given values. Thus, harmonic mean *H,* of *n* observations $x_i i$ = 1, 2, 3,…,n is

$$H= \frac{1}{\frac{1}{n}\sum_{i=1}^{n}(\frac{1}{x_i})}$$

In the case of frequency distribution $x_i|f_i$, (*i* = 1, 2, 3,…,*n)*

$$H = \frac{1}{\frac{1}{N}\sum_{i=1}^{n}(\frac{f_i}{x_i})} \quad , \quad [N = \sum_{i=1}^{n} f_i]$$

3.3.2 Measures of Dispersion

Range

The range is the difference between two extreme. Observations, or the distribution. If *A* and *B* are the greatest and smallest observations respectively in a distribution, then its range is *A-B.* Range is the simples but a crude measure of dispersion. Since it is based on two extreme observations which themselves are subject to change fluctuations, it is not at all a reliable measures of dispersion.

Quartile deviation

Quartile deviation or Semi-quartile range Q is given by

$$Q = \frac{1}{2}\ (Q_3 - Q_1)$$

where Q_3 and Q_1 are the first and third quartiles of the distribution respectively. Quartile deviation is definitely a better measure than the range as it makes use of 50% of the data. But since it ignores the other 50% of the data, it cannot be regarded as a reliable measure.

Mean deviation

If $x_i|f_i$ *i* = 1,2…, *n* is .the frequency distribution, then mean deviation from the average *A,* (usually mean, median or mode), is given by

$$\text{Mean deviation} = \frac{1}{N}\sum f_i|x_i - A|\ , \quad \sum f_i = \text{N}$$

Where $|x_i - A|$ represents the modulus or the absolute value of the deviation *(Xi - A)* when the negative sign is ignored.

Since mean deviation is based on all the observations. it is a better measure of dispersion than range or quartile deviation. But the step of ignoring the signs of the deviations (*Xi - A*)creates artificiality and, renders it useless for –further mathematical treatment.

Standard deviation

Standard deviation, usually denoted by the Greek letter small sigma (σ), is the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean. For the frequency distribution $x_i|f_i$ *i* = 1,2, ... , *n.*

$$\sigma = \sqrt{\frac{1}{n}\sum_i f_i(x_i - \tilde{x})^2}$$

Where $\tilde{x}$ is arithmetic mean of the distribution and $\sum_i f_i$= N

The square of standard deviation is called the variance and is given by

$$\sigma^2 = \frac{1}{N}\sum_i f_i\ (x_i - \tilde{x})^2$$

Root mean square deviation, denoted *by 'S'* is given by

$$S = \sqrt{\frac{1}{N}\sum_i f_i\ (x_i - A)^2}$$

Where A is any arbitrary number $s^2$is called mean, square, deviation.

These statistics are critical for spotting patterns, detecting outliers, and gaining preliminary insights before employing more complicated statistical or machine learning methods. In real-world applications, descriptive statistics are used to make data-driven decisions, publish study findings, and support hypotheses. For example, in healthcare analytics, they might summarise patient demographics or clinical measures, but in business, they can help analyze sales patterns or consumer behaviour.

3.4 Graphical Representation

3.4.1 Bar Graph

Bar graphs are one of the most common and versatile types of charts used to represent categorical data visually. They display data using rectangular bars, where the length or height of each bar corresponds to the value it represents. Bar graphs are widely used in various fields such as business, education, and research to compare different categories or track changes over time. This article explores the different types of bar graphs, their uses, and how to create and interpret them.

Simple Bar Graph

A diagram in which each class or category of data is represented by a group of rectangular bars of equal width is known as a **Simple Bar Diagram**. It is the simplest type of bar diagram. In this diagram, each bar represents one figure only. The number of bars will be equal to the number of figures. These diagrams show only one characteristic of the data, such as sales, production, or population figures for various years.

The magnitude of data is determined by the bar's height (or length). The lower end of the bar touches the base line; therefore, the height of a bar starts from the zero unit. These diagrams can be vertical or horizontal in layout:

1. **Vertical Bar Graph:** The diagram in which the magnitude of the data is presented vertically, i.e., along the Y-axis, is a Vertical Bar Diagram.
2. **Horizontal Bar Graph:** The diagram in which the magnitude of the data is presented horizontally; i.e., along the X-axis is a Horizontal Bar Diagram.

Multiple Bar Graph

The Multiple Bar Diagram is used to compare two or more variables such as revenue and expenditure, import and export for different years, marks obtained in different subjects in different grades, and so on. It is often referred to as a Compound Bar Diagram. The method for creating multiple bar diagrams is the same as for making a Simple Bar Diagram. However, to

distinguish the bars from each other, different bars are differentiated by different shades or colours.

Sub-Divided Bar Graph

In these diagrams, the bar corresponding to each phenomenon is divided into several components. Each part or component occupies a proportional part of the bar to its share in the total. **For example,** the bar corresponding to the number of students enrolled in a course can be further sub-divided into boys and girls.

Percentage Bar Graph

A Percentage Bar Diagram is a sub-divided bar diagram that indicates the total percentage of each component rather than the magnitude. The absolute magnitudes of several components are presented using a subdivided diagram. These magnitudes can be converted into relative values by describing them as a percentage of the total.

Each data component is expressed as a percentage of the corresponding total. Thus, in a percentage bar diagram, all of the bars are of height 100, while the different segments of the bar representing the various components vary in size depending on their % value of the total. Just like in the sub-divided bar diagram, in the percentage bar diagram, different components can be differentiated by different shades or colours.

3.4.2 Histogram

A histogram is a graphical representation of the frequency distribution of continuous series using rectangles. The x-axis of the graph represents the class interval, and the y-axis shows the various frequencies corresponding to different class intervals. A histogram is a two-dimensional diagram in which the width of the rectangles shows the width of the class intervals, and the length of the rectangles depicts the corresponding frequency.

There are no gaps between two consecutive rectangles based on the fact that histograms can be drawn when data are in the form of the frequency distribution of a **continuous series.** No histogram can be drawn for a data set in the form of discrete series, and this makes histograms different from bar graphs as they can be plotted for both discrete and continuous series. The

major difference between a histogram and a bar graph is that the former is two-dimensional; i.e., both the width and length of the rectangles are used for comparison, whereas the latter is one-dimensional, which means only the length of the rectangles is used for comparison. A histogram is used to determine the value of **the Mode** of a data set in the form of a continuous series.

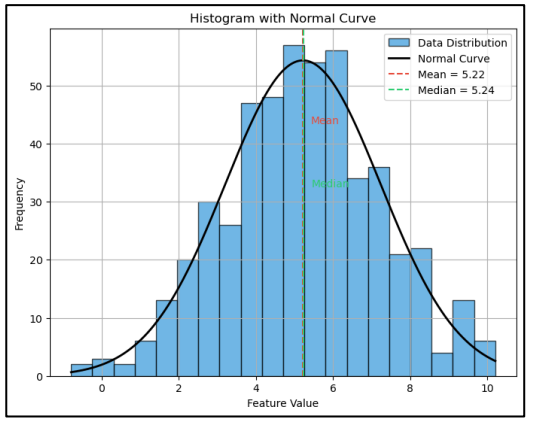


Fig 2.5 Histogram

Graphical representation is an important part of data analysis, machine learning, and deep learning because it converts complex data into visual formats that are easier to grasp and interpret. In data analysis, graphs such as histograms, bar charts, box plots, and scatter plots are used to investigate data distributions, identify trends, locate outliers, and comprehend variable relationships.

These visual tools help you make informed decisions and identify trends that may not be visible in raw data. Confusion matrices, ROC curves, and feature significance plots are examples of visualizations used in machine learning to evaluate model performance, understand findings, and compare different techniques. Similarly, in deep learning, graphical representations such as training vs. validation loss curves, activation maps, and learning rate plots are critical for tracking model training, diagnosing overfitting concerns, and increasing model accuracy. Overall, graphical representation improves the clarity, communication, and effectiveness of data-driven insights at every step of analysis and modeling.

3.5 Correlation

Correlation is a statistical measure that describes the extent to which two variables change together. In other words, it assesses the strength and direction of the linear relationship between them. Correlation does not imply causation. A strong correlation between two variables does not necessarily mean that one causes the other. Pearson's correlation measures linear relationships. It may not accurately reflect non-linear relationships. It is important to visualize data with scatter plots, to have a visual representation of the correlation. Correlation can be heavily affected by outliers.

Pearson's Correlation Coefficient (r)

This is the most widely used measure of linear correlation. It quantifies the strength and direction of a linear relationship between two continuous variables. The value of r ranges from -1 to +1. +1 indicates a perfect positive linear relationship. -1 indicates a perfect negative linear relationship. 0 indicates no linear relationship.

Formula 1 (using standard deviations and covariance):

$$r = \frac{cov(x, y)}{s_x s_y}$$

Where

- r is the Pearson correlation coefficient.
- $cov(x, y)$ is the covariance of x and y.
- s_x is the standard deviation of x.
- s_y is the standard deviation of y.

Formula 2 (more direct calculation)

$$r = \frac{\sum(x_i - y_i)(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where

- x_i and y_i are the individual data points.
- $\bar{x}$ and $\bar{y}$ are the means of x and y, respectively.

Explanation of the Formulas

Covariance (cov(x, y))

Measures how much two variables change together. A positive covariance indicates that the variables tend to increase or decrease together. A negative covariance indicates that one variable tends to increase when the other decreases.

Standard Deviation (s_x, s_y)

Measures the dispersion of a variable's values. It normalizes the covariance, allowing for a standardized measure of correlation. The second formula breaks down the covariance and standard deviation calculations into more elemental sums, which is useful for hand calculations.

Correlation heat-map

Data Preparation

The process starts with a dataset containing multiple continuous variables. These variables represent measurements or scores that can take on a range of numerical values. The goal is to understand how these variables relate to each other.

Correlation Matrix Calculation

A correlation matrix is computed. This matrix quantifies the pair wise linear relationships between all the variables. The most common method used is Pearson's correlation, which measures the strength and direction of linear association.

This is the most widely used measure of linear correlation. It quantifies the strength and direction of a linear relationship between two continuous variables. The value of r ranges from -1 to +1. But we have to know that correlation does not imply causation. A strong correlation between two variables does not necessarily mean that one causes the other.

Heat-map Visualization

The correlation matrix is then visualized as a heat-map. This visual representation uses a color-coded grid to display the correlation values.

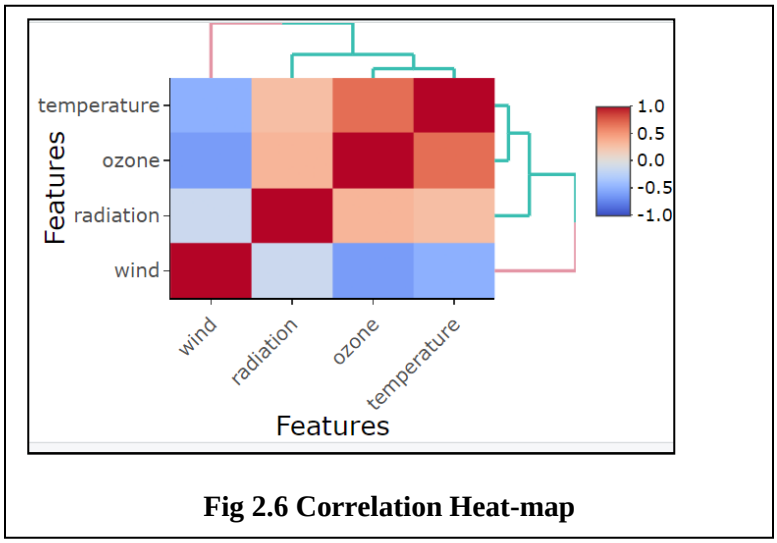


Fig 2.6 Correlation Heat-map

Heat-map Interpretation

The diagonal of the heat-map (from top-left to bottom-right) always shows a perfect positive correlation (represented by a strong color, typically red). This confirms that each variable is perfectly correlated with itself.

Off-Diagonal Analysis

The off-diagonal cells represent the correlations between different pairs of variables.

Color Intensity

The intensity of the color indicates the strength of the correlation. Darker colors signify stronger relationships.

Color Direction

The color itself indicates the direction of the correlation. One color (often red) represents positive correlations, where variables tend to increase together. Another color (often blue) represents negative correlations, where one variable increases as the other decreases. Light or neutral colors indicate weak or no linear correlation.

Numerical Values

The numerical values displayed within each cell provide the precise correlation coefficients, allowing for a more accurate interpretation.

Relationship Identification

By examining the colors and numerical values, we can identify pairs of variables that exhibit strong positive, strong negative, or weak correlations. This helps in understanding which variables tend to vary together.

Overall Pattern Recognition

The heat-map provides a comprehensive overview of the correlation patterns within the dataset, allowing for quick identification of key relationships.

Application

The insights gained from the correlation heat-map can be used for various purposes. Identifying potential predictors for modeling. Understanding the underlying structure of the data. Detecting multi-collinearity (high correlation between predictor variables). Guiding further statistical analysis.

Correlation is an important tool in data analysis since it measures the strength and direction of correlations between variables. It aids in the identification of relevant patterns, such as associations between age and disease risk in healthcare or advertising and sales in business. Strong correlations can help drive feature selection in machine learning, as well as detect multi-collinearity, which can have an impact on model performance. Overall, correlation facilitates exploratory investigation, hypothesis testing, and sound decision-making.

3.6 Proportion test

A proportion test is a statistical hypothesis test used to compare observed proportions to expected proportions or to compare proportions between two or more groups. It helps determine if there's a statistically significant difference between these proportions**.**

Two-Proportion Z-Test

This test is used to determine if there's a statistically significant difference between the proportions of a specific characteristic in two independent groups. Essentially, it compares two sample proportions to see if they likely came from populations with different underlying proportions.

Hypothesis

Null Hypothesis (H_0): The proportions in the two populations are equal ($p_1 = p_2$).

Alternative Hypothesis (H_1):

- Two-tailed: The proportions in the two populations are not equal ($p_1 \neq p_2$).
- Right-tailed: The proportion in population 1 is greater than the proportion in population 2 ($p_1 > p_2$).
- Left-tailed: The proportion in population 1 is less than the proportion in population 2 ($p_1 < p_2$).

Formula

The test statistic (z-score) is calculated using the following formula:

$$Z = \frac{\left(\hat{P}_1 - \hat{P}_2\right)}{\sqrt{\hat{P}\left(1 - \hat{P}\right)\left(\frac{1}{n_1} - \frac{1}{n_2}\right)}}$$

Where

- $\hat{P}_1$ is the sample proportion from group 1.
- $\hat{P}_2$ is the sample proportion from group 2.
- n_1 is the sample size of group 1.

- n_2 is the sample size of group 2.
- $\widehat{P}$ is the pooled sample proportion, calculated as

$$\hat{P} = \frac{x_1 + x_2}{n_1 + n_2}$$

Where x_1 is the number of successes in group 1, and x_2 is the number of successes in group 2.

The proportion test is a statistical approach that compares observed proportions to expected proportions or proportions between two or more groups in order to assess whether there is a statistically significant difference. One popular form is the Two-Proportion Z-Test, which analyzes the proportions of a feature in two independent groups to see if their underlying populations differ. The test involves establishing a null hypothesis (H_0) that the proportions are equal and an alternative hypothesis (H_1) indicating a difference. The Z-score formula incorporates the difference in sample proportions, adjusted for sample size, and the pooled proportion of successes across both groups.

The Z-score formula contains the difference in sample proportions, adjusted for sample size, and the pooled proportion of successes across both groups. This test is frequently employed in a variety of sectors, including evaluating success rates across different treatments, marketing techniques, and demographic groupings, and it gives a quantitative method for determining whether observed differences in proportions are due to chance or indicate a meaningful effect.

Applications

This test is commonly used in a variety of applications. For example, in healthcare, it can be used to compare the efficacy of two distinct therapies in terms of recovery rates. In marketing, it may be useful to assess whether the proportion of customers who prefer one product differs considerably from the proportion who prefer another. It is also used in social sciences to compare proportions among various demographic groups. The percentage test is useful for decision-making and strategy formulation because it allows analysts to determine whether observed differences are likely to be random or reflect a meaningful influence.

3.7 One-way Analysis of Variance

A one-way ANOVA (Analysis of Variance) is a statistical test used to determine whether there are any statistically significant differences between the means of three or more independent groups. The primary goal is to examine if the independent variable (categorical, with multiple levels/groups) has an effect on the dependent variable (continuous). When a researcher wants to compare the means of numerous groups based on a single independent categorical variable, he or she will utilize One-Way ANOVA. For example, a researcher could use ANOVA to compare the average test scores of students from different schools or the average wealth across regions. The primary assumption is that the groups being compared are independent, and ANOVA aids in assessing whether observed differences in group averages are statistically significant or occurred by chance. This assists in identifying factors that have a significant impact on the dependent variable.

- **Independent Variable (Factor) -** The categorical variable that defines the groups being compared.
- **Dependent Variable (Response Variable) -** The continuous variable that is measured.
- **Groups/Levels -** The different categories or values of the independent variable.

Hypothesis

- **Null Hypothesis (H_0):** The means of all groups are equal.
- **Alternative Hypothesis (H_1):** At least one group mean is different from the others.

ANOVA partitions the total variability of the data into two components:

- **Between-group variability:** The variability between the means of the different groups.
- **Within-group variability:** The variability within each group (due to random error).

It calculates an F-statistic, which is the ratio of between-group variability to within-group variability. A large F-statistic indicates that the between-group variability is significantly greater than the within-group variability, suggesting that the group means are different. The F-statistic is then used to determine a p-value, which indicates the probability of observing the data if the null hypothesis were true.

Assumptions

- **Independence -** The observations within each group are independent.
- **Normality -** The dependent variable is normally distributed within each group.
- **Homogeneity of Variance (Homoscedasticity) -** The variances of the dependent variable are equal across all groups.

Steps

1. **State the Hypotheses -** Define the null and alternative hypotheses.
2. **Calculate the F-statistic**

This involves calculating the sum of squares between groups (SSB), the sum of squares within groups (SSW), the mean square between groups (MSB), and the mean square within groups (MSW).

The F-statistic is then calculated as

$$F = \frac{MSB}{MSW}$$

3. **Determine the p-value -** Find the p-value associated with the calculated F-statistic.
4. **Make a Decision -** Compare the p-value to the significance level (alpha, typically 0.05).

- If the p-value ≤ alpha, reject the null hypothesis.
- If the p-value > alpha, fail to reject the null hypothesis.

In machine learning, One-Way ANOVA is a popular feature selection strategy, particularly when working with categorical data. For example, while developing a classification model, ANOVA can assist in determining which categorical factors (such as client groups, geographies, or marketing techniques) have a substantial impact on the target variable. Analysts can select characteristics that contribute the most to predictive performance by testing the variation in the dependent variable at various levels of each categorical feature.

ANOVA Table

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F-statistic
Between Groups	SSB	k – 1	MSB = SSB / (k - 1)	F = MSB / MSW
Within Groups	SSW	N – k	MSW = SSW / (N - k)	
Total	SST = SSB + SSW	N – 1		

SSB (Sum of Squares Between)

$$SSB = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

Where

- k = number of groups
- n_i = number of observations in group i
- $\bar{x}_i$ = mean of group i
- $\bar{x}$ = overall mean

SSW (Sum of Squares Within)

$$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i\right)^2$$

Where

- x_{ij} = jth observation in group i

SST (Sum of Squares Total)

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}\right)^2$$

Or simply: SST = SSB + SSW

k: Number of groups.

N: Total number of observations.

MSB (Mean Square Between)

$$MSB = \frac{SSB}{k-1}$$

MSW (Mean Square Within)

$$MSW = \frac{SSW}{N-k}$$

F (F-statistic)

$$F = \frac{MSB}{MSW}$$

P-value

The probability of obtaining the observed F-statistic (or a more extreme value) if the null hypothesis is true. This is obtained from the F-distribution using the calculated F-statistic and the degrees of freedom.

ANOVA is frequently used in preprocessing to determine whether different features (such as marketing channels) should be addressed differently in the model. Although deep learning models often work with big, unstructured datasets (such as images or text), One-Way ANOVA can still be useful in feature engineering.

3.8 Linear Regression

Linear regression is a statistical method used to model the linear relationship between a dependent variable and one or more independent variables.

Core Concept

The goal is to find the "best-fit" line (or hyperplane in multiple regression) that describes how the dependent variable changes as the independent variable(s) change. This line can then be used to predict the value of the dependent variable for given values of the independent variable(s).

Types

Simple Linear Regression - Involves one independent variable. The relationship is modeled by a straight line.

Multiple Linear Regression - Involves two or more independent variables. The relationship is modeled by a hyperplane.

The Linear Equation

Simple Linear Regression

$$y = mx + b$$

Where

- y: dependent variable
- x: independent variable
- m: slope of the line (how much y changes for a one-unit change in x)
- b: y-intercept (the value of y when x is 0)

Multiple Linear Regression

$$y = b_0 + b_1x_1 + b_2x_2 + ... + b_nx$$

Where

- y: dependent variable
- x_1, x_2, ..., x_n : independent variables
- b_0: y-intercept
- b_1, b_2, ..., b_n coefficients (slopes) for each independent variable

Key Aspects

Finding the "Best-Fit" Line. The coefficients (m and b in simple regression, b_0, b_1, etc., in multiple regression) are determined by minimizing the "least squares." This means minimizing the sum of the squared differences between the actual y-values and the predicted y-values.

Assumptions

- Linearity - The relationship between the variables is linear.
- Independence - The observations are independent.
- Homoscedasticity - The variance of the errors is constant.
- Normality - The errors are normally distributed.

Evaluation

R-squared - Measures the proportion of the variance in the dependent variable that is explained by the independent variable(s). Values range from 0 to 1, with higher values indicating a better fit.

P-values - Used to assess the statistical significance of the coefficients. Low p-values indicate that the coefficients are likely not zero, suggesting a significant relationship.

Residual analysis - Checking the residuals, (the difference between the observed and predicted values) for violations of the assumptions.

F statistic – Used to test the overall significance of the regression model.

Predicting sales based on advertising spending. Estimating house prices based on square footage and other factors. Analyzing the relationship between temperature and ice cream sales.

Suppose the outcome of any process is denoted by a random variable y , called as dependent (or study) variable, depends on k independent (or explanatory) variables denoted by $X_1, X_2, X_3, \dots\dots, X_k$, Suppose the behaviour of y can be explained by a relationship given by

$$y = f(X_1, X_2, X_3, \dots\dots, X_k, \beta_1, \beta_2, \beta_3, \dots\dots, \beta_k)$$

Where f is some well-defined function and $\beta_1, \beta_2, \beta_3, \dots\dots, \beta_k$are the parameters which characterize the role and contribution of$X_1, X_2, X_3, \dots\dots, X_k$, respectively. The term ε reflects the stochastic nature of the relationship between y and $X_1, X_2, X_3, \dots\dots, X_k$and indicates that such a relationship is not exact in nature. When $\varepsilon = 0$, then the relationship is called the mathematical model otherwise the statistical model. The term "model" is broadly used to represent any phenomenon in a mathematical framework.

A model or relationship is termed as linear if it is linear in parameters and non-linear, if it is not linear in parameters. In other words, if all the partial derivatives of y with respect to each of the parameters $\beta_1, \beta_2, \beta_3, \dots\dots, \beta_k$are independent of the parameters, then the model is called as a linear model. If any of the partial derivatives of y with respect to any of the $\beta_1, \beta_2, \beta_3, \dots\dots, \beta_k$is not independent of the parameters, the model is called non-linear. Note that the linearity or non-linearity of the model is not described by the linearity or non-linearity of explanatory variables in the model.

In machine learning, regression approaches are used to create predictive models, especially for supervised learning tasks. Linear regression is one of the simplest and most extensively used algorithms, with the purpose of predicting a continuous target variable (e.g., price, temperature) using one or more features. To avoid over-fitting, polynomial regression and more advanced approaches such as Ridge and Lasso regression are utilized in conjunction with more intricate interactions and regularization.

Regression is used to study and model the relationships between variables. In machine learning, it is used for predictive modeling and forecasting, particularly when the output is continuous. Deep learning uses regression within neural networks to model complicated non-linear relationships and predict continuous outcomes, especially on huge and unstructured datasets. In all circumstances, regression aids in quantifying correlations and producing credible forecasts that can be used to drive decision-making processes.

3.9 Chi-Square test

The Chi-Square test is one of the most widely used techniques for hypothesis testing. Whether you're analyzing categorical data, testing for independence between variables, or assessing goodness-of-fit, the Chi-Square test provides a robust method for validating assumptions and drawing meaningful conclusions. The Chi-Square test is a statistical procedure for determining the difference between observed and expected data. It can also be used to decide whether the data correlates with our categorical variables. Thus, it helps determine whether a difference between two categorical variables is due to chance or a relationship between them.

A Chi-Square or comparable nonparametric test is required to test a hypothesis regarding the distribution of a categorical variable. Categorical variables, which indicate categories such as animals or countries, can be nominal or ordinal. They cannot have a normal distribution because they have only a few particular values.

Chi-Square Test Formula

$$\chi_c^2 = \frac{\sum(O_i - E_i)^2}{E_i}$$

where

c = Degrees of freedom

O = Observed Value

E = Expected Value

The degrees of freedom in a statistical calculation represent the number of variables that can vary. The degrees of freedom can be calculated to ensure that Chi-Square tests are statistically valid. These tests frequently compare observed data with data expected to be obtained if a particular hypothesis were true.

- The Observed values are those you gather yourselves.
- The Expected values are the anticipated frequencies, based on the null hypothesis.

There are two main types of Chi-Square tests

Independence

The Chi-Square test of Independence is a derivable (also known as inferential) statistical test that examines whether the two sets of variables are likely to be related to each other or not. This test is used when we have counts of values for two nominal or categorical variables and is considered a non-parametric test. A relatively large sample size and independence of observations are the required criteria for conducting this test.

Goodness-of-Fit

In statistical hypothesis testing, the Chi-Square Goodness-of-Fit test determines whether a variable is likely to come from a given distribution. We must have a set of data values and an idea of the distribution of this data. We can use this test when we have value counts for categorical variables. This test demonstrates a way of deciding if the data values have a "good enough" fit for our idea or if it is a representative sample data of the entire population.

Define the Hypothesis

H_0: There is no link between gender and political party preference.

H_1: There is a link between gender and political party preference.

You compare the obtained statistics to the critical ones in the chi-square table. As you can see, for an alpha level of 0.05 and two degrees of freedom, the critical statistic is 5.991, less than our obtained statistic of 9.83. You can reject our null hypothesis because the critical statistic is higher than your obtained statistic.

The Chi-Square Test of Homogeneity is used to determine whether various populations have the same distribution of a categorical variable. For example, it can determine whether different age groups have similar preferences for a specific product. This test evaluates the distribution of a categorical variable in two or more populations to see if they are similar or different.

3.10 Logistic Regression

Logistic regression is a statistical method used to model the relationship between a dependent (or response) variable and one or more independent (or explanatory) variables when the dependent variable is categorical. It is commonly used for binary classification problems, where the outcome variable has two possible values (e.g., success/failure, 0/1, yes/no).

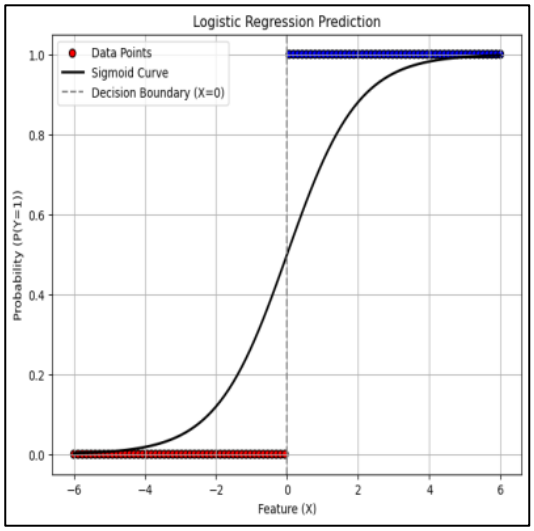


Fig 2.7 Logistic Regression Plot

Mathematical Formulation

Let Y be the binary response variable that takes values 0 or 1. The probability that Y=1 is given by:

$$P(Y = 1 \parallel X_1, X_2, \dots, X_k) = p$$

Where $X_1, X_2, \dots, X_k$ are the independent variables. Instead of modeling p directly, logistic regression models the log-odds (logit) of the probability as a linear function of the independent variables

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are the regression coefficients that determine the effect of each independent variable on the log-odds of the outcome.

The Uses of Logistic Regression

1. **Handles Non-linearity in Probability** - Unlike linear regression, logistic regression models the probability using a logistic function (sigmoid curve), ensuring that predicted values always lie between 0 and 1.

2. **Interpretable Coefficients** - The coefficients β represent the change in log-odds for a one-unit change in the predictor variable.

3. **Useful for Classification** - The predicted probability p can be converted into a binary decision using a threshold (e.g., 0.5).

Logistic Function (Sigmoid Function)

To ensure the output is between 0 and 1, logistic regression applies the sigmoid function

The hypothesis function is

$$P(Y = 1|X) = H_\beta(X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}}$$

Where

- $H_\beta(X)$: Predicted probability that Y=1,
- β_0: Intercept (bias term),
- $\beta_0 + \beta_1 + \beta_2 + \cdots + \beta_n$: Coefficients (weights) for each predictor,
- $x_1 + x_2 + \cdots + x_n$: Input features (e.g., symptoms, history of urinary problems, etc.).

Interpretation

- The expression $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$ is the linear combination of predictors.
- The sigmoid function maps this linear combination to a value between 0 and 1.

- If $H_\beta(X) \geq 0.5$, we classify the outcome as 1 (e.g., UTI positive),
- If $H_\beta(X) < 0.5$, we classify the outcome as 0 (e.g., UTI negative).

Model Assumptions

- The dependent variable is binary.
- The independent variables influence the log-odds of the dependent variable.
- There is no perfect **multicollinearity** among the independent variables.
- The observations are independent.

Types of Logistic Regression

1. **Binary Logistic Regression** - Used when the dependent variable has two categories (e.g., Yes/No).
2. **Multinomial Logistic Regression** - Used when the dependent variable has three or more unordered categories.
3. **Ordinal Logistic Regression** - Used when the dependent variable has three or more ordered categories.

In real-world applications, logistic regression is still quite useful, especially in sectors that need clear interpretation, quick implementation, and consistent performance with structured data. For example, in healthcare, it is frequently used to forecast the likelihood of diseases such as diabetes or urinary tract infections (as in your study), based on patient history and clinical assessments. Its capacity to produce interpretable data makes it perfect for assisting clinical judgments in which knowing why a prediction was made is just as crucial as the forecast. In finance, logistic regression helps assess credit risk by estimating the likelihood of a borrower defaulting, allowing institutions to make more informed lending decisions. Marketing teams use it to forecast customer behavior, such as the likelihood of purchasing or leaving, based on previous encounters. These domains frequently rely on organized, tabular data and benefit from the model's transparency, efficiency, and usability, particularly when deep learning would be overly complex or difficult to understand.

3.11 Confusion Matrix

A **confusion matrix** is a performance evaluation metric for classification models, especially in **binary classification** (e.g., UTI Diagnosis: Yes/No). It compares the actual values with the predicted values to assess how well a model is performing.

1. Structure of a Confusion Matrix

For a **binary classification** problem, the confusion matrix is a **2x2 table**

Actual \ Predicted	**Predicted: No (0)**	**Predicted: Yes (1)**
Actual: No (0)	True Negative (TN)	False Positive (FP)
Actual: Yes (1)	False Negative (FN)	True Positive (TP)

Where

- **True Positive (TP)** - Correctly predicted **Yes** (e.g., patient has UTI, and model predicts UTI).
- **False Positive (FP)** - Incorrectly predicted **Yes** (e.g., patient doesn't have UTI, but model predicts UTI). Also called **Type I Error**.
- **False Negative (FN)** - Incorrectly predicted **No** (e.g., patient has UTI, but model predicts No UTI). Also called **Type II Error**.
- **True Negative (TN)** - Correctly predicted **No** (e.g., patient doesn't have UTI, and model predicts No UTI).

Performance Metrics Derived from Confusion Matrix

Using the confusion matrix, we can compute several key performance metrics

1. Accuracy

The overall correctness of the model

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- High accuracy means the model is making mostly correct predictions.
- Drawback: Accuracy can be misleading if the dataset is imbalanced (e.g., 95% "No UTI" and 5% "UTI").

2. Precision (Positive Predictive Value - PPV)

Precision is how many of the predicted **Yes (UTI)** cases were actually correct

$$Precision = \frac{TP}{TP + FP}$$

- High precision means fewer false positives.
- Important when false positives are costly (e.g., misdiagnosing a disease when the patient is healthy).

3. Recall (Sensitivity / True Positive Rate - TPR)

Recall is how many of the actual **Yes (UTI)** cases were correctly identified

$$Recall = \frac{TP}{TP + FN}$$

- High recall means fewer false negatives.
- Important when false negatives are costly (e.g., missing a cancer diagnosis).

4. Specificity (True Negative Rate - TNR)

Specificity is how many of the actual **No (No UTI)** cases were correctly identified

$$Specificity = \frac{TN}{TN + FP}$$

- High specificity means fewer false positives.

5. F1-Score

Balances **Precision** and **Recall** into a single metric

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

- Useful when data is **imbalanced** (e.g., 90% "No UTI", 10% "UTI").

6. False Positive Rate (FPR)

The proportion of actual **No (healthy patients)** incorrectly classified as **Yes (UTI patients)**:

$$FPR = \frac{FP}{FP + TN}$$

7. False Negative Rate (FNR)

The proportion of actual **Yes (UTI patients)** incorrectly classified as **No (healthy patients)**

$$FNR = \frac{FN}{FN + TP}$$

The Confusion Matrix and Performance Metrics are crucial for assessing machine learning and deep learning models, especially in classification tasks. They help determine how successfully the model classifies cases and where errors occur. Applications include healthcare diagnosis, fraud detection, image classification, and sentiment analysis. Performance measures such as Precision, Recall, and F1-Score are critical for evaluating models, particularly when dealing with imbalanced data. These indicators help to fine-tune models and improve performance. Confusion Matrices identify problems like as False Positives and False Negatives, which aid in the model improvement process. AUC-ROC is used in recommender systems to assess ranking performance. In time series forecasting, confusion matrices aid in categorization problems. Overall, these methods ensure that models provide accurate and consistent predictions in real-world scenarios.

3.12 AUC-ROC Curve

The **AUC-ROC Curve** (Area under the Receiver Operating Characteristic Curve) is a crucial metric for evaluating classification models, especially when dealing with imbalanced data. It helps measure how well a model distinguishes between two classes.

1. Understanding ROC Curve

The **Receiver Operating Characteristic (ROC) Curve** is a plot of:

- **True Positive Rate (TPR) = Sensitivity = Recall** on the **Y-axis**
- **False Positive Rate (FPR)** on the **X-axis**

Where

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

- **TPR (Recall/Sensitivity)**: Measures how many actual positive cases were correctly identified.
- **FPR**: Measures how many actual negative cases were incorrectly classified as positive.

How the ROC Curve is Created

- The model predicts probabilities instead of strict 0/1 classifications.
- Different probability thresholds (e.g., 0.2, 0.5, 0.8) are tested.
- At each threshold, the **TPR and FPR** are calculated.

2. Understanding AUC (Area Under Curve)

The **AUC (Area Under the Curve)** quantifies the overall ability of the model to distinguish between classes.

Interpreting AUC Values

AUC Score	Interpretation
0.5	No discrimination (random guessing)
0.6 - 0.7	Poor model
0.7 - 0.8	Fair model
0.8 - 0.9	Good model
0.9 - 1.0	Excellent model

- AUC = 1.0 → Perfect classifier (ideal scenario).
- AUC = 0.5 → Model is making random predictions.
- AUC < 0.5 → Model is worse than random guessing (incorrectly classifies more than 50% of the cases).

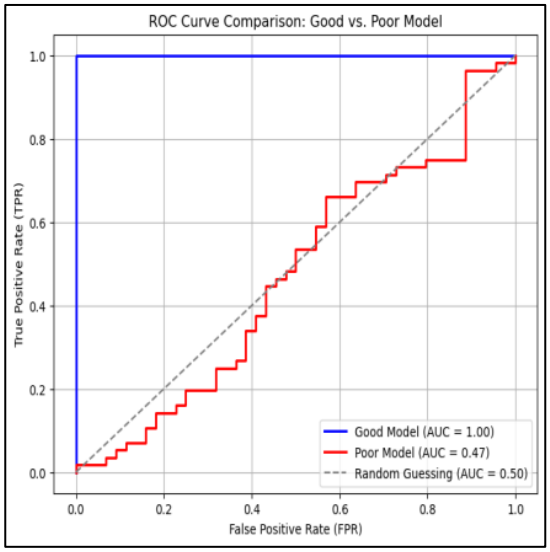


Fig 2.8 ROC Curve

3.13 Random Forest

Random Forest is a powerful and versatile machine learning algorithm that belongs to the ensemble learning family. It's used for both classification and regression tasks. Random Forest is a powerful, non-parametric ensemble learning algorithm that is mostly used for classification and regression. It works by creating a large number of decision trees and aggregating their predictions to get a final output. This methodology takes advantage of the bootstrap aggregation (bagging) technique and random feature selection, resulting in lower variance and better generalization.

Model Hypothesis

Unlike parametric models (e.g., logistic regression), Random Forest does not rely on a closed-form hypothesis function. Instead, its predictive hypothesis is defined as an ensemble of base learners (decision trees).

Classification Hypothesis

$$\hat{Y} = mode\{h_1(X), h_2(X), \dots, h_T(X)\}$$

Where

- $h_T(X)$: Prediction from the t^{th} decision tree,
- T: Total number of trees in the forest,
- $\hat{Y}$: Final predicted value or class.

Each individual tree is trained on a **bootstrap sample** from the training data, and at each split, a random subset of predictors is considered, enhancing model diversity.

Statistical Strengths

- **Non-linearity** - Captures complex interactions and non-linear relationships without explicit specification.
- **Robustness** - Reduces overfitting through averaging, improving out-of-sample performance.

- **Feature Importance -** Offers variable importance measures, aiding in model interpretation and feature selection.
- **Scalability -** Well-suited for large datasets and high-dimensional predictor spaces.

Limitations

- **Interpretability -** While more accurate than single decision trees, the model sacrifices transparency, often being treated as a "black box."
- **Computational Cost -** Training a large number of trees can be computationally intensive, particularly for high-dimensional data.
- **Sensitivity to Imbalanced Data -** Requires strategies such as resampling (e.g., SMOTE) or class weighting to handle imbalanced classification tasks effectively.

Random Forest is especially well-suited for situations that value prediction accuracy over interpretability. Its robustness to noise and capacity to handle mixed-type data make it a common choice in biological and epidemiological studies, such as predicting the presence of Urinary Tract Infection (UTI) based on clinical and behavioral characteristics.

Core Concept

- Random Forest builds multiple decision trees and combines their predictions to achieve a more accurate and stable result.
- It leverages the "wisdom of the crowd" principle, where the aggregated predictions of many individual trees are generally more reliable than the prediction of a single tree.

Key Features

Ensemble Learning - It creates a "forest" of decision trees. Each tree is trained on a random subset of the data.

Random Subsampling by **Bagging (Bootstrap Aggregating) -** Each tree is trained on a random sample of the training data with replacement. This means some data points are used multiple times, and some are not used at all.

Feature Randomness - When building each tree, only a random subset of features is considered at each node for splitting. This adds further diversity to the trees.

Voting (Classification) - For classification, the Random Forest aggregates the predictions of all trees by majority voting. The class that receives the most votes is the final prediction.

Averaging (Regression) - For regression, the Random Forest averages the predictions of all trees to produce the final prediction.

Advantages

- **High Accuracy -** Generally provides high prediction accuracy.
- **Robustness to Overfitting -** The ensemble approach and random feature selection reduce overfitting.
- **Handles High-Dimensional Data -** Can handle datasets with a large number of features.
- **Feature Importance** - Provides a measure of feature importance, indicating which features are most influential in the predictions.
- **Handles Missing Values -** Can handle missing values in the data.
- **Versatility** - Works well for both classification and regression tasks.

How It Works (Simplified)

1. **Bootstrap Sampling -** Create multiple random subsets of the training data with replacement.
2. **Tree Building -**For each subset, build a decision tree. At each node, select a random subset of features and find the best split.
3. **Prediction -** For a new data point, pass it through all the trees in the forest.
4. **Aggregation**
    - For classification, take the majority vote of the trees' predictions.
    - For regression, take the average of the trees' predictions.

Parameters

Number of Trees (n_estimators) - The number of trees in the forest. More trees generally improve accuracy but increase computation time.

Maximum Depth of Trees (max_depth) - Limits the depth of each tree, controlling overfitting.

Minimum Samples Split (min_samples_split) - The minimum number of samples required to split an internal node.

Minimum Samples Leaf (min_samples_leaf) – The minimum number of samples required [1] to be at a leaf node.

Maximum Features (max_features) - The number of features to consider when looking for the best split.

Random Forest is a widely used and effective algorithm in machine learning, particularly when dealing with complex datasets and the need for accurate predictions.

Random Forest is a versatile machine learning method with several real-world applications. In healthcare, it predicts illness outcomes such as cancer and diabetes by evaluating patient data and categorizing individuals into risk groups. In finance, it is used to detect fraud by recognizing questionable transactions based on past behavior. Random Forest analyzes customer data to anticipate which customers will depart, allowing firms to implement proactive retention efforts. In e-commerce, it fuels recommendation systems that suggest things based on client behavior.

The technique is also used in stock market forecasting, which involves evaluating historical data to predict market movements. In agriculture, Random Forest forecasts crop output by analyzing environmental parameters like as weather and soil characteristics. The capacity of Random Forest to categorize objects and patterns is also useful for image classification in computer vision problems. Finally, it is utilized in environmental monitoring to forecast air quality and pollutant levels

3.14 XG-Boost

XG-Boost (eXtreme Gradient Boosting) is a highly efficient and popular machine learning algorithm, particularly known for its speed and performance. It's a refined implementation of the gradient boosting framework, designed for both classification and regression problems. XG-Boost is a powerful and scalable implementation of gradient-boosted decision trees that is commonly used for structured (tabular) data prediction applications. It belongs to the boosting family of ensemble methods, which combines weak learners (usually shallow trees) in a sequential fashion to produce a robust predictive model. XG-Boost is designed for performance and efficiency, with speed and accuracy tuned, making it a top choice for machine learning contests and real-world applications.

Statistical Foundation and Hypothesis

In **gradient boosting**, the model is built additively

$$\widehat{Y}_i = \sum_{t=1}^{T} f_t(x_i)\,, f_t \in \mathcal{F}$$

Where:

- $\widehat{Y}_i$: Predicted output for observation iii,
- f_t: A regression tree in the functional space $\mathcal{F}$,
- T: Number of boosting rounds (trees),
- x_i: Feature vector for observation i.

Each new tree f_t is trained to predict the **residual errors** (gradients) of the previous ensemble's prediction with respect to a **differentiable loss function**, such as logistic loss for classification or squared error for regression.

Key Advantages

High Predictive Accuracy - Consistently outperforms traditional models in structured data problems. **Regularization -** L1 and L2 regularization built into the objective to reduce over-

fitting. **Handling Missing Values -** Automatically learns the best path for missing values during training. **Parallelization -** Supports distributed and parallel computation, enabling scalability. **Feature Importance -** Provides insightful feature importance metrics via gain, cover, and frequency.

Limitations

- **Model Interpretability:** As an ensemble of many trees, XGBoost is considered less transparent than linear models.
- **Hyper-parameter Tuning:** Requires careful tuning of learning rate, depth, regularization terms, and boosting rounds.
- **Computational Load:** Though optimized, training can be resource-intensive for very large datasets.

XG-Boost is especially well suited for medical and health analytics, such as forecasting the likelihood of urinary tract infection (UTI), where data might be noisy, high-dimensional, and feature relationships are complex. Its ability to handle class imbalances (via built-in class weighting and scale_pos_weight) makes it particularly useful in healthcare categorization challenges.

Core Concept

XG-Boost, like other gradient boosting algorithms, works by sequentially building decision trees. However, it focuses on optimizing the boosting process to achieve superior performance. It emphasizes speed and accuracy through various techniques.

Key Features

Gradient Boosting – XG-Boost builds trees in a sequential manner, where each new tree aims to correct the errors made by the previous trees. It uses gradient descent to minimize a loss function, guiding the model towards better predictions.

Regularization – XG-Boost incorporates L1 and L2 regularization to prevent overfitting. This helps to create more generalized models that perform well on unseen data.

Handling Sparse Data- XG-Boost efficiently handles sparse data, which is common in many real-world datasets, by automatically learning the best direction for missing values.

Parallel Processing – XG-Boost supports parallel processing, enabling faster training times, especially for large datasets.

Tree Pruning – XG-Boost uses a technique called "pruning" to remove unnecessary branches from trees, further preventing overfitting.

Weighted Quantile Sketch - This algorithm helps XG-Boost to handle very large datasets efficiently.

System Optimization – XG-Boost is engineered for computational speed and memory efficiency.

The performance of XG-Boost

Initial Prediction - The model starts with an initial prediction. **Error Calculation -** It calculates the errors (residuals) between the predictions and the actual values.

Tree Building - A new decision tree is built to predict the errors. **Model Update -** The predictions are updated by adding the predictions of the new tree, weighted by a learning rate. **Iteration -** Steps 2-4 are repeated until a stopping criterion is met. **Final Prediction -** The final prediction is the sum of the predictions from all the trees.

The XG-Boost Is Popular Classification method

High Performance - It consistently delivers state-of-the-art results in various machine learning competitions and real-world applications. **Speed and Efficiency -** Its optimized implementation makes it fast and efficient, even with large datasets. **Robustness** - Its regularization techniques make it robust to overfitting. **Flexibility** - it supports various objective functions and evaluation metrics, making it adaptable to different tasks. In summary, XGBoost is a powerful gradient boosting algorithm that combines accuracy, speed, and robustness, making it a valuable tool for machine learning practitioners.

3.15 Neural network

A **neural network** is a machine learning model inspired by the human brain, designed to recognize patterns and make predictions. It consists of layers of interconnected **neurons (or nodes)** that process input data and learn from it. A neural network (NN) is a computational model based on the structure and function of the human brain. It is commonly utilized in both supervised and unsupervised learning tasks, especially when dealing with non-linear relationships and complicated patterns. Neural networks are particularly effective for classification, regression, and feature extraction in fields such as image processing, natural language processing, and biological signal analysis.

Structure of a Neural Network

A typical **feedforward neural network (FNN)** consists of

- **Input Layer:** Receives the raw features X = $(x_1 + x_2 + \cdots + x_n)$,
- **Hidden Layers:** One or more layers of neurons that apply transformations using weights and activation functions,
- **Output Layer:** Produces the final output $\hat{Y}$ either as a class label (classification) or continuous value (regression).

Each **neuron** computes a weighted sum of inputs and passes it through a **non-linear activation function** (e.g., sigmoid, ReLU, tanh).

Mathematical Formulation

For a neuron in layer l, the output is computed as:

$$a_j^{(l)} = \phi\left(\sum_i \omega_{ji}^{(l)} a_i^{(l-1)} + b_j^{(l)}\right)$$

Where:

- $a_i^{(l)}$: Activation (output) of the j^{th} neuron in layer l,

- $\omega_{ji}^{(l)}$): Weight from neuron i in layer $l-1$ to neuron j in layer l,
- $b_j^{(l)}$: Bias term,
- ϕ: Activation function (e.g., ReLU, sigmoid).

The final predicted output $\hat{Y}$ is computed after forward propagation through all layers.

Learning Process (Backpropagation)

The network uses a **loss function** (e.g., cross-entropy for classification, MSE for regression) to measure prediction error. Through back-propagation, the gradient of the loss with respect to each parameter is computed using the chain rule. **Gradient descent** (or its variants like Adam) updates the weights to minimize the loss.

Strengths of Neural Networks

- **Non-linearity:** Can model complex, non-linear decision boundaries.
- **Feature Learning:** Automatically learns feature representations without manual feature engineering.
- **Scalability:** Suitable for large-scale datasets and high-dimensional data.

Limitations

Black-box nature - Interpretability is limited compared to simpler models (like logistic regression).

Data requirements - Requires a large amount of labeled data to perform well.

Over-fitting risk - Prone to over-fitting without regularization or dropout, especially with small datasets.

Hyper-parameter sensitivity -Requires careful tuning (e.g., number of layers, neurons, learning rate).

Neural networks are rapidly being used in clinical decision support to predict UTI occurrence based on patient history, symptoms, and laboratory results. When paired with regularization and interpretability tools (such as SHAP and LIME), they can be effective and dependable in healthcare analytics.

Key Components of a Neural Network

1. **Input Layer** - The first layer that receives input features (e.g., variables from your dataset). Each node represents a feature.

2. **Hidden Layers** - Intermediate layers that perform computations. Each neuron applies a **weight**, **bias**, and **activation function** to transform the input.

3. **Output Layer** - Produces the final prediction (e.g., classification labels or regression values). For binary classification (e.g., predicting UTI presence: Yes/No), it usually has one neuron with a sigmoid activation function.

How Neural Networks Learn

Each neuron applies a weighted sum of inputs and passes it through an activation function. The network adjusts the weights using back propagation, which minimizes the error using an optimization algorithm like **Gradient Descent**. It uses a **loss function** to measure the difference between predicted and actual values.

Types of Neural Networks for Your Project

Since you are working with binary logistic regression and clustering, you can use

1. **Feedforward Neural Network (FNN)** – A basic model suitable for classification tasks like binary logistic regression.

2. **Convolutional Neural Network (CNN)** – If your data includes images or spatial features.

3. **Recurrent Neural Network (RNN)** – For sequential data (like time series).

4. **Autoencoders** – If you need unsupervised learning for clustering.

Aspect	Neural Network	Statistical Model (Logistic Regression, K-Means, etc.)
Interpretability	Low (black-box)	High (coefficients explain relationships)
Assumptions	Few (flexible, handles non-linearity)	Many (e.g., linearity, independence)
Accuracy	Often higher with complex patterns	Good for simple relationships
Overfitting	Possible, needs regularization	Less prone, more stable
Feature Selection	Automatic via hidden layers	Requires manual selection/testing

Application

Neural networks are widely used in a variety of industries due to their capacity to learn complicated patterns from data. In healthcare, they are used to diagnose diseases based on medical imaging and forecast patient outcomes. Neural networks enable autonomous vehicles perceive things and make real-time judgments to ensure safe navigation. They provide speech recognition for virtual assistants such as Siri and Google Assistant. Neural networks drive image and video identification in security systems, as well as automatic photo tagging on social networking platforms. Machine translation, sentiment analysis, and chatbots are all made possible by natural language processing.

Neural networks are used in financial services for stock price prediction, credit risk assessment, and fraud detection. Recommendation systems on sites such as Amazon and Netflix use neural networks to provide individualized recommendations. In gaming, they are used to create intelligent NPCs and train AI in complex games. Finally, neural networks can increase energy efficiency in smart grids by forecasting energy demand and supply. These applications demonstrate the transformational power of neural networks across different industries.

3.16 Unsupervised Learning Methods

Unsupervised Learning is a type of machine learning technique that works with unlabeled data to uncover underlying patterns, structures, or distributions within it. Unlike supervised learning, there are no predefined target labels, and the goal is to explore the data's intrinsic structure without being guided by specific output expectations.

Key Goals of Unsupervised Learning

1. Clustering - Grouping data points into distinct clusters or categories based on similarity.
2. Dimensionality Reduction - Reducing the number of features while retaining the important information, often for visualization or preprocessing.

Popular Unsupervised Learning Methods

Clustering Algorithms

K-Means Clustering - Partitions data into k clusters by minimizing the within-cluster variance.

$$min(C) = \sum_{i=1}^{k} \sum_{x_j \in C_i} \left\| x_j - \mu_i \right\|^2$$

Where C_i is the set of points in cluster i, and μ_i is the centroid of cluster i.

Agglomerative Clustering - is a type of hierarchical clustering algorithm that builds a hierarchy of clusters by iteratively merging the closest pairs of clusters. It is particularly useful for creating a dendrogram, which visually represents the data's hierarchical structure, allowing analysts to observe the optimal number of clusters and their relationships.

Applications of Unsupervised Learning

1. Market Segmentation - In business, clustering can help segment customers into groups with similar purchasing behaviors.

2. Medical Diagnostics - Clustering and anomaly detection can assist in identifying patterns in medical imaging or patient data, such as clustering patients with similar symptoms for disease prediction.

Advantages and Limitations

Advantages

No need for labeled data, making it suitable for exploratory analysis. Can discover hidden patterns that supervised methods may miss. Useful for feature extraction and data preprocessing.

Limitations

Harder to evaluate the quality of the results since there are no ground truth labels. May require domain knowledge to interpret results meaningfully. Sensitive to the choice of hyper-parameters (e.g., the number of clusters in K-means).

Application

Unsupervised learning and clustering have various real-world applications because they can uncover hidden patterns and structures in data that lacks labeled outputs. Clustering is a marketing technique that divides clients into groups with similar tendencies, allowing for more tailored promotions. In healthcare, it aids in the identification of patient subgroups with similar symptoms or diseases, allowing for better treatment planning. Unsupervised learning is used in finance to detect anomalies such as fraudulent transactions or odd financial activity. Clustering is used in social networks to discover groupings of users who share common interests or habits. Natural language processing (NLP) uses unsupervised learning to group related documents, hence improving search and recommendation systems. Retailers utilize clustering to optimize product placement by recognizing client purchase trends. Image compression and compression techniques use unsupervised learning to reduce data storage requirements while maintaining quality. Genomics employs clustering to group similar genetic patterns, which aids in medication development. Clustering is useful in customer service for analyzing support tickets and identifying common issues. These applications show how unsupervised learning and clustering may extract useful insights from unlabeled data, resulting in increased efficiency and creativity across industries.

3.17 Software Tools used for Analysis

Minitab

Minitab 18 is versatile statistical software that improves project processes by offering capabilities for data analysis, visualization, and statistical inference throughout the project lifecycle. It facilitates data-driven decision-making with tools such as exploratory data analysis, hypothesis testing, regression analysis, Design of Experiments, Statistical Process Control, and Measurement Systems Analysis. The "Assistant" feature assists users in selecting and interpreting appropriate statistical tools, making Minitab 18 an invaluable tool for carrying out the statistical aspects of methodologies such as DMAIC or Lean to understand data, identify root causes, optimize processes, and ensure long-term improvements.

Python

Python is a high-level, interpretable, general-purpose programming language known for its readability and adaptability. Its design philosophy emphasizes code readability through the use of extensive indentation. Python is an interpreted language that executes code line by line, making development and debugging easier. Its general-purpose nature allows it to be used in a variety of domains, including web development (with frameworks like Django and Flask), data science (with libraries like Pandas, NumPy, and Scikit-learn), machine learning, artificial intelligence, scientific computing, scripting, automation, and even game development.

Python 3.12 in Spyder and IDLE provide separate development environments for developing and running Python code. Spyder, a powerful IDE, offers a complete interface that includes a multi-pane structure for code editing, variable exploration, debugging, and an IPython console, making it ideal for data research and larger projects. IDLE, on the other hand, is a simpler, lightweight integrated development environment that ships with Python. It includes a rudimentary text editor with syntax highlighting, a shell for interactive execution, and a debugger, making it perfect for beginners and rapid scripting tasks. While both can run Python 3.12, Spyder has a more feature-rich and organized workflow, whilst IDLE provides a simpler and less complex environment.

CHAPTER 4

ANALYSIS AND INTERPRETATION

The analysis and interpretation of raw data yield useful insights. It displays findings objectively utilizing tables, graphs, and statistics. It then evaluates these findings, linking them to research issues and existing literature. Patterns, correlations, and unexpected findings are discussed, along with their significance and limitations. This chapter connects the data to the final findings, providing preliminary insights into the project's results.

4.1 EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis is an iterative process of discovering and comprehending data in order to establish the framework for more formal analysis and modeling. It entails summarizing a dataset's essential properties, which is frequently accomplished through visual means. The fundamental purpose is to comprehend the data structure, discover trends, detect abnormalities, and develop hypotheses. EDA assists in making educated decisions on data cleaning, transformation, and the selection of relevant statistical models or machine learning algorithms. Techniques Used in EDA are

Handling Missing Values - Identifying and addressing missing data through visualization (e.g., heat-maps) and determining imputation or removal procedures.

Univariate analysis - involves analyzing individual variables to understand their distribution. This can be done graphically (histograms, box plots) or non-graphically (calculations of mean, median, mode, standard deviation, and frequency distributions).

Bivariate analysis - examines the link between two variables. Techniques include scatter plots (for two numerical variables), bar charts (for one categorical and one numerical variable), and box plots (for a categorical and a numerical variable). Correlation analysis can also be applied to numerical variables.

Multivariate Analysis - Investigating links among three or more variables. This may use scatter plot matrices, heat-maps (to illustrate correlations), parallel coordinates, and dimensionality reduction techniques such as Principal Component Analysis (PCA) for visualization.

4.1.1 Classification and Tabulation of data

Classification of data is the process of categorizing data according to its features, sensitivity, and relevance. The fundamental purpose is to make data easier to understand, manage, secure, and efficiently use. We can classify data based on data type, by content, by source, by sensitivity, and usage of the data.

The specific method of classification is determined by the environment and aims of the organization or investigation. Effective data classification is an essential step toward data governance, security, compliance, and effective data management.

Category - This column groups the numerical variables and categorical variables based on the type of information they represent. This helps in understanding the different aspects of the study participants being measured numerically.

Numerical/ Categorical Variables - This column lists the specific variables that were recorded within each category. Numerical variables are those that can be measured and have a quantitative value (e.g., age in years, BMI as a continuous value, frequency of commode usage). And Categorical Variables such as education, Occupation and etc.

No. of Variables - This column indicates the count of variables within each category.

The following tables represent the classification of the variables based on the features of the variable using statistical analysis for dataset.

Table 4.1 Classification of the numerical variables based on the features

Category	**Numerical Variables**	**No. of Variables**
Demographic Information	Age	1
Health Indicators	BMI	1
Lifestyle Factors	Indian_commode, Western_commode	2

Table 4.2 Classification of the categorical variables based on the features

Category	**Categorical Variables**	**No. of Variables**
Demographic Information	Age_Interval, Education, Occupation, M_Status	4
Health Indicators	BMI_Categorised, Diabetes, Hypertension, Coronary_heart_disease, Chronic_pulmo_lung_disease, Back_vertebral_spinal_pain, Any_Neurological_disorder, Parkinsons, Multiple_sclerosis, Brain_tumor, Stroke, Any_other, Persistent_severe_lower_pain, Other_chronic_disease	14
Lifestyle Factors	Simplified_COMMODE, Coffee, Tea, Softdrinks, Butter_milk, Fruit_Juices, Liquid_consumption	7
Behavioral Factors	empty_bladded_frequently, Fluid_restriction, Empty_bladder_prior_go_out, Lookout_for_toilet, Pad_protection_to_management	5
Reproductive Health	No_of_children, Concatinate_Delivary, Menopause_attained, Hysterectomy_done, Ovaries_removed	5
Target Variable	Urinary_Track_Infection	1

Interpretation of classification of data and tabulation of data

The dataset is divided into six main categories to allow for a more structured investigation of the factors that influence urinary tract infections (UTIs). Demographic information includes personal data such as age (numerical) and education, occupation, and marital status (categorical). Health indicators focus on an individual's health state and pre-existing illnesses, which can be numerical (such as BMI) or categorical (such as diabetes or

hypertension). Lifestyle Factors are categorical variables that describe daily routines and preferences, such as coffee and tea consumption. Behavioral Factors look at bladder function and hygiene practices, such as commode type and frequency of emptying the bladder.

Reproductive Health comprises characteristics related to reproductive history, particularly for women, such as Number of Children (numerical) and Menopause Attained (categorical). Finally, the Target Variable shows whether a UTI is present or not (categorical). Variables are grouped into two types: numerical variables, which represent measurable values such as age and BMI, and categorical variables, which describe attributes or groupings such as education level and occupation. This category is important because it allows for a systematic approach to investigating the links between various types of variables and the prevalence of UTIs, allowing for a more structured and meaningful investigation of potential influencing factors.

This categorization is strategic because it enables targeted statistical analyses within each domain. For example, in Demographic Information, we can use correlation to investigate continuous relationships with age and chi-square tests to compare UTI prevalence across different educational or occupational groups.

The detailed breakdown within Health Indicators allows for the examination of specific co-morbidities as potential risk factors. Similarly, the segmentation of Lifestyle and Behavioral Factors facilitates the analysis of changeable habits and behaviors. The establishment of a separate Reproductive Health category acknowledges the unique physiological and hormonal aspects that may increase UTI risk in women. Ultimately, this hierarchical classification intends to promote a comprehensive knowledge of the diverse etiology of UTIs and to identify possible targets for prophylactic or therapeutic strategies.

Data tabulation integrates complicated information into simple tables, making it easier to interpret and compare directly. This organized style emphasizes critical information and exposes trends, making data available for statistical analysis and visualization. Tabulation saves space and time by presenting data clearly, allowing for more informed decision-making across a variety of sectors. It is a critical step in converting raw data into meaningful insights.

4.1.2 Diagrammatic representation categorical variables

Diagrammatic representation is the use of visual tools such as charts, graphs, and diagrams to show and analyze data in a more natural and accessible manner than tables or text, allowing for rapid discovery of patterns, trends, links, and comparisons. These graphics improve knowledge, facilitate transmission of discoveries to larger audiences, aid in exploratory data analysis, clearly describe big datasets, and promote engagement. In a research like yours on UTIs, these representations would effectively demonstrate prevalence across different groups as well as the distribution of critical variables, allowing for a better understanding of potential risk factors.

The following diagrammatic representations show the Clustered Bar Plots for the categorical variables with showing the number of individual have UTI.

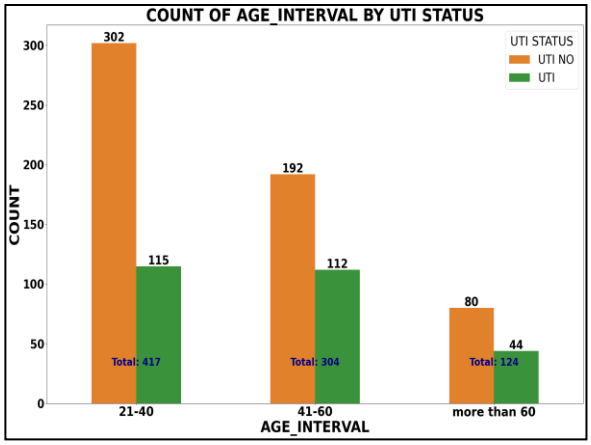


Fig 4.1 Bar plot for Age_Interval

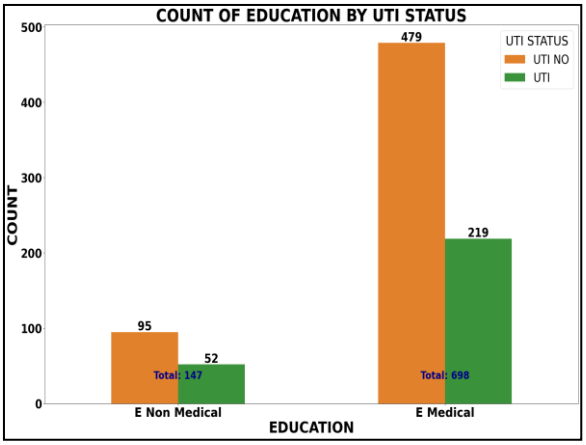


Fig 4.2 Bar plot for Education

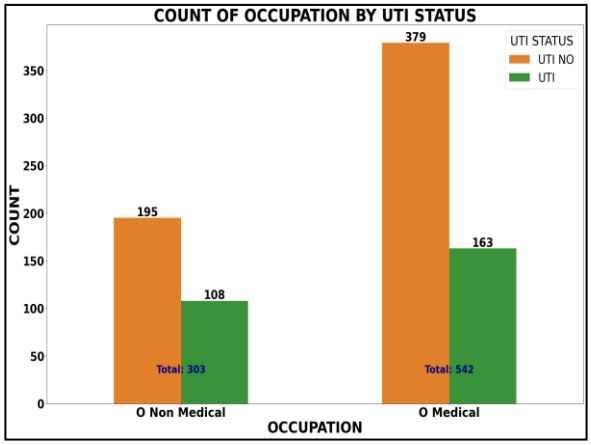


Fig 4.3 Bar plot for Occupation

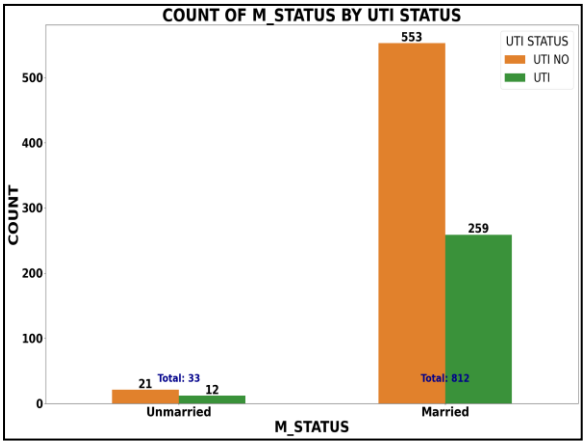


Fig 4.4 Bar plot for M_Status

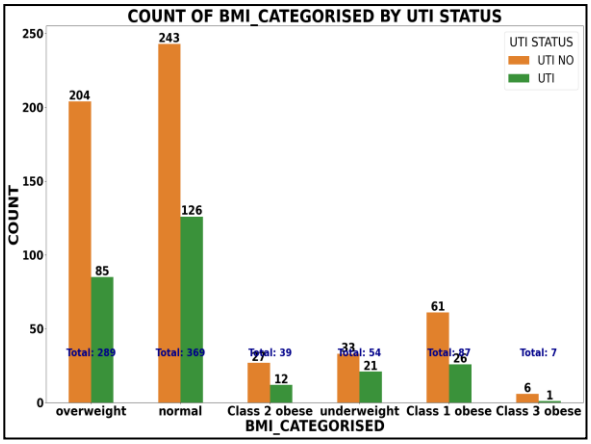


Fig 4.5 Bar plot for BMI_Categorical

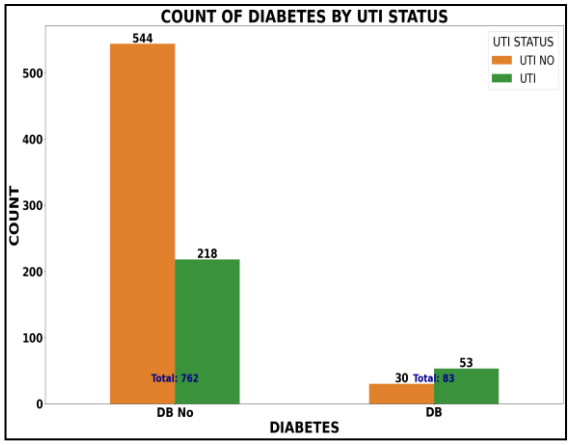


Fig 4.6 Bar plot for Diabetes

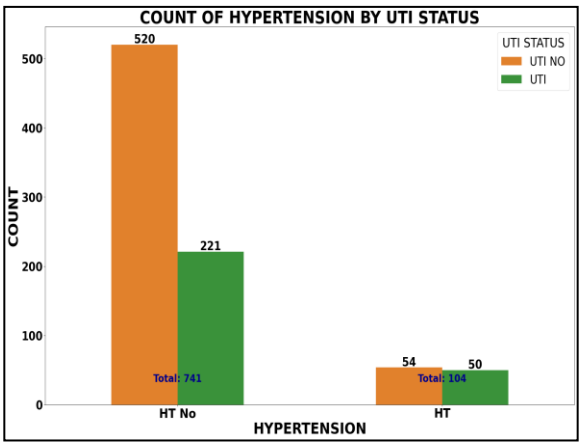


Fig 4.7 Bar plot for Hypertension

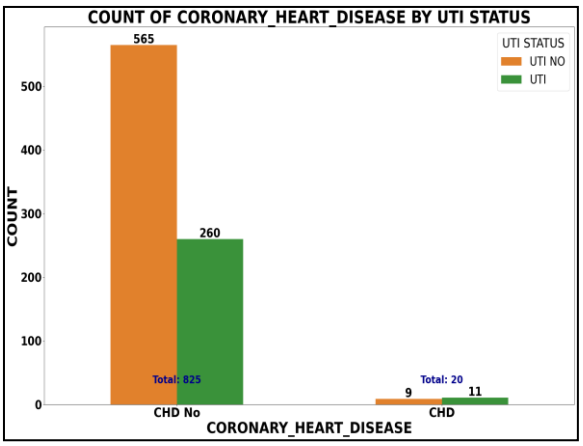


Fig 4.8 Bar plot for Coronary_Heart_Disease

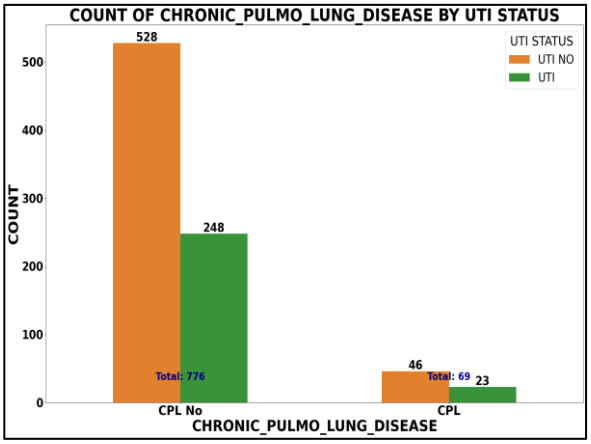


Fig 4.9 Bar plot for Chorinic_Pulmo_Lung_Disease

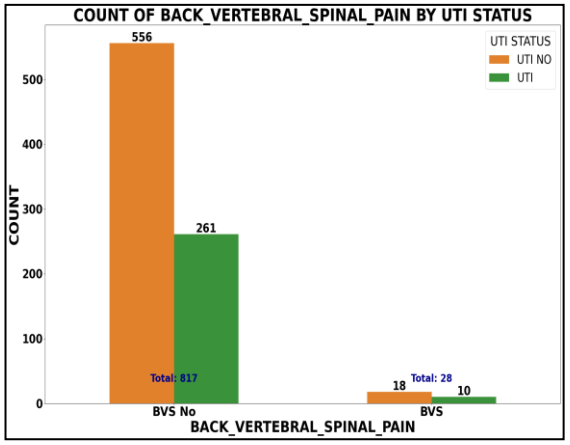


Fig 4.10 Bar plot for Back_Vertebral_Pain

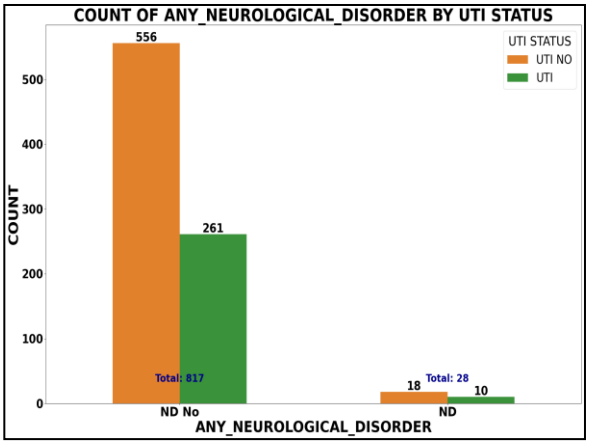


Fig 4.11 Bar plot for Any_Neurallogical_Disorder

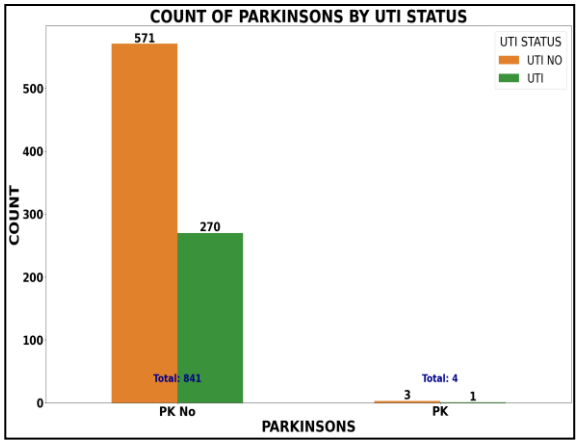


Fig 4.12 Bar plot for Parkinsons

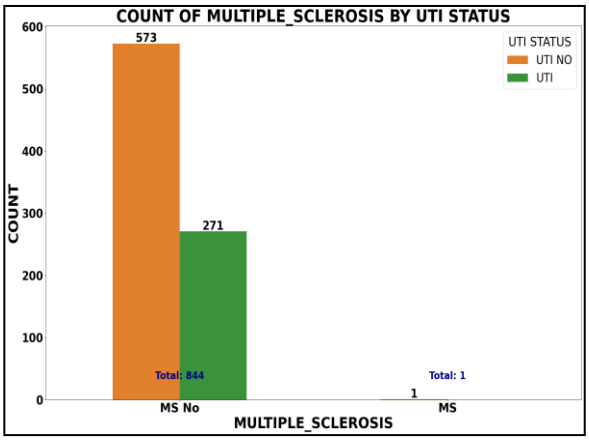


Fig 4.13 Bar plot for Multiple_Sclerosis

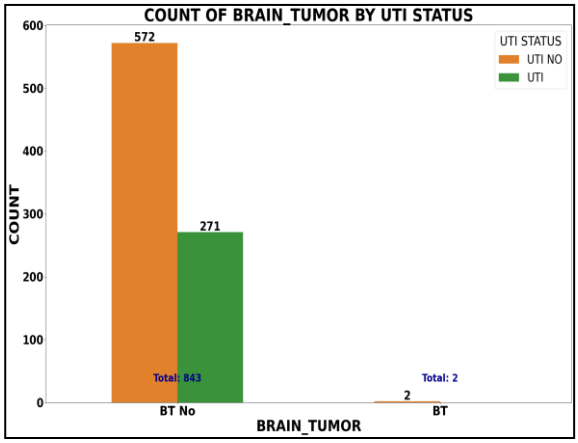


Fig 4.14 Bar plot for Brain_Tumor

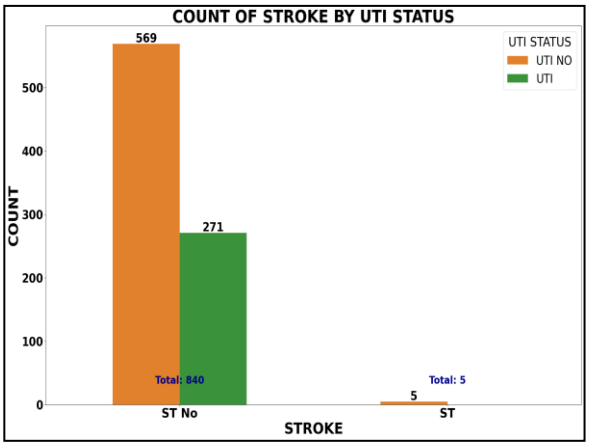


Fig 4.15 Bar plot for Stroke

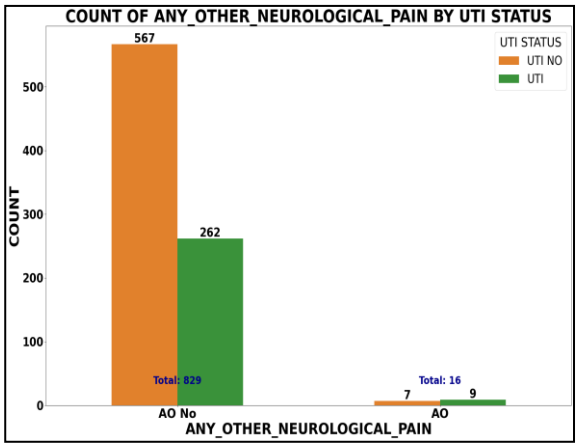
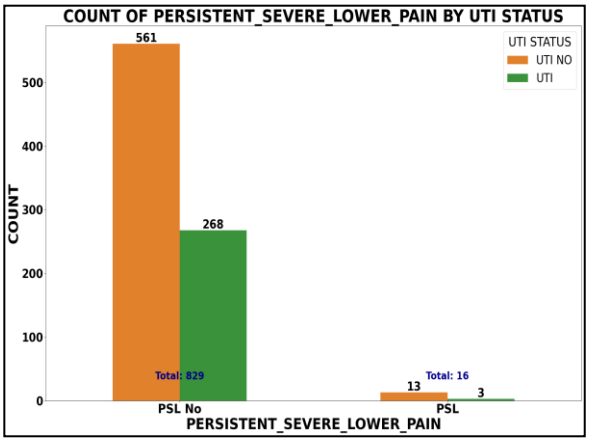


Fig 4.16 Bar plot for Any_Other_Neurological_Pain



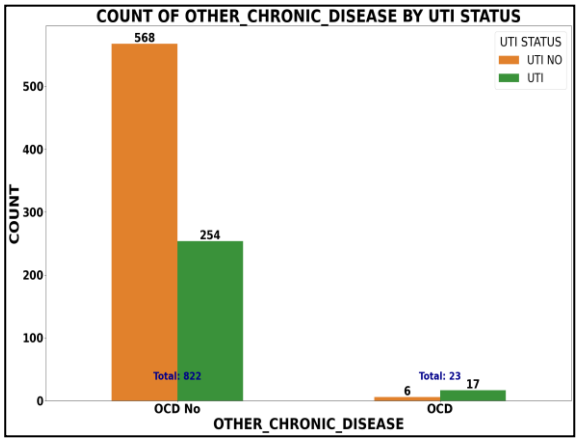
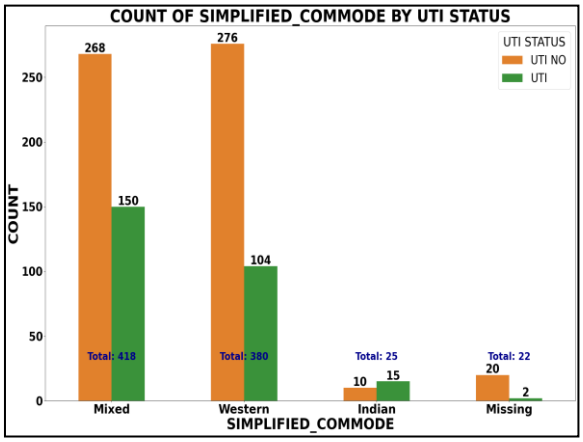


Fig 4.17 Bar plot for Persistent_Severe_Lower_Pain **Fig 4.18 Bar plot for Other_Cronic_Disease**



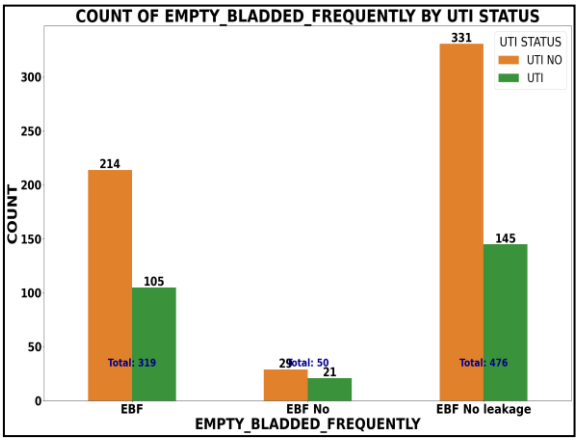
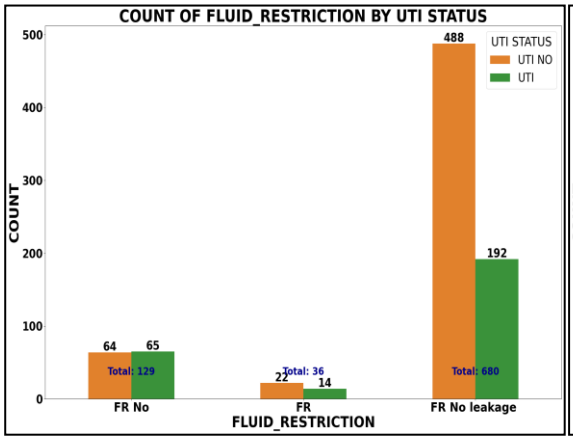


Fig 4.19 Bar plot for Simplefied_Commode **Fig 4.20 Bar plot for Empty_Bladded**



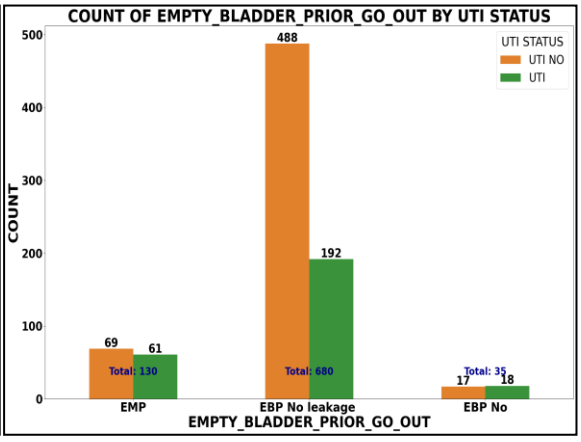


Fig 4.21 Bar plot for Fluid_Restriction **Fig 4.22 Bar plot for Empty_Bladded_Prior_Go_Out**

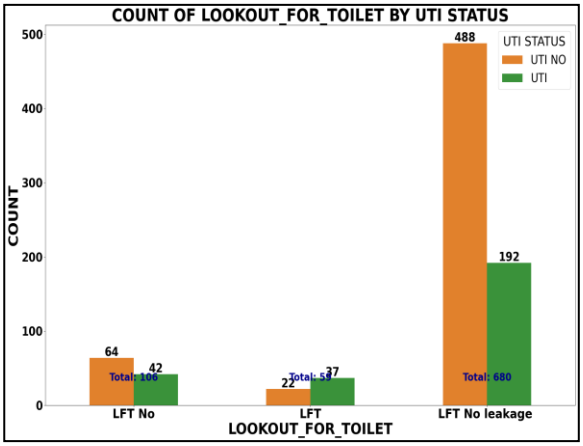


Fig 4.23 Bar plot for Lookout_For_Toilet

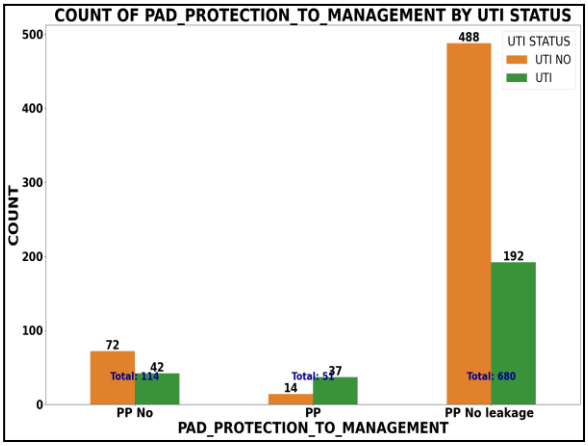


Fig 4.24 Bar plot for Pad_Protection_To_Management

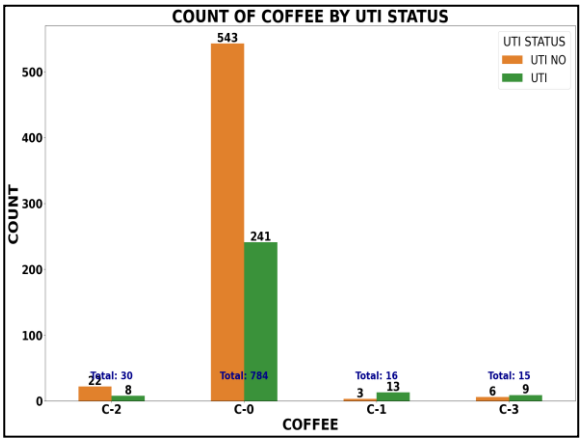


Fig 4.25 Bar plot for Coffee

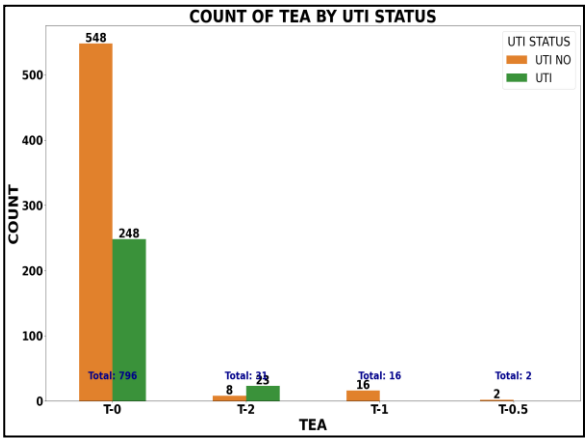


Fig 4.26 Bar plot for Tea

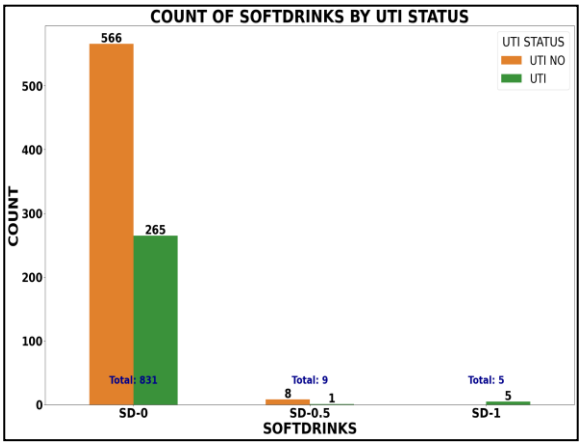


Fig 4.27 Bar plot for Softdrinks

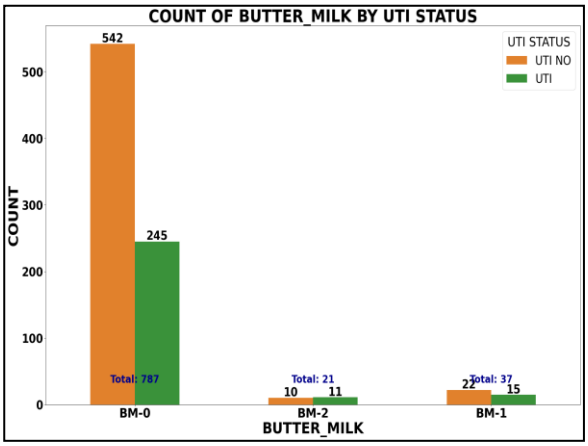
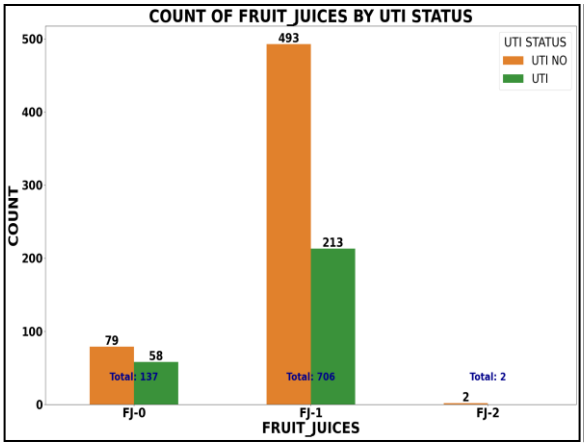


Fig 4.28 Bar plot for Butter_Milk



4.29 Bar plot for Fruit_Juices

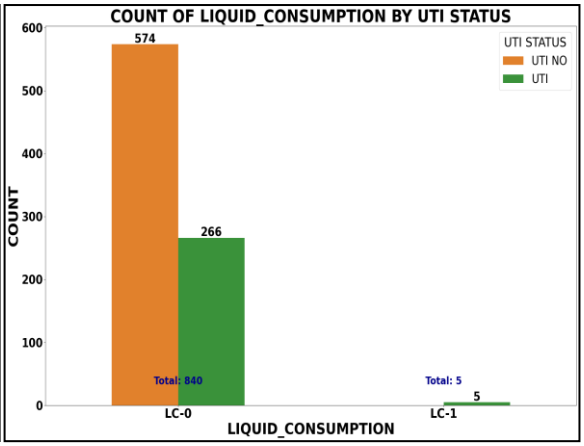
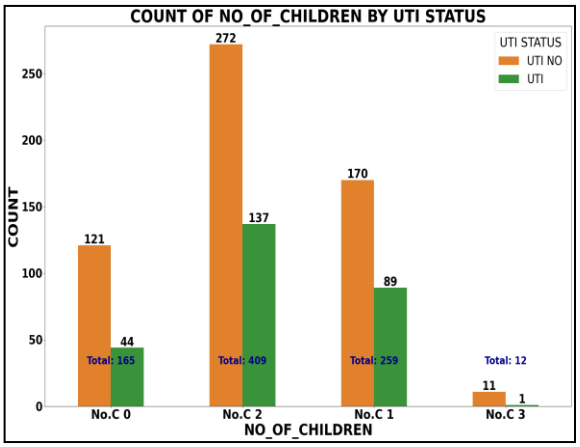


Fig 4.30 Bar plot for Liquid_Consumption



4.31 Bar plot for No_of_Children

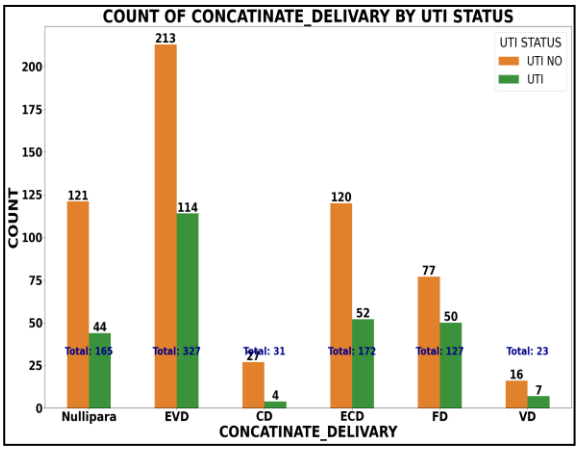
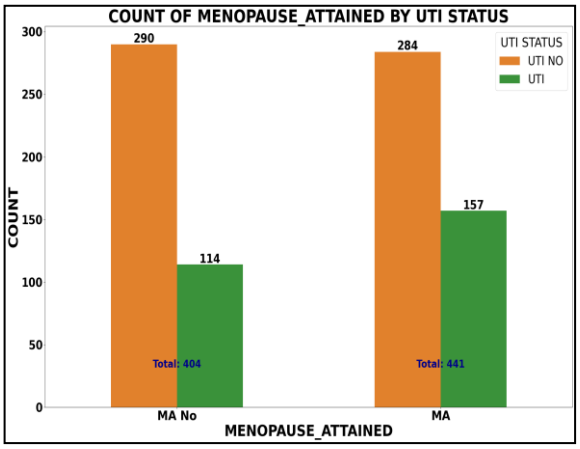


Fig 4.32 Bar plot for Concatinate_Delivary



4.33 Bar plot for Menopause_Attained

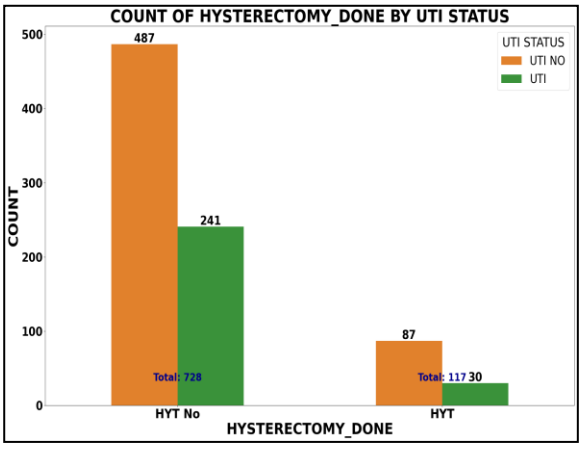
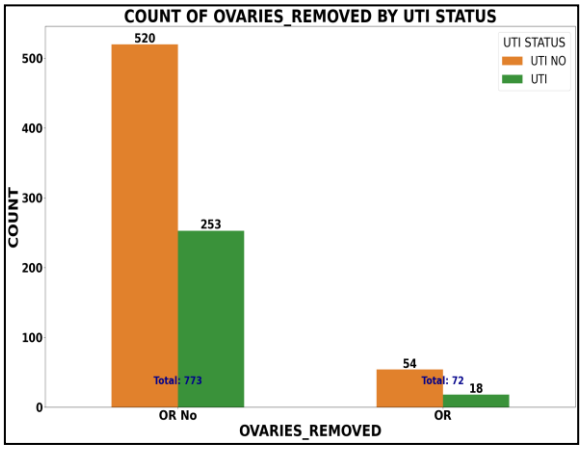


Fig 4.34 Bar plot for Hysterectomy_Done



4.35 Bar plot for Ovaries_Removed

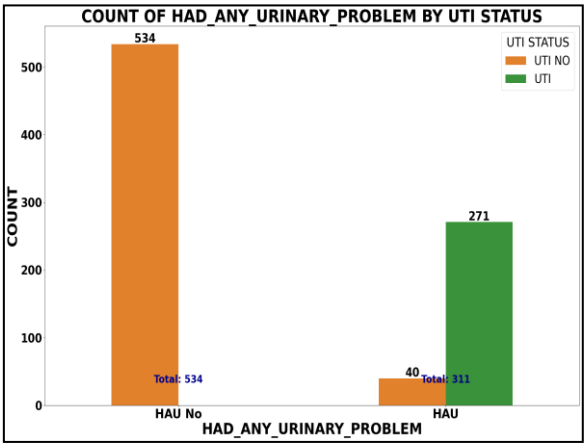


Fig 4.36 Bar plot for Had_Any_Uninary_Problem

Interpretation of clustered bar plots for all categorical variables

The 21-40 age groups have the most UTI cases, while the incidence appears to be fairly stable across all age categories. Individuals with non-medical degrees and occupations had a somewhat greater rate of UTI than those with medical credentials. BMI classifications vary, with potentially greater proportions in the Underweight and Normal weight categories. Several health issues appear to be linked to UTI. Notably, people with diabetes have a significantly greater proportion of UTIs. Hypertension and Coronary Heart Disease have greater UTI proportions, albeit the CHD group has a much smaller sample size. Other chronic conditions, such as Chronic Pulmonary Lung Disease, Back/Vertebral/Spinal Pain, Neurological Disorders, Parkinson's, Multiple Sclerosis, Brain Tumor, Stroke, Other Neurological Pain, Persistent Severe Lower Pain, and Other Chronic Diseases, have small sample sizes, making it difficult to draw firm conclusions about their relationship to UTI. Hygiene and behaviors appear to play an important role. The Indian commode type is related with a higher incidence of UTI. Not emptying the bladder frequently and not emptying it before going out result in greater UTI proportions. Pad usage for management is connected with an increased risk of UTI. Coffee and tea use have increased UTI proportions in certain categories. Interestingly, not restricting fluids appears to be connected with a higher risk of UTI. In terms of reproductive history, reaching menopause is related with a larger proportion of UTI, whereas having a hysterectomy or ovaries removed appears to be associated with a lower proportion. Finally, a prior history of any urinary issue is strongly associated with the current UTI.

4.1.3 Descriptive Statistics

Descriptive statistics are used to describe and summarize the main features of a dataset. They provide simple summaries about the sample and the measures. Some common descriptive statistics include measures of central tendency (mean, median, mode), measures of variability or spread (range, variance, standard deviation), and measures of distribution shape (skewness, kurtosis).

Table 4.3 Descriptive statistics for numerical variables

Variable	Mean	StDev	Variance	CoefVar	Minimum	Maximum	Mode
Age	42.785	14.996	224.873	35.05	21	70	25
BMI	25.345	5.081	25.813	20.05	12.5	45.3	26
Indian_commode	13.353	15.091	227.736	113.02	0	60	0
Western_commode	22.77	14.621	213.762	64.21	0	71	25

Variable	IQR	Q1	Median	Q3	Skewness	Kurtosis	SE Mean
Age	27	28	41	55	0.26	-1.3	0.516
BMI	6.4	21.9	24.9	28.3	0.57	0.61	0.175
Indian_commode	25	0	8	25	0.72	-0.66	0.519
Western_commode	17.5	12	25	29.5	0.49	0.41	0.503

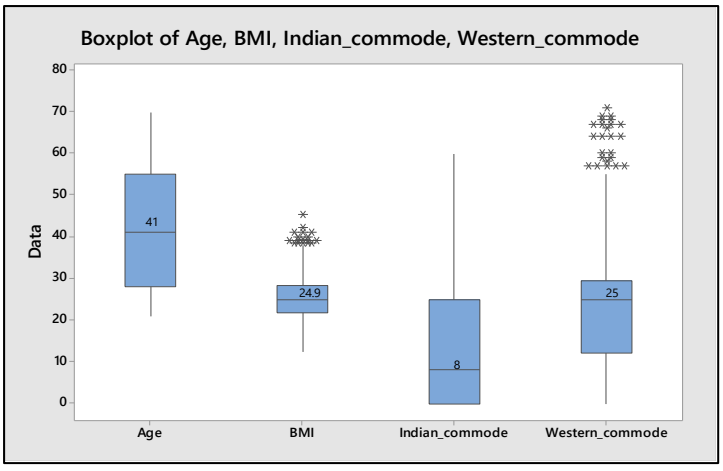


Fig 4.37 Box plot of Age, BMI, Indian_commode and Western_commode

The following is the summary plot of the numerical variables in the dataset.

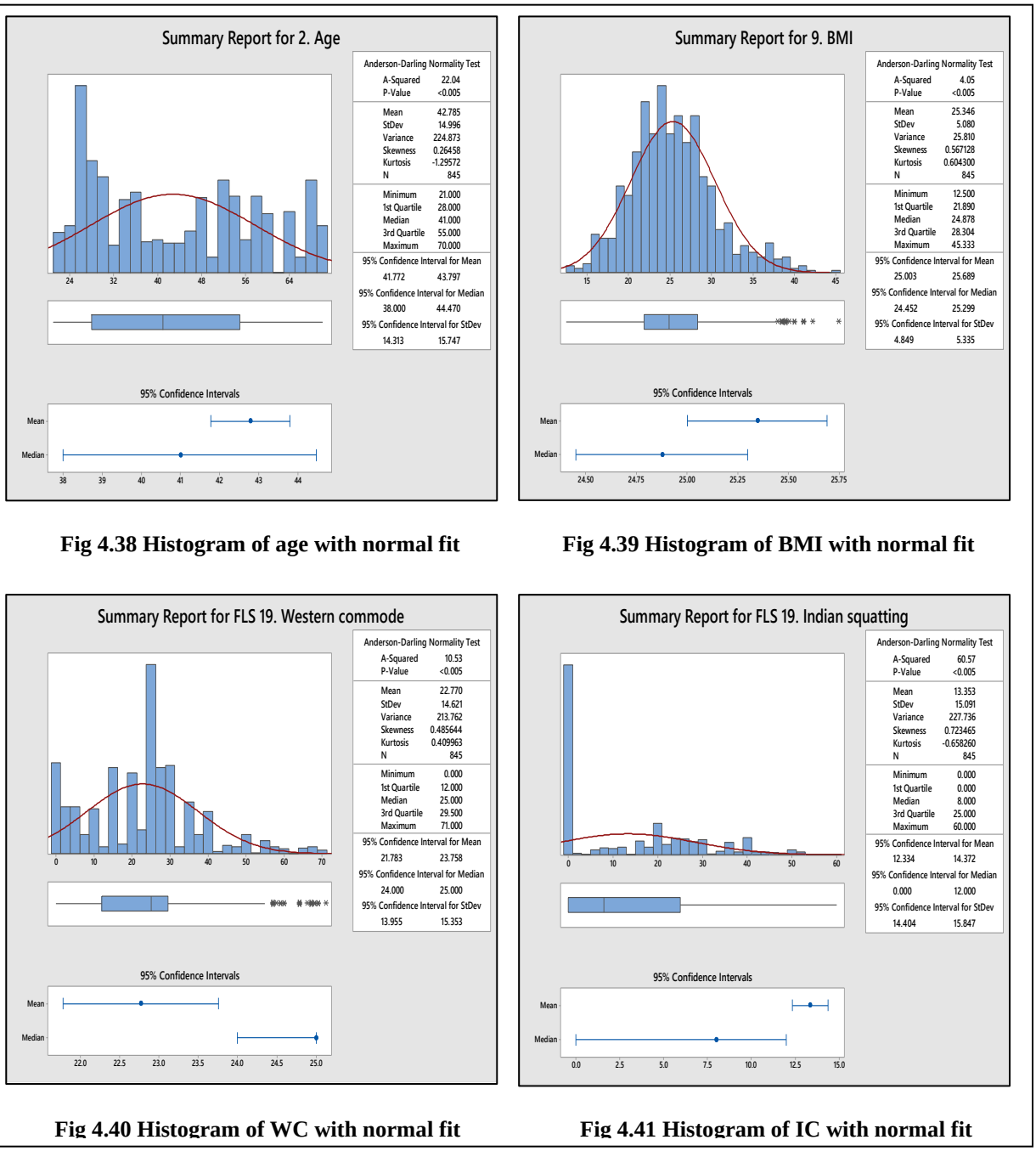


Fig 4.38 Histogram of age with normal fit

Fig 4.39 Histogram of BMI with normal fit

Fig 4.40 Histogram of WC with normal fit

Fig 4.41 Histogram of IC with normal fit

Interpretations of descriptive statistics

Age - The age variable (mean=42.79, SD=15.00) has moderate heterogeneity within the patient population. The slight positive skewness (0.26) indicates a slight trend toward older ages in the distribution, however the negative kurtosis (-1.30) implies a flatter, platykurtic distribution as compared to a normal curve. The box plot reveals a fairly symmetrical distribution with no noticeable outliers. The Anderson-Darling normality test ($p < 0.005$) contradicts the null hypothesis of a normal distribution, suggesting that age-related statistical analyses should account for the non-normality.

BMI - The Body Mass Index (BMI) variable (Mean=25.35, SD=5.08) indicates that the average BMI is overweight, with a reasonable degree of variation. The positive skewness (0.57) implies a bias toward higher BMI values, whereas the positive kurtosis (0.60) indicates a more peaked, leptokurtic distribution with heavier tails than a normal distribution. The box plot indicates the existence of outliers at higher BMI levels. The Anderson-Darling normality test ($p < 0.005$) reveals a statistically significant divergence from a normal distribution. This suggests that BMI analysis should account for non-normality and outliers.

Indian_commode - The Indian_commode variable (mean = 13.35, SD = 15.09) has a high relative variability (CV = 113.02%). The positive skewness (0.72) implies that the distribution is concentrated at lower values, with a tail spreading upwards. The negative kurtosis (-0.66) points to a flatter, platykurtic distribution. The box plot graphically verifies the skewness, with the median positioned in the lower quartiles. The Anderson-Darling normality test ($p < 0.005$) strongly rejects the assumption of normality, emphasizing the necessity for non-parametric or distribution-appropriate approaches in future analyses with this variable.

Western_commode - The Western_commode variable (mean=22.77, SD=14.62) similarly has a high relative variability (CV = 64.21%). The positive skewness (0.49) indicates a skew toward higher values. The positive kurtosis (0.41) suggests a peaked distribution with heavier tails than a normal distribution. The box plot shows outliers at higher levels. The Anderson-Darling normality test ($p < 0.005$) confirms a statistically significant deviation from normality, indicating that analysis of Western_commode should account for its non-normal distribution and the presence of outliers.

4.2 INFERENTIAL STATISTICS

Inferential statistics is a type of statistical analysis that uses data from a sample to draw inferences or make predictions about the larger population. Instead of evaluating the entire population (which is frequently difficult or impossible), inferential statistics enables us to make educated assumptions or inferences about the entire group based on a smaller, more representative selection.

Key Analytical methods

Correlation Analysis - Measuring the strength and direction of the linear relationship between two variables.

Analysis of Variance (ANOVA) - Comparing the means of two or more groups to see if there are statistically significant differences.

Confidence Intervals - Estimating a range of values within which a population parameter is likely to lie.

Hypothesis Testing - Determining if there is enough statistical evidence to reject a null hypothesis about a population parameter.

Regression Analysis - Examining the relationship between variables and predicting the value of a dependent variable based on independent variables.

Chi-Square Tests - Examining relationships between categorical variables.

The validity of inferences drawn from inferential statistics heavily relies on the **representativeness** of the sample. A biased or non-random sample can lead to inaccurate conclusions about the population. Inferential statistics is the foundation for drawing broader conclusions from restricted data. Its power stems from probability theory and a grasp of sampling distributions, which allows us to measure uncertainty and make sound generalizations. Inferential statistics offers a wide range of analytical tools, including correlation to investigate linear relationships, ANOVA to compare group means, confidence intervals to estimate population parameters, hypothesis testing to validate assumptions, regression to model and predict variable relationships, and chi-square tests to analyze categorical associations.

4.2.1 Correlation

The following plot represents the correlation heat-map which the relations between numerical variables Age, BMI, Indian commode, and Western commode

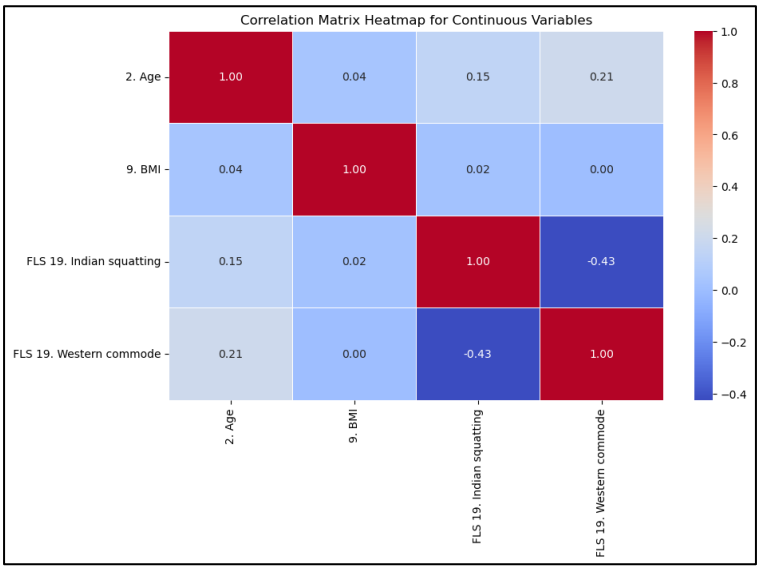


Fig 4.42 Correlation heat-map numerical variables

Interpretation of correlation

Correlation Values

Age vs. BMI (0.04) - Very weak positive correlation, Age vs. Indian squatting (0.15): Weak positive correlation, Age vs. Western commode (0.21) - Weak positive correlation, BMI vs. Indian squatting (0.02) - Very weak positive correlation, BMI vs. Western commode (0.00) - No correlation, Indian squatting vs. Western commode (-0.43) - Moderate negative correlation. Age shows weak positive connections with "Indian squatting" and "Western commode." This indicates that older people may have slightly better scores in these qualities, but the association is not strong. BMI has poor or no association with other variables. This suggests that BMI is relatively unaffected by age and the two functional limitation ratings. Indian squatting and Western commode have a moderate negative association (-0.43). This means that people who excel in "Indian squatting" often struggle with "Western commode," and vice versa. This could imply that these two qualities represent conflicting aspects of functional constraints or preferences.

4.2.2 One-Way ANOVA

The following tables show the Welch's test for One-way ANOVA and confidence interval for the numerical variable.

Table 4.4 One – Way ANOVA for numerical variables

Welch's Test				
Variables	DF	DF Den	F-Value	P-Value
Age	1	519.48	5.92	0.015
BMI	1	548.653	2.47	0.116
Indian_Commode	1	576.061	4.7	0.031
Western_Commode	1	582.486	6.64	0.01

Table 4.5 One – Way ANOVA Confidence Interval

Variables	**Urinary_Track_Infection**	**Mean**	**StDev**	**95% CI**
Age	UTI_No	41.918	14.848	(40.701, 43.135)
	UTI	44.62	15.168	(42.806, 46.434)
BMI	UTI_No	25.531	5.139	(25.110, 25.953)
	UTI	24.951	4.942	(24.360, 25.542)
Indian_Commode	UTI_No	12.606	15.484	(11.337, 13.876)
	UTI	14.934	14.121	(13.245, 16.622)
Western_Commode	UTI_No	23.625	15.036	(22.393, 24.858)
	UTI	20.959	13.549	(19.339, 22.580)

Hypothesis

- Null Hypothesis (H_0): The means of all groups are equal.
- Alternative Hypothesis (H_1): At least one group mean is different from the others.

Welch's Test Rationale - The use of Welch's ANOVA indicates that the initial assumption of equal variances across the groups (UTI vs. No UTI) was likely not met. Welch's test provides a more robust analysis under such conditions.

Variables Analyzed - The analysis examines the differences in mean values of Age, BMI, Indian_Commode usage, and Western_Commode usage between individuals with Urinary Tract Infections (UTI) and those without (No UTI).

Test Statistics and Significance

The table presents the degrees of freedom (DF, DF Den), F-values, and p-values for each variable. The p-value is the crucial indicator of statistical significance. A p-value less than the significance level (typically 0.05) suggests that there's a statistically significant difference in the means of the variable between the UTI and No UTI groups.

The following Diagrammatic representation is the Interval plot for individual numerical variables.

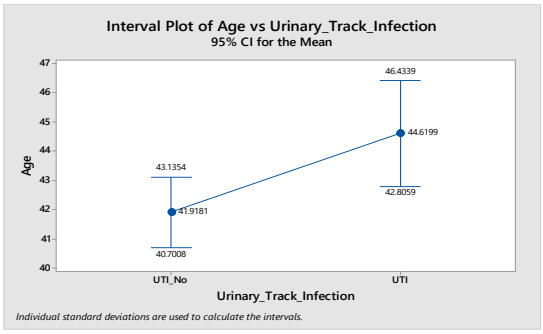


Fig 4.43 Interval plot of Age

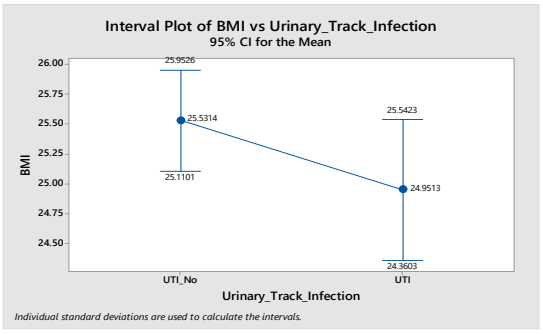


Fig 4.44 Interval plot of BMI

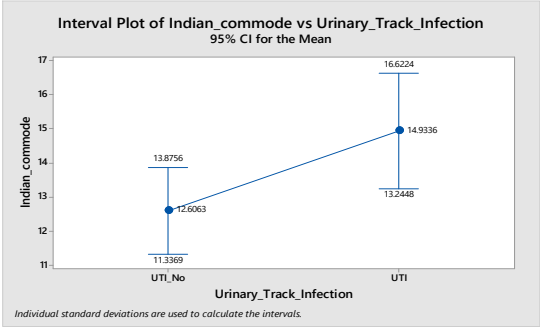


Fig 4.45 Interval plot of IC

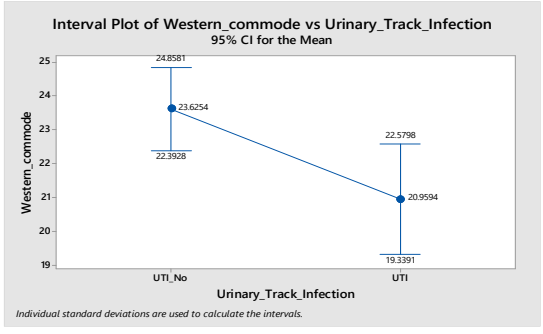


Fig 4.46 Interval plot of WC

Interpretation of one way ANOVA table

Independent samples t-tests comparing the means of Age, BMI, Indian_commode usage, and Western_commode usage between individuals with and without UTIs revealed statistically significant differences for Age ($p = 0.015$), Indian_commode usage ($p = 0.031$), and Western_commode usage ($p = 0.01$), with p-values less than the 0.05 significance threshold. The mean age of the two groups differs significantly, as does the mean reported usage of both Indian and Western commodes. However, there was no statistically significant difference in mean BMI ($p = 0.116$) between individuals with and without UTIs in this study. The ANOVA results show that age and commode usage (both Indian and Western) are strongly related to the presence of UTIs in this dataset. Individuals with UTIs are slightly older on average. There are considerable disparities in how people with and without UTIs use Indian and Western commodes. However, there is no statistically significant difference in BMI between the two groups, implying that it may not be a significant predictor of UTI recurrence in this sample.

Interval plots from a one-way ANOVA test investigating the link between urinary tract infection (UTI) and other parameters show intriguing tendencies. The age plot indicates a slightly higher average age among those with UTIs, but the overlapping confidence intervals imply that this difference may not be statistically significant. In contrast, the BMI plot indicates a slightly lower average BMI in the UTI group, with overlapping confidence intervals indicating probable non-significance. When it comes to toilet usage, the plot for Indian commode usage shows that individuals with UTIs had a greater mean usage, with low overlap in confidence intervals implying a possibly larger difference. In contrast, the Western commode usage figure indicates a decreased mean usage in the UTI group, with some overlap in confidence intervals.

Overall, the plots indicate possible connections between UTI status, age, BMI, and commode usage; however, the overlapping confidence intervals for BMI suggest that these differences may not be statistically significant. The commode usage graphs, notably for the Indian commode, show less overlap, indicating a possibly stronger correlation that justifies additional statistical analysis using the ANOVA test findings.

4.2.3 Proportion test

The following table shows the proportion of getting UTI for different usage of commodes in overall data.

Table 4.6 Proportion of UTI in overall data

Simplified_COMMODE (N=845)				
Simplified_COMMODE	UTI	UTI No	Commode Total	Proportion of UTI
Indian	15	10	25	60.0%
Missing	2	20	22	9.1%
Mixed	150	268	418	35.9%
Western	104	276	380	27.4%
UTI Total	271	574	845	32.1%

Among the total participants, 49% (418/845) were mixed users (used both IC and WC), 45% (380/845) were exclusive WC users, 3% (25/845) were exclusive IC users, and 3% (22/845) did not specify their usage. The following table is the proportion test for the Mixed Vs Western commode for 95% confidence for overall data.

Table 4.7 Proportion test for Mixed Vs Western commode for allover data

Proportion Test	Mixed Vs Western		H0: p1 - p2 = 0
	P-value =	0.006	H1: p1 - p2 > 0

UTI was reported by 35.9% (150/418) of the mixed users and 27.4% (104/380) of exclusive WC users (P-value: 0.006).

The following table shows the proportion of getting UTI for different usage of commodes in the age interval 21-40 years Age old women.

Table 4.8 Proportion of UTI in Age interval 21-40

Age_Interval = 21-40 (N=417)				
Simplified_COMMODE	UTI	UTI No	Commode Total	Proportion of UTI
Indian	8	3	11	72.7%
Missing	0	10	10	0.0%
Mixed	63	128	191	33.0%
Western	44	161	205	21.5%
UTI Total	115	302	417	27.6%

Among the participants in the Age Interval (21-40) years, 46% (191/417) were mixed users (used both IC and WC), 49% (205/417) were exclusive WC users, 2.6% (11/417) were exclusive IC users, and 2.4% (10/417) did not specify their usage.

The following table is the proportion test for the Mixed Vs Western commode for 95% confidence for age interval 21-40 years Age old women.

Table 4.9 Proportion test for Mixed Vs Western commode for age interval 21-40

Proportion Test	Mixed Vs Western		H0: p1 - p2 = 10
	P-value =	0.007	H1: p1 - p2 > 0

Among the participants in the Age Interval (21-40) years, UTI was reported by 33% (63/191) of the mixed users and 21.5% (44/205) of exclusive WC users (P-value: 0.006).

The following table shows the proportion of getting UTI for different usage of commodes in the age interval 41-60 years Age old women.

Table 4.10 Proportion of UTI in Age interval 41-60

Age_Interval = 41-60 (N=304)				
Simplified_COMMODE	UTI	UTI No	Commode Total	Proportion of UTI
Indian	6	7	13	46.2%
Missing	2	7	9	22.2%
Mixed	64	96	160	40.0%
Western	40	82	122	32.8%
UTI Total	112	192	304	36.8%

Among the participants in the Age Interval (41-60) years, 53% (160/304) were mixed users (used both IC and WC), 40% (122/304) were exclusive WC users, 4% (13/304) were exclusive IC users, and 3% (9/304) did not specify their usage.

The following table is the proportion test for the Mixed Vs Western commode for 95% confidence for age interval 41-60 years Age old women.

Table 4.11 Proportion test for Mixed Vs Western commode for age interval 41-60

Proportion Test	Mixed Vs Western		H0: p1 - p2 = 0
	P-value =	0.131	H1: p1 - p2 > 0

Among the participants in the Age Interval (41-60) years, UTI was reported by 40% (64/160) of the mixed users and 32.8% (40/122) of exclusive WC users (P-value: 0.131).

The following table shows the proportion of getting UTI for different usage of commodes in the age interval of women age more than 60 years.

Table 4.12 Proportion of UTI in Age interval more than 60

Age_Interval = more than 60 (N=124)				
Simplified_COMMODE	UTI	UTI No	Commode Total	Proportion of UTI
Indian	1	0	1	100.0%
Missing	0	3	3	0.0%
Mixed	23	44	67	34.3%
Western	20	33	53	37.7%
UTI Total	44	80	124	35.5%

Among the participants in the Age Interval (more than 60) years, 54% (67/124) were mixed users (used both IC and WC), 43% (53/124) were exclusive WC users, 1% (1/124) were exclusive IC users, and 2% (3/124) did not specify their usage.

The following table is the proportion test for the Mixed Vs Western commode for 95% confidence for age interval 41-60 years Age old women.

Table 4.13 Proportion test for Mixed Vs Western commode for age interval more than 60

Proportion Test	Mixed Vs Western		H0: p1 - p2 = 0
	P-value =	0.422	H1: p1 - p2 < 0

Among the participants in the Age Interval (more than 60) years, UTI was reported by 34.3% (23/67) of the mixed users and 37.7% (20/53) of exclusive WC users (P-value: 0.422).

Interpretation of Mixed vs. Western Commode Comparison

Overall Commode Usage and UTI Proportion (N=845)

A proportion test between Mixed and Western commode users revealed a statistically significant difference (p-value = 0.006), with a higher proportion of UTIs in the Mixed group.

The alternative hypothesis (H1: p1 - p2 > 0) implies that the test was explicitly meant to determine whether the proportion of UTIs was higher in the Mixed commode group versus the Western commode group.

Age Interval Analysis - The analysis is further broken down into three age intervals

Age Interval 21-40 (N=417)

The Mixed commode group had a higher UTI proportion (33.0%) than the Western group (21.5%), and this difference was statistically significant (p-value = 0.007).

Age Interval 41-60 (N=304)

The Mixed commode group again showed a higher UTI proportion (40.0%) than the Western group (32.8%), but this difference was not statistically significant (p-value = 0.131).

Age Interval >60 (N=124)

In this older age group interestingly, the Western commode group had a little higher UTI proportion (37.7%) than the Mixed group (34.3%), although the difference was not statistically significant (p-value = 0.422). The alternative hypothesis (H1: p1 - p2 < 0) suggests the test was designed to examine if the Western group had a smaller proportion, which was not supported.

Interpretation of proportion tests

The data reveal a strong relationship between commode preference and UTI risk, which may be greater among Indian commode users. Younger women (21-60), particularly those who used mixed commodes, had a greater proportion of UTIs than main Western commode users. Age appears to modulate this link, with the correlation between commode type and UTI being stronger in younger women (21-40) and lessening with age (41-60 revealing non-significant data). The non-linear association for Indian commode usage in the 41-60 group, as well as the variation in model fit across ages, suggest that age is a substantial impact modifier, with additional variables likely influencing UTI risk, particularly in elderly people. Interpretations for categories with limited sample sizes should be used with caution. Further multivariate analysis is required to determine independent effects and interactions.

4.2.4 Regression

The following table shows the regression equations for probability of getting affected by UTI in different age groups.

Table 4.14 Regression equation for age intervals

Age interval 21-40	IC	P(UTI) = 0.2215 + 0.005065 Indian_commode
	WC	P(UTI) = 0.3768 - 0.004922 Western_commode
Age interval 41-60	IC	P(UTI) = 0.3674 + 0.000321 Indian_commode- 0.000008 Indian_commode^2
	WC	P(UTI) = 0.4280 - 0.002510 Western_commode

Interpretation of regression Equation

Logistic regression equations that predict the chance of Urinary Tract Infection (UTI) as a function of toilet usage (Indian and Western) within two age intervals: 21-40 and 41-60. These equations allow us to assess how commode preference may influence UTI risk while accounting for age.

Age Interval 21-40

- **Indian Commode (IC)**: P(UTI) = 0.2215 + 0.005065 * Indian_commode
- The frequency of Indian commode use increases, the model predicts a slight increase in the probability of UTI in this age group.
- **Western Commode (WC)**: P(UTI) = 0.3768 - 0.004922 * Western_commode
- This implies that in younger individuals, increased Western commode usage is associated with a slight decrease in the predicted probability of UTI.

Age Interval 41-60

- **Indian Commode (IC)**: P(UTI) = 0.3674 + 0.000321 * Indian_commode - 0.000008 * Indian_commode^2

- This equation is quadratic, indicating a non-linear relationship. The interpretation is more complex, suggesting that the effect of Indian commode usage on UTI probability isn't constant in this age group and may increase at a decreasing rate.
- **Western Commode (WC)**: P(UTI) = 0.4280 - 0.002510 * Western_commode
- Similar to the younger age group, increased Western commode usage is associated with a decrease in UTI probability, although the magnitude of the decrease is different.

Rate of increase in the Age Intervals

The following table shows the probability of getting affected by UTI in the regression equation over two different age groups and usage of commode increasing over years.

Table 4.15 Rate of increase for comparing the commodes

	Age interval 21-40		Age interval 41-60	
No. of years	IC	WC	IC	WC
(Adj) R-square	85.3	58.2	0.01	54.7
10	27.2%	32.8%	37.0%	40.3%
20	32.3%	27.8%	37.1%	37.8%
30	37.3%	22.9%	37.0%	35.3%
40	42.4%	18.0%	36.7%	32.8%
Rate of Increase	15.2%	-14.8%	-0.2%	-7.5%

Interpretation of rate of increase table

For the younger age group (21-40), the corrected R-squared values suggest that commode type has a substantial predictive capacity on UTI risk, with IC showing an increasing likelihood (from 27.2% to 42.4%) and WC showing a decreasing chance (from 32.8% to 18.0%) across the observed timeframe. In contrast, the older age group (41-60) had significantly lower adjusted R-squared values, indicating a reduced impact of commode type on UTI risk. Within this older group, the chance of IC is relatively steady (about 37%), but the probability of WC is falling (from 40.3% to 32.8%). The rates of change highlight these tendencies even more, with a significant increase for IC and a drop for WC in the younger group, whereas the older group exhibits negligible change for IC and a moderate decline for WC.

Comparing Age Interval 21-40 IC Vs WC

The following diagrammatic representation shows the fitted regression equation lines for age interval 21-40 years old women.

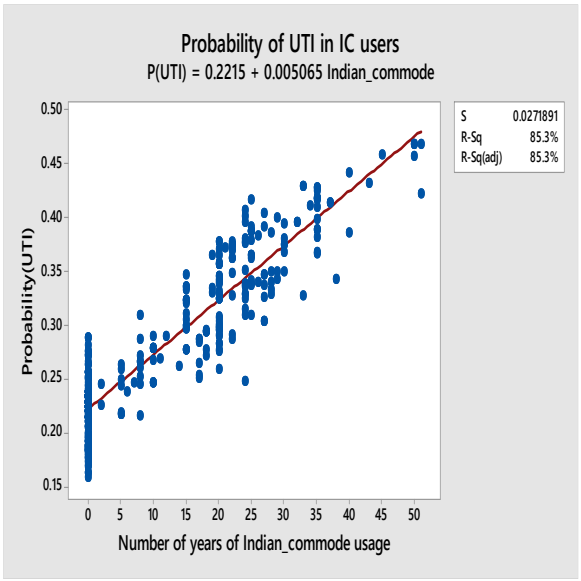


Fig 4.47 Linear fitted line for IC

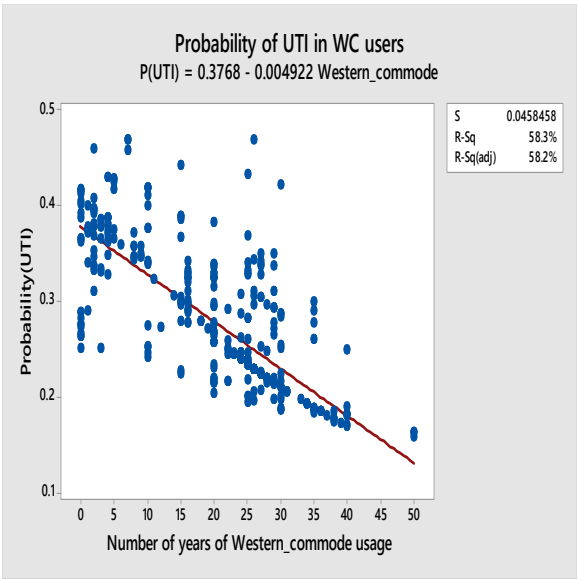


Fig 4.48 Linear fitted line for WC

Interpretation of comparing Age Interval 21-40 (IC Vs WC)

For Indian commode (IC) users, the strong positive linear connection (R-squared = 85.3%) indicates that the number of years women use an Indian commode increases dramatically, as does their projected likelihood of contracting a UTI. Each extra year of IC use is related with an approximately 0.5% increase in the likelihood of developing UTI. Conversely, for Western commode (WC) users, the moderate negative linear association (R-squared = 58.3%) indicates that as the number of years women use a Western toilet grows, their projected likelihood of contracting a UTI reduces. Each extra year of WC use is associated with a 0.5% decrease in the likelihood of developing a UTI. The baseline projected likelihood of being impacted by a UTI when using for zero years is significantly greater for WC users (approximately 37.7%) than for IC users (about 22.2%). The model for IC users explains a greater fraction of the variability in UTI probability depending on years of use than the model for WC users. This shows that the duration of Indian commode usage is a better linear predictor of UTI risk than Western commode usage.

Comparing Age Interval 41-60 IC Vs WC

The following diagrammatic representation shows the fitted regression equation lines for age interval 41-60 years old women.

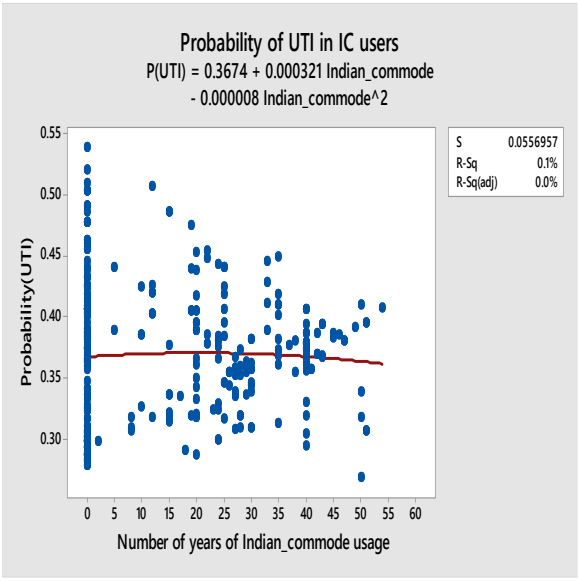


Fig 4.49 Quadratic fitted line for IC

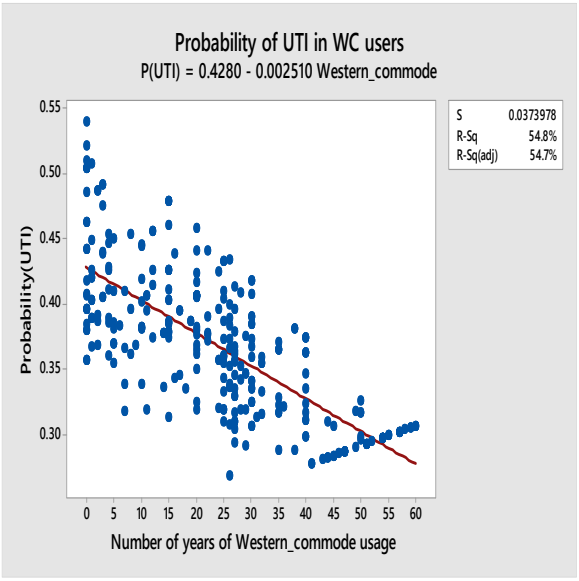


Fig 4.50 linear fitted line for WC

Interpretation of comparing Age Interval 41-60 (IC Vs WC)

The regression analyses for Indian commode (IC) and Western commode (WC) usage on the probability of UTI in women aged 41-60 show significant differences in the nature and degree of their connections. The quadratic model for IC usage had an extremely poor fit (R-squared = 0.1%), implying that the number of years using an Indian commode, even with a non-linear factor, is a minor predictor of UTI probability in this age range. In contrast, the linear model for WC usage showed a moderately strong negative connection (R-squared = 54.8%), indicating that increasing the length of WC usage is related with a significant decrease in the projected chance of UTI. This significant contrast suggests that, among this age group, the duration of Western commode use is a far more relevant factor in predicting UTI risk than the duration of Indian commode use, as represented by these specific regression models. The near-flat and poorly fitting curve for IC suggests that other unmeasured variables are far more important, but the evident downward trend for WC indicates a possible protective connection with longer-term use.

Commode comparison

The following diagrammatic representation shows the probability getting affected by UTI in different age groups.

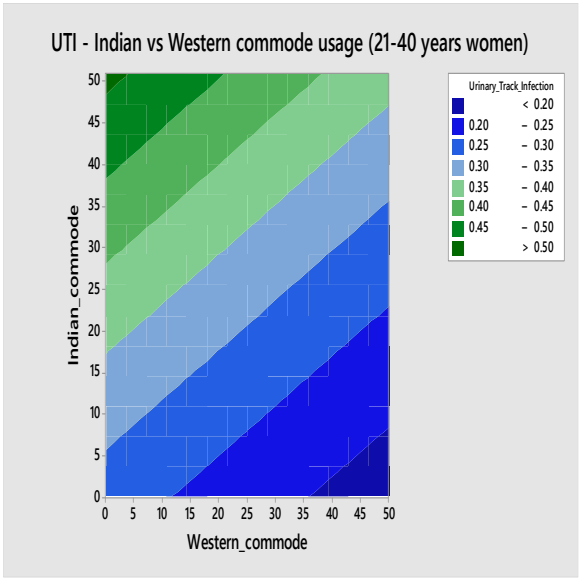


Fig 4.51 Contour plot for age group 21-40

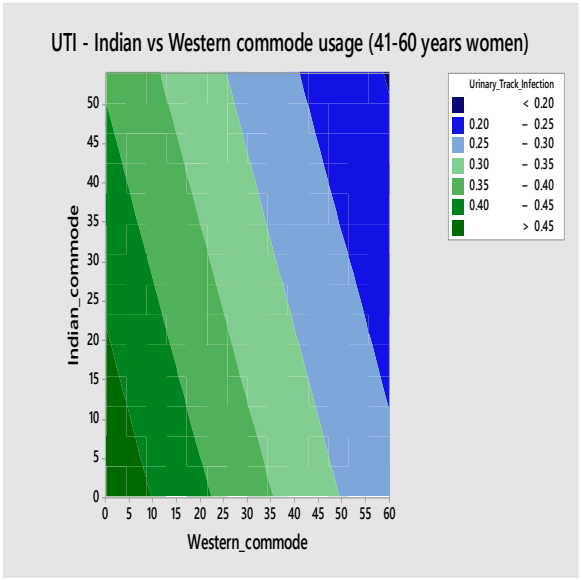


Fig 4.52 Contour plot for age group 41-60

Interpretation commode comparison with contour plot

The comparison of UTI risk between two age groups, women aged 21-40 and 41-60, based on commode usage, reveals significant patterns supported by probabilities. In the 21-40 age group, women with high Indian commode usage (40-50 times) and low Western usage (0-10 times) have a UTI probability larger than 0.45, indicating a higher infection risk. Conversely, those with low Indian usage (0-10 times) and high Western usage (40-50 times) have a UTI likelihood less than 0.20, indicating a preventive effect. Western commode use is almost exclusively responsible for UTI risk in people aged 41 to 60. Women who use Western commodes less than 10 times per period, regardless of Indian usage, have UTI likelihood greater than 0.45, but those who use Western commodes 50-60 times have a probability less than 0.20. Indian commode use has little influence on this senior demographic. These findings suggest that, whereas both commode types influence UTI risk in younger women, Western commode usage is a stronger protective factor in older women, most likely due to the increased physical difficulty and cleanliness concerns associated with Indian toilets as age progresses.

4.2.5 Chi-square test for association for categorical data

The table presents the results of chi-square tests of independence for various categorical variables against the binary outcome variable, UTI Status (UTI vs. UTI No). **Variables -** Lists the categorical independent variables. **Elements -** Specifies the categories within each variable being compared. **UTI -** Count of patients with a UTI in each category. **UTI No -** Count of patients without a UTI in each category. **Chi-Square Value -** The calculated chi-square statistic for the test. **Chi-Square p-value -** The probability of observing a chi-square statistic as extreme as, or more extreme than, the one calculated, assuming no association between the variables (null hypothesis). A p-value below a chosen significance level (commonly 0.05) suggests rejecting the null hypothesis and concluding a statistically significant association. **Fisher's Exact Test p-value -** An alternative test for independence, particularly useful when expected cell counts in the chi-square test are small (typically < 5).

This chi-square summary table provides a more rigorous statistical foundation for analyzing the potential correlations between the variables studied and UTI status. Variables having statistically significant p-values merit additional research using advanced statistical modeling approaches. The chi-square test for association, a non-parametric approach appropriate for categorical data, evaluates variable independence by quantifying the difference between observed and predicted frequencies under the assumption of no relationship. A larger chi-square statistic indicates a greater difference, whereas a significant p-value (usually < 0.05) rejects the null hypothesis of independence, indicating a relationship exists. However, because the test is sensitive to sample size, large samples can produce significant results for weak associations, whereas small samples may lack the power to detect real effects; importantly, chi-square association does not imply causation, and interpreting the practical significance necessitates examining the observed patterns within the data in addition to the statistical significance.

Fisher's precise Test p-value is an alternative to the chi-square test, particularly when predicted cell counts are low. It calculates the precise probability of witnessing the present or more extreme table configurations under the null hypothesis of no connection. It's especially handy with small sample sizes, because the chi-square approximation may be incorrect.

The following table shows the number persons affected by UTI in all categorical variables and give the Chi-square value and its p-value and Fisher's exact p-value.

Table 4.16 Chi-Square association test table

Variables	Elements	UTI	UTI No	Chi-Square value	Chi-Square p-value	Fisher's Exact Test
Age_Interval	21-40	115	302	0.88	0.02	-
	41-60	112	192			
	more than 60	44	80			
Education	E Medical	219	479	2.77	0.35	0.38
	E Non Medical	52	95			
Occupation	O Medical	163	379	2.77	0.10	0.11
	O Non Medical	108	195			
M_Status	Married	259	553	0.29	0.59	0.57
	Unmarried	12	21			
BMI_Categorised	Class 1 obese	26	61	4.06	0.54	-
	Class 2 obese	12	27			
	Class 3 obese	1	6			
	Normal	126	243			
	Overweight	85	204			
	underweight	21	33			
Diabetes	DB	53	30	42.68	0.00	0.00
	DB No	218	544			
Hypertensio	HT	50	54	13.95	0.00	0.00
	HT No	221	520			
Coronary_heart_disease	CHD	11	9	4.94	0.03	0.05
	CHD No	260	565			
Chronic_pulmo_lung_disease	CPL	23	46	0.06	0.82	0.79
	CPL No	248	528			
Back_vertebral_spinal_pain	BVS	10	18	0.18	0.67	0.68
	BVS No	261	556			

Variables	Elements	UTI	UTI No	Chi-Square value	Chi-Square p-value	Fisher's Exact Test
Any_Neurological_disorder	ND	10	18	0.18	0.67	0.68
	ND No	261	556			
Parkinsons	PK	1	3	0.09	0.76	1.00
	PK No	270	571			
Multiple_sclerosis	MS	0	1	0.47	-	1.00
	MS No	271	573			
Brain_tumor	BT	0	2	0.95	-	1.00
	BT No	271	572			
Stroke	ST	0	5	2.38	0.12	0.18
	ST No	271	569			
Any_other	AO	9	7	4.38	0.04	0.05
	AO No	262	567			
Persistent_severe_lower_pain	PSL	3	13	1.33	0.25	0.29
	PSL No	268	561			
Other_chronic_disease	OCD	17	6	19.00	0.00	0.00
	OCD No	254	568			
Simplified_COMMODE	Indian	15	10	20.93	0.00	-
	Missing	2	20			
	Mixed	150	268			
	Western	104	276			
empty_bladded_frequently	EBF	105	214	2.93	0.23	-
	EBF No	21	29			
	EBF No leakage	145	331			
Fluid_restriction	FR	14	22	25.23	0.00	-
	FR No	65	64			
	FR No leakage	192	488			
Empty_bladder_prior_go_out	EBP No	18	17	23.78	0.00	-
	EBP No leakage	192	488			
	EMP	61	69			

Variables	Elements	UTI	UTI No	Chi-Square value	Chi-Square p-value	Fisher's Exact Test
Lookout_for_toilet	LFT	37	22	32.79	0.00	-
	LFT No	42	64			
	LFT No leakage	192	488			
Pad_protection_to_management	PP	37	14	44.14	0.00	-
	PP No	42	72			
	PP No leakage	192	488			
Coffee	C-0	241	543	24.17	0.00	-
	C-1	13	3			
	C-2	8	22			
	C-3	9	6			
Tea	T-0	248	548	34.05	-	-
	T-0.5	0	2			
	T-1	0	16			
	T-2	23	8			
Softdrinks	SD-0	265	566	12.42	0.00	-
	SD-0.5	1	8			
	SD-1	5	0			
Fruit_Juices	FJ-0	58	79	8.74	-	-
	FJ-1	213	493			
	FJ-2	0	2			
Liquid_consumption	LC-0	266	574	10.65	0.00	0.00
	LC-1	5	0			
No_of_children	No.C 0	44	121	6.32	0.10	-
	No.C 1	89	170			
	No.C 2	137	272			
	No.C 3	1	11			
Concatinate_Delivary	CD	4	27	12.01	0.04	-
	ECD	52	120			
	EVD	114	213			
	FD	50	77			
	Nullipara	44	121			
	VD	7	16			

Variables	Elements	UTI	UTI No	Chi-Square value	Chi-Square p-value	Fisher's Exact Test
Menopause_attained	MA	157	284	5.28	0.02	0.02
	MA No	114	290			
Hysterectomy_done	HYT	30	87	2.58	0.11	0.11
	HYT No	241	487			
Ovaries_removed	OR	18	54	1.81	0.18	0.19
	OR No	253	520			
Had_any_unrinary_problem	HAU	271	40	685.01	0.00	0.00
	HAU No	0	534			

The following graphical representations show the Visualization of Variables Associated with UTI from the Chi-square test for association for categorical data.

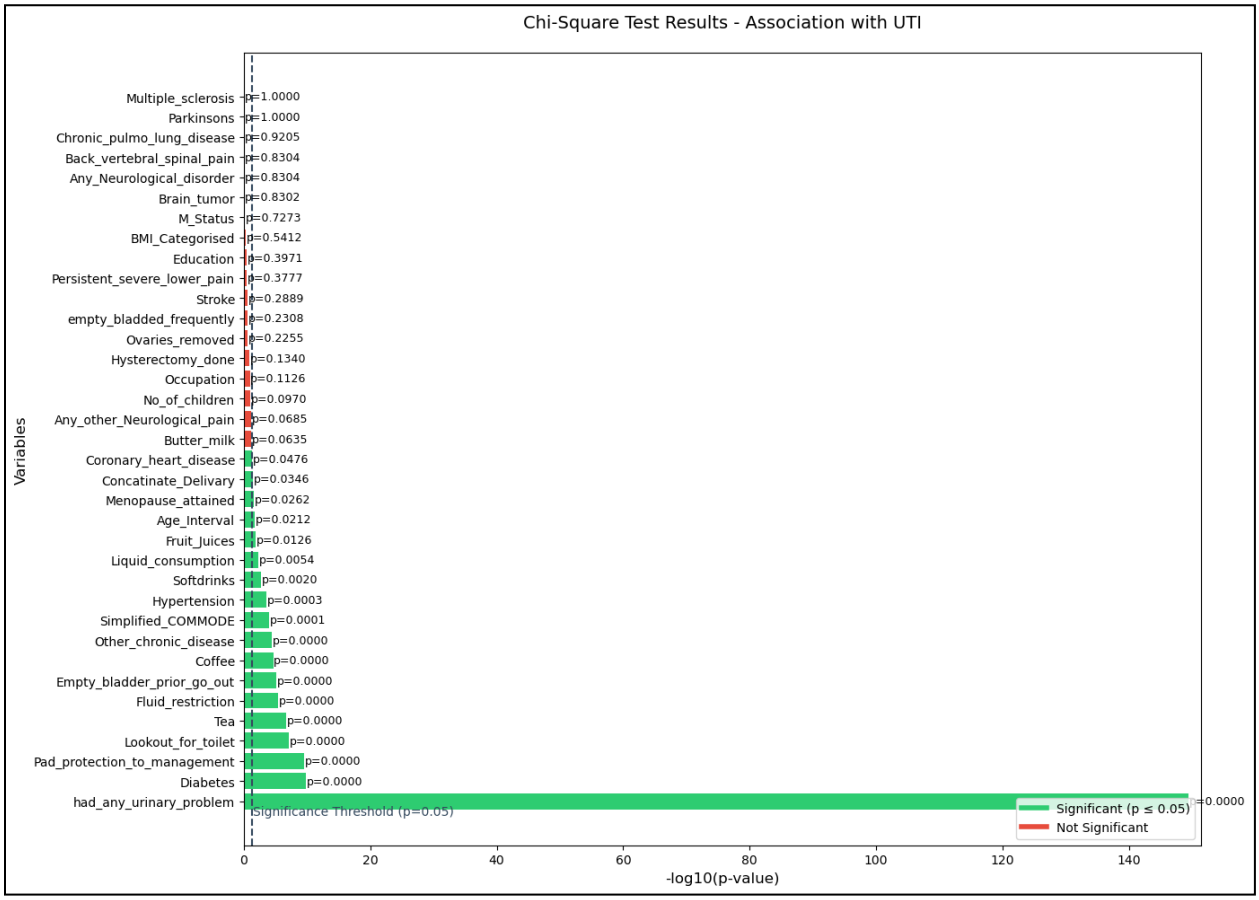


Fig 4.53 Visualization of variables associated with UTI (Chi-Square Test)

Interpretation of Chi-Square table

Summary of Statistically Significant Associations ($p < 0.05$)

The chi-square tests reveal statistically significant associations between UTI status and the following variables - Diabetes, Hypertension, Coronary Heart Disease, Any Other Neurological Pain, Other Chronic Disease, Simplified Commode, Empty Bladder Frequently, Fluid Restriction, Empty Bladder Prior Go Out, Lookout For Toilet, Pad Protection To Management, Coffee Consumption, Tea Consumption. Softdrinks Consumption, Fruit Juices Consumption, Liquid Consumption, Conconcatenate Delivery, Menopause Attained, Had Any Urinary Problem.

Using chi-square testing, some characteristics have statistically significant relationships ($p < 0.05$) with UTI status. Diabetes ($p=0.00$) and a history of earlier urinary issues ($p=0.00$) have the strongest connections, implying that they are substantial risk factors for UTI. Certain hygiene and lifestyle characteristics, such as simplified commode ($p=0.00$), frequency of bladder emptying ($p=0.00$), fluid restriction ($p=0.00$), bladder emptying before leaving the house ($p=0.00$), and pad protection ($p=0.00$), also show extremely significant relationships. Furthermore, other chronic diseases ($p=0.00$), coffee consumption ($p=0.00$), tea consumption ($p=0.00$), soft drink consumption ($p=0.00$), fruit juice consumption ($p=0.00$), liquid consumption ($p=0.00$), delivery history ($p=0.01$), and menopause attainment ($p=0.02$) all show statistically significant relationships. Coronary heart disease ($p=0.03$) and any other neurological discomfort ($p=0.04$) reveal substantial relationships, albeit with limitations due to sample size. In contrast, demographic variables, BMI, and several other chronic illnesses do not have statistically significant relationships ($p > 0.05$). These findings call for additional exploration utilizing multivariate modeling to identify separate effects and probable causal pathways.

Using chi-square testing, several characteristics did not show statistically significant associations ($p > 0.05$) with UTI status. Demographic characteristics include age interval ($p=0.07$), education ($p=0.15$), occupation ($p=0.11$), and marital status ($p=0.29$). Similarly, BMI Categorised ($p=0.54$) and reproductive history characteristics such as number of children ($p=0.10$), hysterectomy performed ($p=0.11$), and ovaries removed ($p=0.19$) had no significant link with UTI. In terms of health conditions, Chronic Pulmonary Lung Disease ($p=0.82$),

Back/Vertebral/Spinal Pain (p=0.67), Any Neurological Disorder (p=0.67), Parkinson's (p=0.76), Multiple Sclerosis (p=1.00), Brain Tumor (p=1.00), Stroke (p=0.12), and Persistent Severe Lower Pain (p=0.29) all showed no significant associations. Non-significant results for disorders with small sample numbers (Parkinson's, Multiple Sclerosis, Brain Tumor, and maybe Stroke and Coronary Heart Disease) should be viewed with caution due to a lack of statistical power to identify a true effect.

The strongest associations point to key risk factors - The extremely significant p-values (p=0.00) for diabetes and a history of previous urinary difficulties clearly imply that these are key and important risk factors for UTIs in this cohort. This emphasizes the need of taking into account prior urinary health and metabolic factors when determining UTI risk.

Hygiene and lifestyle are modifiable risk factors - The cluster of highly significant hygiene and lifestyle features (simple commode, frequency of bladder emptying, fluid restriction, bladder emptying before leaving the home, and pad protection) indicates that these practices play an important influence in UTI incidence. This creates opportunities for prospective treatments and patient education that focus on modifiable risk factors. The "simplified commode" organization calls for greater inquiry into what defines a "simplified" toilet and its potential relationship to hygienic practices.

Systemic Health and UTI Risk - Significant connections with hypertension, coronary heart disease, any other neurological pain, and other chronic diseases point to a possible relationship between general systemic health and UTI susceptibility. This implies that people with these illnesses may have changed physiological states that raise their risk. The emphasis on sample size limits for Coronary Heart Disease and maybe other neurological diseases is critical; while significant in this test, bigger investigations are required for confirmation.

Fluid Intake and Urinary Health - The substantial connections with several components of fluid intake (coffee consumption, tea consumption, softdrink consumption, fruit juice consumption, liquid consumption, fluid restriction) suggest a complicated link between hydration and the risk of UTI. The direction of these correlations (whether higher or lower intake increases risk) would need more investigation (e.g., examination of contingency tables). However, it is obvious that the kind and quantity of fluid consumption are important.

Hormonal and Reproductive Factors - Significant Associations with Conconcatenate Delivery and Menopause. The findings relate to the impact of hormonal shifts and reproductive history on UTI risk in women. Menopause, with its accompanying estrogen reduction, is a recognized cause, and the relation to "Conconcatenate Delivery" might refer to type of delivery.

Demographics Less Directly Influential (in this model) - The lack of significant associations with core demographic variables such as age interval, education, occupation, and marital status suggests that these factors, both individually and within this Chi-Square analysis, are not strong direct predictors of UTI status.

BMI and Direct Reproductive Events - The non-significant results for BMI Categorised, number of children, hysterectomy done, and ovaries removed indicate that these variables, as classified and evaluated here, do not have a direct independent connection with UTI risk. This does not rule out the potential of indirect effects, or their importance in more complicated models.

Specific Chronic Conditions and UTI - The absence of statistically significant associations with specific chronic conditions such as Chronic Pulmonary Lung Disease, Back/Vertebral/Spinal Pain, Any Neurological Disorder (general), Parkinson's, Multiple Sclerosis, Brain Tumor, Stroke, and Persistent Severe Lower Pain suggests that these conditions are not directly related to UTI status in this analysis. However, caution should be exercised when using small sample sizes for any of these disorders, since a real link may exist but the research lacks the power to detect it.

The bar chart displays the results of Chi-Square tests used to determine the relationship between various variables and urinary tract infection (UTI) status. The length of each bar shows the negative logarithm (base 10) of the p-value for each variable. Longer bars indicate a larger statistical significance of the link ($p < 0.05$). Variables with bars that extend above the dashed vertical line (significance level at $p = 0.05$) have a statistically significant association with UTI. Notably, "had_any_urinary_problem" and "Diabetes" have the strongest associations, followed by other significant factors related to bladder habits, hygiene, lifestyle, and some pre-existing conditions, whereas many demographic and other health-related variables have no significant associations with UTI status in this analysis.

4.3 SUPERVISED MACHINE LEARNING

Supervised Machine Learning is a sort of machine learning in which the model is trained with a labeled dataset. In this paradigm, the algorithm learns from input-output pairs, with each input (feature vector) linked to a corresponding output (target label). The goal is to learn a mapping function from inputs to outputs that can be used to predict the outcome of fresh, previously unknown data.

Key Components

1. **Features (Independent Variables):** The input variables used to make predictions.
2. **Target (Dependent Variable):** The output variable the model is trained to predict.
3. **Training Data:** A dataset with known inputs and corresponding outputs used to train the model.
4. **Test Data:** A separate dataset used to evaluate the performance of the trained model.

Types of Supervised Learning Tasks

- **Classification:** Predicts a categorical label.
  Example: Predicting whether a patient has a urinary tract infection (UTI) based on symptoms and medical history (binary classification).
- **Regression:** Predicts a continuous numerical value.
  Example: Estimating blood pressure based on age, weight, and other factors.

Workflow of Supervised Learning

1. **Data Preprocessing:** Handling missing values, encoding categorical variables, scaling features.
2. **Feature Selection/Engineering:** Identifying the most relevant predictors using techniques like Chi-Square test.
3. **Model Selection:** Choosing an appropriate algorithm (e.g., Logistic Regression, Decision Trees, and Random Forest).

4. **Model Training:** Feeding the training data into the model to learn the mapping.
5. **Model Evaluation:** Assessing performance using metrics such as Accuracy, Precision, Recall, F1-score (for classification) or RMSE, MAE (for regression).
6. **Prediction:** Applying the trained model to new data for inference.

Advantages

Clear relationship between inputs and outputs. Model performance is easy to measure. Suitable for real-world applications with labeled datasets.

Common Algorithms

- Logistic Regression
- Decision Trees
- Random Forest
- XG-Boost
- Support Vector Machines
- k-Nearest Neighbors
- Neural Networks

Limitations of Supervised Machine Learning

1. **Requires Labeled Data:** Supervised learning depends on large, labeled datasets. Acquiring high-quality labels can be costly and time-intensive, especially in specialized fields like healthcare.
2. **Overfitting Risk:** Models may memorize training data instead of learning general patterns, reducing performance on unseen data—especially when the dataset is small or noisy.

3. **Limited to Known Patterns:** Supervised models cannot detect new or unexpected patterns outside of what they were trained on, making them less adaptive to changing environments.

4. **Imbalanced Data Issues:** When one class dominates (e.g., fewer UTI cases), models may become biased, leading to poor detection of minority classes unless techniques like SMOTE or reweighting are applied.

5. **Feature Dependence:** Model accuracy heavily depends on the quality and relevance of features. Poor feature selection can degrade performance, while good features require domain expertise.

6. **Scalability Constraints:** Some algorithms struggle with very large or high-dimensional datasets, leading to increased training time and computational load.

7. **Sensitive to Noisy Data:** Mislabels or outliers can mislead the learning process, making it crucial to clean and validate the training data.

8. **Interpretability Challenges:** Complex models like ensembles or neural networks may be hard to interpret, which can be problematic in critical domains where decision transparency is needed.

In supervised machine learning, labeled data serves as the foundation for training algorithms to produce predictions or classifications. This type of data includes of input characteristics and the proper output labels. For example, in image classification, labeled data can comprise photographs of cats labeled as "cat" and images of dogs classified as "dog." During the training phase, the algorithm learns the link between these characteristics and the labels assigned to them. As a result, when supplied with fresh, unlabeled data, the trained model may predict the correct label using the patterns learnt from labeled samples. The accuracy of the supervised learning model is strongly dependent on the quality and amount of labeled training data. Supervised learning approaches include artificial neural networks (ANNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs), which are widely utilized in classification and regression. They need labeled data to learn the mapping function from inputs to outputs.

4.3.1 LOGISTIC REGRESSION

Models Type

- **All Variables:** This model likely includes all the available independent variables in the dataset. The idea is to leverage every piece of information to predict UTI.
- **Chi-square Selected Variables:** This model uses only the variables that were found to have a statistically significant association with UTI in the Chi-square analysis. This variable selection technique aims to improve model parsimony and potentially avoid over fitting.
- **Single Variable/best predictor Variable:** This model is identified for only the highly associated one single best predictor variable.
- **Excluded Single Variable/ Variables to be excluded:** this model takes all the variables except best predictor single variable from the group of Chi square indentified variables.(Ref Fig no 4.53)

4.3.1.1 Logistic Regression for All Variables

Logistic Regression is a supervised classification technique that predicts the likelihood of a binary outcome (such as the presence or absence of disease). The logistic (sigmoid) function, rather than linear regression, is used to model the connection between input features and a

Regression Equation

log(p/(1-p)) = -11465.6769 + (-8430.1075 * Age) + (-1219.9855 * Education) + (8858.5027 * Occupation) + (3159.0289 * M_Status) + (-5757.0875 * BMI) + (-2012.2280 * BMI_Categorised) + (-6986.6835 * Diabetes) + (-7460.5284 * Hypertension) + (-7155.8279 * Coronary_heart_disease) + (-22974.1047 * Chronic_pulmo_lung_disease) + (-22974.1047 * Back_vertebral_spinal_pain) + (12583.4139 * Any_Neurological_disorder) + (-264102.7697 * Parkinsons) + (8033.9537 * Multiple_sclerosis) + (36086.2216 * Brain_tumor) + (39736.8070 * Stroke) + (527.3581 * Any_other_Neurological_pain) + (-13794.9023 * Persistent_severe_lower_pain) + (-31572.3503 * Other_chronic_disease) + (-13794.9397 * Indian_commode) + (-5762.9735 * Western_commode) + (-17314.4039 *

Simplified_COMMODE) + (521263.6342 * empty_bladded_frequently) + (8881.3834 * Fluid_restriction) + (-42712.4486 * Empty_bladder_prior_go_out) + (-509904.4646 * Lookout_for_toilet) + (8715.2228 * Pad_protection_to_management) + (-3844.6043 * Coffee) + (18871.0769 * Tea) + (-3767.1134 * Softdrinks) + (17028.1565 * Butter_milk) + (634630.3161 * Fruit_Juices) + (-4131.9100 * Liquid_consumption) + (-9905.5310 * No_of_children) + (-7318.5808 * Concatinate_Delivary) + (2965.5770 * Menopause_attained) + (-1138.4482 * Hysterectomy_done) + (-111674.8625 * Ovaries_removed)

McFadden's Pseudo R-squared: 1.0000

This **logistic regression equation** models the log-odds of having a UTI based on numerous factors, where a positive coefficient increases the odds and a negative one decreases them. The extremely large coefficients (both positive and negative) and the McFadden's Pseudo R-squared of 1.0000 strongly suggest a model that perfectly separates the two UTI outcomes based on the predictors included. However, such perfect separation in real-world data is highly unusual and often indicative of issues like multi-collinearity, data errors, or an over fit model that might not generalize well to new data. The practical interpretation of individual coefficients is challenging due to their magnitude and the perfect prediction, implying extreme effects of these variables on UTI likelihood within this specific dataset. While scikit-learn (sklearn) in Python is good for creating and assessing machine learning models, including logistic regression, it does not provide a summary result in the same way that statistical tools such as statsmodels do. Sklearn prioritizes prediction and model performance measures over full statistical inference tables containing p-values, standard errors, and confidence intervals for individual coefficients.

Explanation of the statsmodels summary

The result.summary() from statsmodels provides a detailed statistical output, including

Dependent Variable - The variable being predicted (e.g., 'y'). **Model -** The type of model used (e.g., 'Logit'). **Method -** The estimation method (e.g., 'MLE'). **No. Observations -** The number of data points. **Df Residuals -** Degrees of freedom of the residuals. **Df Model -** Degrees of freedom of the model (number of predictors). **Pseudo R-squ -** Measures of goodness-of-fit for logistic regression (e.g., McFadden's R-squared). **Log-Likelihood -** The value of the log-likelihood function at the estimated coefficients. **LL-Null -** The log-likelihood of the null

model. **LLR p-value -** The p-value for the likelihood ratio test, comparing the fitted model to the null model.

This table shows the estimated coefficients for each predictor variable in the model:

- **coef:** The estimated coefficient for each variable. A positive coefficient means that as the variable increases, the log-odds of having a UTI increase (assuming other variables are held constant). A negative coefficient means the log-odds decrease.
- **std err:** The standard error of the coefficient, measuring the precision of the estimate.
- **z:** The z-statistic, calculated as coef / std err, used for hypothesis testing.
- **P>|z|:** The p-value associated with the z-statistic, indicating the probability of observing a z-statistic as extreme as, or more extreme than, the one calculated if the true coefficient were zero (null hypothesis).
- **[0.025 0.975]:** The 95% confidence interval for the coefficient.

Key Observations and Potential Issues

Many 'nan' Values for Standard Errors and P-values: A large number of variables (Age, Education, BMI, BMI_Categorised, Back_vertebral_spinal_pain, Brain_tumor, Stroke, Other_chronic_disease) have 'nan' (Not a Number) for their standard errors, z-statistics, p-values, and confidence intervals. This typically indicates **perfect separation** or **multicollinearity**.

Perfect Separation - This occurs when a predictor variable perfectly predicts the outcome. For example, if all individuals with a certain characteristic always have a UTI and all individuals without it never do, the model can't estimate a stable coefficient.

Multicollinearity - This occurs when predictor variables are highly correlated with each other, making it difficult for the model to isolate the independent effect of each variable.

In logistic regression, quasi-complete separation occurs when the result variable almost perfectly separates one or more predictor factors. This means that for most, but not all, predictor(s) values or categories, the outcome is always one (either 0 or 1).

The following summary shows the Results of Logistic Regression for all variables model.

```
=== All Variables ===
                           Logit Regression Results
==============================================================================
Dep. Variable:     Urinary_Track_Infection   No. Observations:                  918
Model:                               Logit   Df Residuals:                      879
Method:                                MLE   Df Model:                           38
Date:                     Mon, 14 Apr 2025   Pseudo R-squ.:                 1.000
Time:                             07:03:49   Log-Likelihood:           -1.1481e-07
converged:                            True   LL-Null:                       -636.30
Covariance Type:                nonrobust   LLR p-value:                2.139e-242
=================================================================================================
                                    coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
Age                            -1.147e+04        nan        nan        nan         nan         nan
Education                      -8430.1075        nan        nan        nan         nan         nan
Occupation                     -1219.9855   9.02e+10  -1.35e-08      1.000   -1.77e+11    1.77e+11
M_Status                        8858.5027   1.21e+12   7.31e-09      1.000   -2.37e+12    2.37e+12
BMI                             3159.0289        nan        nan        nan         nan         nan
BMI_Categorised                -5757.0875        nan        nan        nan         nan         nan
Diabetes                       -2012.2280   5.82e+10  -3.45e-08      1.000   -1.14e+11    1.14e+11
Hypertension                   -6986.6835   8.11e+10  -8.61e-08      1.000   -1.59e+11    1.59e+11
Coronary_heart_disease         -7460.5284   2.07e+10  -3.61e-07      1.000   -4.05e+10    4.05e+10
Chronic_pulmo_lung_disease     -7155.8279   2.68e+11  -2.67e-08      1.000   -5.25e+11    5.25e+11
Back_vertebral_spinal_pain     -2.297e+04        nan        nan        nan         nan         nan
Any_Neurological_disorder      -2.297e+04        nan        nan        nan         nan         nan
Parkinsons                      1.258e+04    2.1e+12   5.99e-09      1.000   -4.12e+12    4.12e+12
Multiple_sclerosis             -2.641e+05   7.59e+12  -3.48e-08      1.000   -1.49e+13    1.49e+13
Brain_tumor                     8033.9537        nan        nan        nan         nan         nan
Stroke                          3.609e+04        nan        nan        nan         nan         nan
Any_other_Neurological_pain     3.974e+04        nan        nan        nan         nan         nan
Persistent_severe_lower_pain     527.3581    4.9e+11   1.08e-09      1.000    -9.6e+11     9.6e+11
Other_chronic_disease          -1.379e+04        nan        nan        nan         nan         nan
Indian commode                 -3.157e+04   8.11e+10  -3.89e-07      1.000   -1.59e+11    1.59e+11

Western_commode                -1.379e+04        nan        nan        nan         nan         nan
Simplified_COMMODE             -5762.9735    7.8e+10  -7.39e-08      1.000   -1.53e+11    1.53e+11
empty_bladded_frequently       -1.731e+04   1.45e+11  -1.19e-07      1.000   -2.84e+11    2.84e+11
Fluid_restriction               5.213e+05        nan        nan        nan         nan         nan
Empty_bladder_prior_go_out      8881.3834   1.28e+11   6.94e-08      1.000   -2.51e+11    2.51e+11
Lookout_for_toilet             -4.271e+04   1.82e+11  -2.34e-07      1.000   -3.57e+11    3.57e+11
Pad_protection_to_management   -5.099e+05   6.84e+10  -7.46e-06      1.000   -1.34e+11    1.34e+11
Coffee                          8715.2228   1.39e+11   6.26e-08      1.000   -2.73e+11    2.73e+11
Tea                            -3844.6043   1.49e+11  -2.58e-08      1.000   -2.92e+11    2.92e+11
Softdrinks                      1.887e+04        nan        nan        nan         nan         nan
Butter_milk                    -3767.1134        nan        nan        nan         nan         nan
Fruit_Juices                    1.703e+04    5.7e+10   2.99e-07      1.000   -1.12e+11    1.12e+11
Liquid_consumption              6.346e+05   3.28e+12   1.94e-07      1.000   -6.42e+12    6.42e+12
No_of_children                 -4131.9100    1.9e+10  -2.18e-07      1.000   -3.72e+10    3.72e+10
Concatinate_Delivary           -9905.5310   1.07e+11  -9.27e-08      1.000   -2.09e+11    2.09e+11
Menopause_attained             -7318.5808   1.35e+11  -5.44e-08      1.000   -2.64e+11    2.64e+11
Hysterectomy_done               2965.5770   5.27e+11   5.63e-09      1.000   -1.03e+12    1.03e+12
Ovaries_removed                -1138.4482   4.36e+11  -2.61e-09      1.000   -8.55e+11    8.55e+11
had_any_urinary_problem        -1.117e+05        nan        nan        nan         nan         nan
=================================================================================================
```

Fig.4.54 The summary of logistic regression for all variables model

The following graphical representations show the Receiver Operating Characteristic curve, Confusion matrix and Classification report for Logistic regression's **all variables** model.

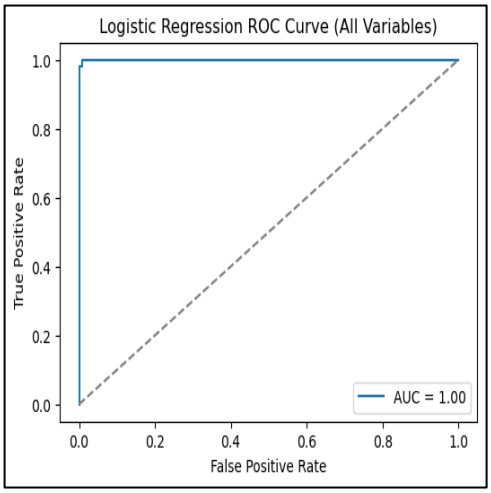


Fig 4.55 ROC curve for all variables model of LR

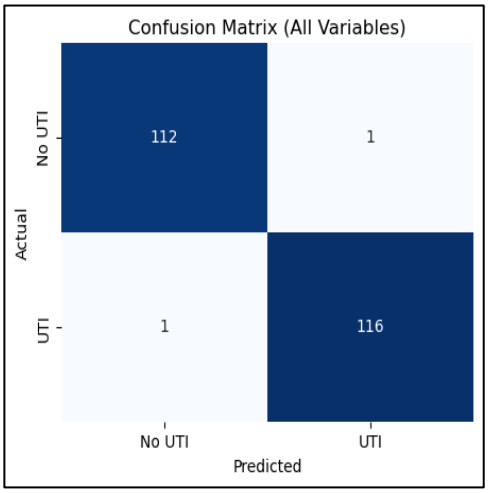


Fig 4.56 Confusion matrix for all variables model of LR

```
Accuracy (All Variables): 0.9913
False Positive Rate (All Variables): 0.0088
Classification Report (All Variables):

              precision    recall  f1-score   support

           0       0.99      0.99      0.99       113
           1       0.99      0.99      0.99       117

    accuracy                           0.99       230
   macro avg       0.99      0.99      0.99       230
weighted avg       0.99      0.99      0.99       230
```

Fig 4.57 Classification report for all variables model of LR

Interpretation of Logistic Regression (All Variables) Results

This logistic regression model, which uses all available variables, has near-perfect classification performance with an AUC of 1.00, surpassing models that use only one variable or exclude a vital one. The confusion matrix indicates very few misclassifications (True Positives=116, True Negatives=112, False Positives=1, False Negatives=1). The classification report validates this, with an accuracy of 0.9913, a low False Positive Rate of 0.0088, and F1-scores of 0.99 in both groups. This highly dependable model, which achieves nearly perfect separation, reduces the danger of missing positive cases, making it incredibly helpful for applications such as UTI detection that require high accuracy and recall. The findings highlight the significance of considering all relevant factors.

4.3.1.2 Logistic Regression for Chi-Square Selected Variables

Regression Equation

log(p/(1-p)) = -0.5860 + (-1.2043 * Diabetes) + (2.2767 * Hypertension) + (2.5227 * Lookout_for_toilet) + (1.2773 * Pad_protection_to_management) + (0.3617 * Coffee) + (1.6865 * Tea) + (-2.0921 * Butter_milk) + (30.9325 * Liquid_consumption) + (0.3058 * Menopause_attained) + (-62.7283 * had_any_urinary_problem)

McFadden's Pseudo R-squared: 0.8595

The following summary shows the Results of Logistic Regression for Chi-Square selected variables model.

```
=== Chi-Square Selected Variables ===
                           Logit Regression Results
==============================================================================
Dep. Variable:     Urinary_Track_Infection   No. Observations:                  918
Model:                               Logit   Df Residuals:                      907
Method:                                MLE   Df Model:                           10
Date:                     Mon, 14 Apr 2025   Pseudo R-squ.:                 0.8595
Time:                             07:04:38   Log-Likelihood:                -89.414
converged:                           False   LL-Null:                       -636.30
Covariance Type:                 nonrobust   LLR p-value:                1.161e-228
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                          -0.5860      1.030     -0.569      0.570      -2.606       1.434
Diabetes                       -1.2043      0.724     -1.663      0.096      -2.624       0.215
Hypertension                    2.2767      0.601      3.787      0.000       1.098       3.455
Lookout_for_toilet              2.5227      1.000      2.523      0.012       0.563       4.482
Pad_protection_to_management    1.2773      0.943      1.355      0.175      -0.570       3.125
Coffee                          0.3617      0.849      0.426      0.670      -1.302       2.025
Tea                             1.6865      0.841      2.005      0.045       0.038       3.335
Butter_milk                    -2.0921      0.651     -3.215      0.001      -3.368      -0.817
Liquid_consumption             30.9325   5.63e+04      0.001      1.000    -1.1e+05     1.1e+05
Menopause_attained              0.3058      0.465      0.658      0.511      -0.605       1.217
had_any_urinary_problem       -62.7283   2.26e+11  -2.78e-10      1.000   -4.43e+11    4.43e+11
================================================================================================
```

Fig.4.58 The summary of logistic regression for Chi-Square Selected Variables model

The following graphical representations show the Receiver Operating Characteristic curve, Confusion matrix and Classification report for Logistic regression's **Chi-Square Selected Variables** model.

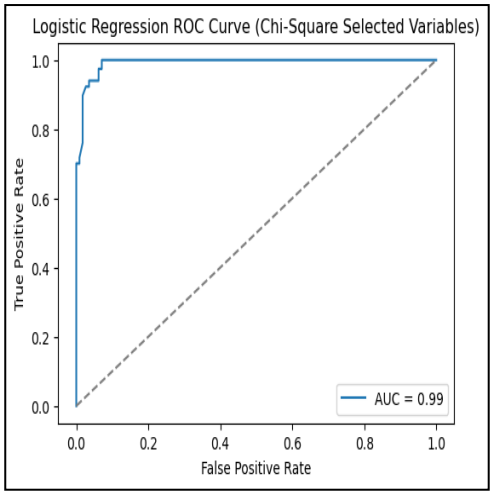


Fig 4.59 ROC curve for Chi-Square Variables model of LR

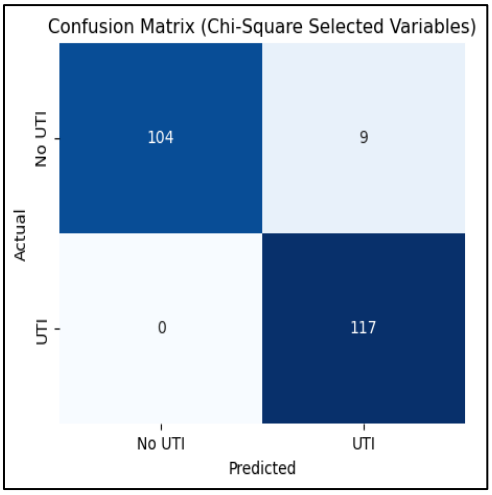


Fig 4.60 Confusion matrix for Chi-Square Selected Variables model of LR

```
Accuracy (Chi-Square Selected Variables): 0.9609
False Positive Rate (Chi-Square Selected Variables): 0.0796
Classification Report (Chi-Square Selected Variables):

              precision    recall  f1-score   support

           0       1.00      0.92      0.96       113
           1       0.93      1.00      0.96       117

    accuracy                           0.96       230
   macro avg       0.96      0.96      0.96       230
weighted avg       0.96      0.96      0.96       230
```

Fig 4.61 Classification report for Chi-Square Selected Variables model of LR

Interpretation of Logistic Regression (Chi-Square Selected Variables) Results

This Logistic Regression model, utilizing Chi-Square selected variables, exhibits excellent discrimination (AUC = 0.99) and high classification accuracy (96%). The ROC curve indicates a strong ability to distinguish between UTI and No UTI cases. The confusion matrix shows accurate classification with high True Positives (117) and True Negatives (104), and crucially, no False Negatives (0), indicating perfect identification of all UTI cases. The classification report confirms these findings with high precision (93% for UTI, 100% for No UTI), perfect recall for UTI (100%), and high recall for No UTI (92%), resulting in balanced F1-scores (0.96 for both).

4.3.1.3 Logistic Regression for Single Variable

Regression Equation

log(p/(1-p)) = -26.6625 + (-22.4009 * had_any_urinary_problem)

McFadden's Pseudo R-squared: 0.8186

The following summary shows the Results of Logistic Regression for Single variable

```
=== Single Variable ===
                           Logit Regression Results
==============================================================================
Dep. Variable:     Urinary_Track_Infection   No. Observations:                  918
Model:                               Logit   Df Residuals:                      916
Method:                                MLE   Df Model:                            1
Date:                     Mon, 14 Apr 2025   Pseudo R-squ.:                  0.8186
Time:                             07:04:40   Log-Likelihood:                -115.44
converged:                           False   LL-Null:                       -636.30
Covariance Type:                 nonrobust   LLR p-value:                1.534e-228
============================================================================================
                               coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                    -26.6633   4.32e+06  -6.17e-06      1.000   -8.48e+06    8.48e+06
had_any_urinary_problem  -22.4015    3.3e+06  -6.79e-06      1.000   -6.47e+06    6.47e+06
============================================================================================
```

Fig.4.62 The summary of logistic regression for Single Variable model

The following graphical representations show the Receiver Operating Characteristic curve, Confusion matrix and Classification report for Logistic regression's **Single Variable**

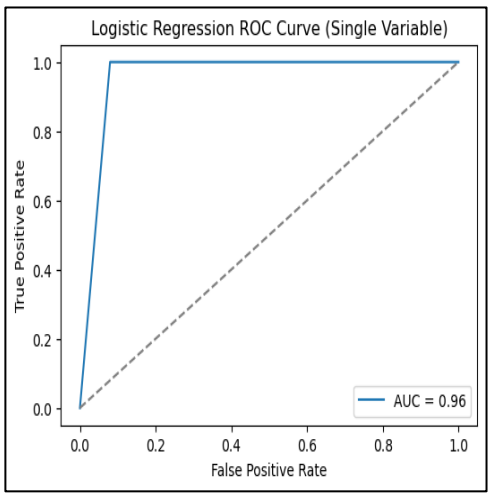


Fig 4.63 ROC curve for single Variable model of LR

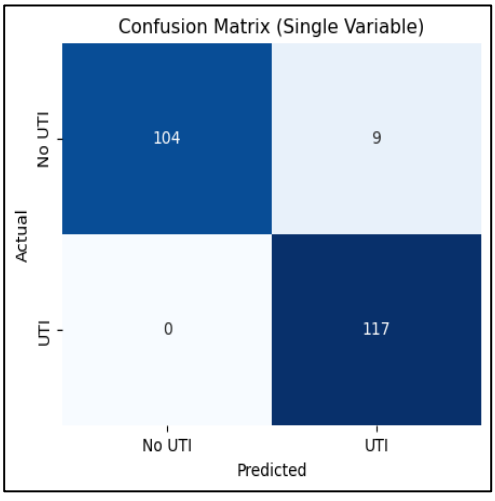


Fig 4.64 Confusion matrix for single Variable model of LR

```
Accuracy (Single Variable): 0.9609
False Positive Rate (Single Variable): 0.0796
Classification Report (Single Variable):

              precision    recall  f1-score   support

           0       1.00      0.92      0.96       113
           1       0.93      1.00      0.96       117

    accuracy                           0.96       230
   macro avg       0.96      0.96      0.96       230
weighted avg       0.96      0.96      0.96       230
```

Fig 4.65 Classification report for single Variable model of LR

Interpretation of Logistic Regression (Single Variable) Results

This Logistic Regression model with a single strong predictor performs well in classification, with an AUC of 0.96. The steep ROC curve suggests an excellent balance of sensitivity and specificity. The confusion matrix reveals a high True Positive (117) and True Negative (104) count, with no False Negatives, indicating that all positive cases have been detected. The classification report indicates exceptional accuracy (96.09%), flawless recall for the positive class (1.00), and outstanding F1-scores (0.96 in both classes). This shows that a single variable is a highly effective predictor, resulting in a promising and dependable model, which is especially useful in medical settings where recognizing all positive instances is crucial.

The advantages of employing only one predictor variable for logistic regression include simplicity and easy of interpretation. A model with a single predictor is simple to understand, allowing for a clear assessment of that variable's influence on the outcome's probability. It is also computationally efficient and less prone to overfitting, particularly when dealing with restricted data sets.

However, the disadvantages are enormous. A single variable model frequently fails to reflect the complex relationships in real-world data, resulting in underfitting and lower predicted accuracy. It ignores the potential influence of other relevant elements, which may result in biased or incomplete insights and restrict the model's practical value.

4.3.1.4 Logistic Regression for Excluded Single Variable

Regression Equation

log(p/(1-p)) = 0.0531 + (-0.3529 * Diabetes) + (-0.1807 * Hypertension) + (-0.3035 * Lookout_for_toilet) + (-0.2248 * Pad_protection_to_management) + (-0.2644 * Coffee) + (-0.1727 * Tea) + (0.0405 * Butter_milk) + (1.3439 * Liquid_consumption) + (-0.1442 * Menopause_attained)

McFadden's Pseudo R-squared: 0.0677

The following summary shows the Results of Logistic Regression for Excluded Single variable model.

```
=== Excluded Single Variable ===
                           Logit Regression Results
==============================================================================
Dep. Variable:     Urinary_Track_Infection   No. Observations:                  918
Model:                               Logit   Df Residuals:                      908
Method:                                MLE   Df Model:                            9
Date:                     Mon, 14 Apr 2025   Pseudo R-squ.:                0.06766
Time:                             07:04:42   Log-Likelihood:                -593.25
converged:                           False   LL-Null:                       -636.30
Covariance Type:                 nonrobust   LLR p-value:                 9.826e-15
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                            0.0531      7.507      0.007      0.994     -14.660      14.766
Diabetes                        -0.3529      0.084     -4.224      0.000      -0.517      -0.189
Hypertension                    -0.1807      0.077     -2.351      0.019      -0.331      -0.030
Lookout_for_toilet              -0.3035      0.205     -1.479      0.139      -0.706       0.099
Pad_protection_to_management    -0.2248      0.221     -1.018      0.309      -0.658       0.208
Coffee                          -0.2644      0.093     -2.850      0.004      -0.446      -0.083
Tea                             -0.1727      0.110     -1.573      0.116      -0.388       0.043
Butter_milk                      0.0405      0.079      0.513      0.608      -0.114       0.195
Liquid_consumption               1.3440     97.293      0.014      0.989    -189.347     192.035
Menopause_attained              -0.1442      0.072     -2.014      0.044      -0.284      -0.004
================================================================================================
```

Fig.4.66 The summary of logistic regression for Excluded Single Variable model

The following graphical representations show the Receiver Operating Characteristic curve, Confusion matrix and Classification report for Logistic regression's **Excluded Single Variable** model.

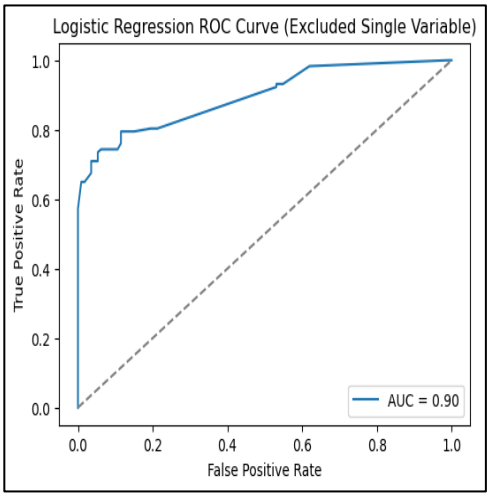


Fig 4.67 ROC curve for Excluded single Variable model of LR

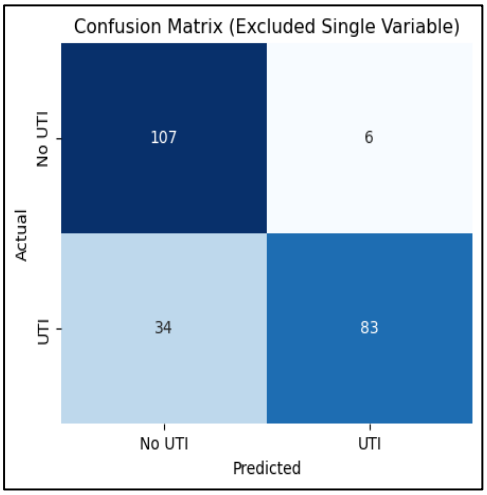


Fig 4.68 Confusion matrix for Excluded single Variable model of LR

```
Accuracy (Excluded Single Variable): 0.8261
False Positive Rate (Excluded Single Variable): 0.0531
Classification Report (Excluded Single Variable):

              precision    recall  f1-score   support

           0       0.76      0.95      0.84       113
           1       0.93      0.71      0.81       117

    accuracy                           0.83       230
   macro avg       0.85      0.83      0.82       230
weighted avg       0.85      0.83      0.82       230
```

Fig 4.69 Classification report for Excluded single Variable model of LR

Interpretation of Logistic Regression (Excluded Single Variable) Results

Excluding a single variable greatly reduces the Logistic Regression model's effectiveness. The AUC falls to 0.90 from 0.96, showing a decrease in discriminative capacity. The confusion matrix shows a significant increase in false negatives (34), resulting in a decrease in recall (0.71 for class 1). Accuracy drops to 0.8261, and F1-scores decrease in both classes.

This emphasizes the important role of the excluded variable in correctly detecting positive cases, which is especially concerning in medical applications such as UTI diagnosis, where missing positive cases might have serious effects. The analysis emphasizes the importance of include this variable in order to return the model to its original, higher performance.

Overall Performance Evaluation (Including Chi-Square Association Variables Model)

The "All Variables" and "Chi-Square Association Variables" models show near-perfect and very strong performance, respectively, indicating the value of incorporating all pertinent features and the Chi-Square feature selection method in identifying important predictors; the "Single Variable" model showed surprising performance, highlighting the importance of that single variable; crucially, removing this single variable resulted in a significant decline in performance, especially in recall, highlighting its significance; the "All Variables" and "Chi-Square" models are robust and reliable, making them highly desirable for medical applications because they reduce false negatives; the Chi-Square method is a very effective feature selection technique that produces a parsimonious but accurate model.

Overall Conclusion

With an emphasis on the importance of pertinent features, the "All Variables" and "Chi-Square Association Variables" models provide the best accuracy and dependability. Key variables were successfully discovered by chi-square selection, which produced performance that was almost identical to that of the "All Variables" model. While depending only on the most important variable produces a good but suboptimal model, excluding a critical variable significantly impairs model performance. The actual application and if the minor performance benefit is worth the additional complexity and computational expense of using all variables will determine which model is used.

Performance Metrics

These metrics assess how well each model performs in predicting UTI

- **Accuracy:** The overall proportion of correctly classified cases (both UTI and No UTI). A higher accuracy indicates better overall classification performance.
- **Precision:** The proportion of correctly predicted UTI cases out of all cases predicted as UTI. High precision means that when the model predicts someone has a UTI, it's usually correct.

- **Sensitivity (Recall):** The proportion of actual UTI cases that are correctly identified by the model. High sensitivity means the model is good at finding most of the people who truly have a UTI.
- **Specificity:** The proportion of actual No UTI cases that are correctly identified by the model. High specificity means the model is good at correctly identifying people who do not have a UTI.
- **Negative Predictive Value (NPV):** The proportion of correctly predicted No UTI cases out of all cases predicted as No UTI. High NPV means that when the model predicts someone does not have a UTI, it's usually correct.
- The **False Negative Rate (FNR)**, also known as the miss rate, is the proportion of actual positive cases that are incorrectly identified as negative by a classification model.
- **F1-score:** The harmonic mean of precision and sensitivity. It provides a balanced measure of a model's accuracy, especially when the classes are imbalanced. A higher F1-score indicates a better balance between precision and sensitivity.

Over-fitting - happens when a machine learning model learns the training data too well, including the noise and random fluctuations inherent in that dataset. This results in a model that performs incredibly well on training data but poorly on fresh, previously unseen data since it memorized the training examples rather than learning the underlying, generalizable patterns.

Under-fitting - occurs when a machine learning model is too simplistic to detect the underlying patterns in training data. It fails to understand the critical links between the input characteristics and the target variable, leading in poor performance not only on unseen data but also on training data. The model is overly generic and fails to reflect the intricacies of the data.

In machine learning, bias and variance are two major sources of mistake that impair a model's ability to generalize to new, previously unknown data. They represent a trade-off that data scientists must consider when developing effective models.

4.3.1.5 Logistic Regression for Comparing All Models

The following graphical representations show the Receiver Operating Characteristic curve and Performance Metrics table for comparing all **Logistic regression** models.

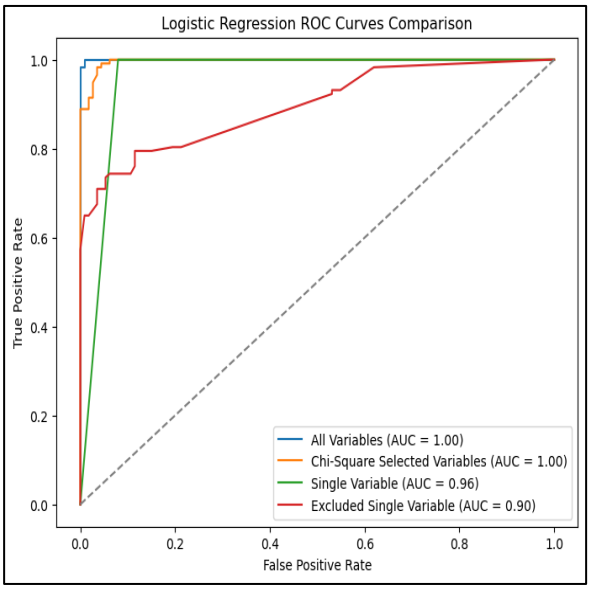


Fig 4.70 ROC curve for comparing all regression models

Logistic Regression						
Model Type	Accuracy	Precision	Sensitivity	Specificity	Negative Predictive Value	False Negative Rate
All Variables	99.2%	99.1%	99.1%	99.2%	99.2%	0.9%
Chi-square selected	96.1%	92.9%	100.0%	92.0%	100.0%	0.0%
Single Variable	96.1%	92.9%	100.0%	92.0%	100.0%	0.0%
Excluded Single Variable	81.9%	93.3%	69.7%	94.7%	74.8%	30.3%

Table 4.17 Performance metrics for logistic regression models

Interpretation of Overall Logistic Regression Results

The ROC curves visually confirm that the "All Variables" and "Chi-Square Selected Variables" models exhibit near-perfect discrimination (AUC=1.00), outperforming the "Single

Variable" (AUC=0.96) and significantly better than the "Excluded Single Variable" model (AUC=0.90). The performance metrics table reinforces this hierarchy: "All Variables" achieves >99% across all metrics; "Chi-Square Selected Variables" shows strong performance with perfect sensitivity (100%); "Single Variable" performs well but slightly lower; and "Excluded Single Variable" demonstrates significantly reduced sensitivity (69.7%) and overall performance. Clinically, the top two models are preferred due to high accuracy and sensitivity, while the excluded variable model is unsuitable due to high false negatives.

The Chi-Square method effectively selected key features, yielding near-optimal performance comparable to using all variables. The excluded and single variables were both shown to be very important predictors. This observed hierarchy in model effectiveness is quantitatively supported by the performance metrics table that follows: the "All Variables" model achieves performance above 99% across all assessed metrics, indicating exceptional predictive power; the "Chi-Square Selected Variables" model performs well with the noteworthy achievement of perfect sensitivity (100%), indicating its ability to correctly identify all positive cases; the "Single Variable" model performs well but shows slightly lower performance than the top two; and the "Excluded Single Variable" model shows a significant decrease in sensitivity (69.7%), suggesting a higher rate of false negatives and, ultimately, a compromised overall performance.

The top two models, "All Variables" and "Chi-Square Selected Variables," are unquestionably preferred from a clinical standpoint because of their high levels of sensitivity and accuracy, which reduce the possibility of misclassification. In contrast, the "Excluded Single Variable" model is considered unsuitable for practical application because of its unacceptable high rate of false negatives, which could result in missed diagnoses. A model with near-optimal performance that is very comparable to the model using all available variables is produced by the Chi-Square feature selection method, which notably seems to have successfully identified the most relevant predictor variables. This suggests an effective and efficient approach to feature selection. It's interesting to note that both the single variable used in the "Single Variable" model and the single variable excluded in the "Excluded Single Variable" model were individually demonstrated to be highly significant predictors of the outcome, despite their lower overall model performance when used in isolation or when one was excluded.

4.3.2 RANDOM FOREST

Random Forest is a powerful, non-parametric ensemble learning algorithm that is mostly used for classification and regression. It works by creating a large number of decision trees and aggregating their predictions to get a final output. This methodology takes advantage of the bootstrap aggregation (bagging) technique and random feature selection, resulting in lower variance and better generalization.

4.3.2.1 Random forest for All Variables

The following graph is one random decision tree for the Random Forest's All Variables model.

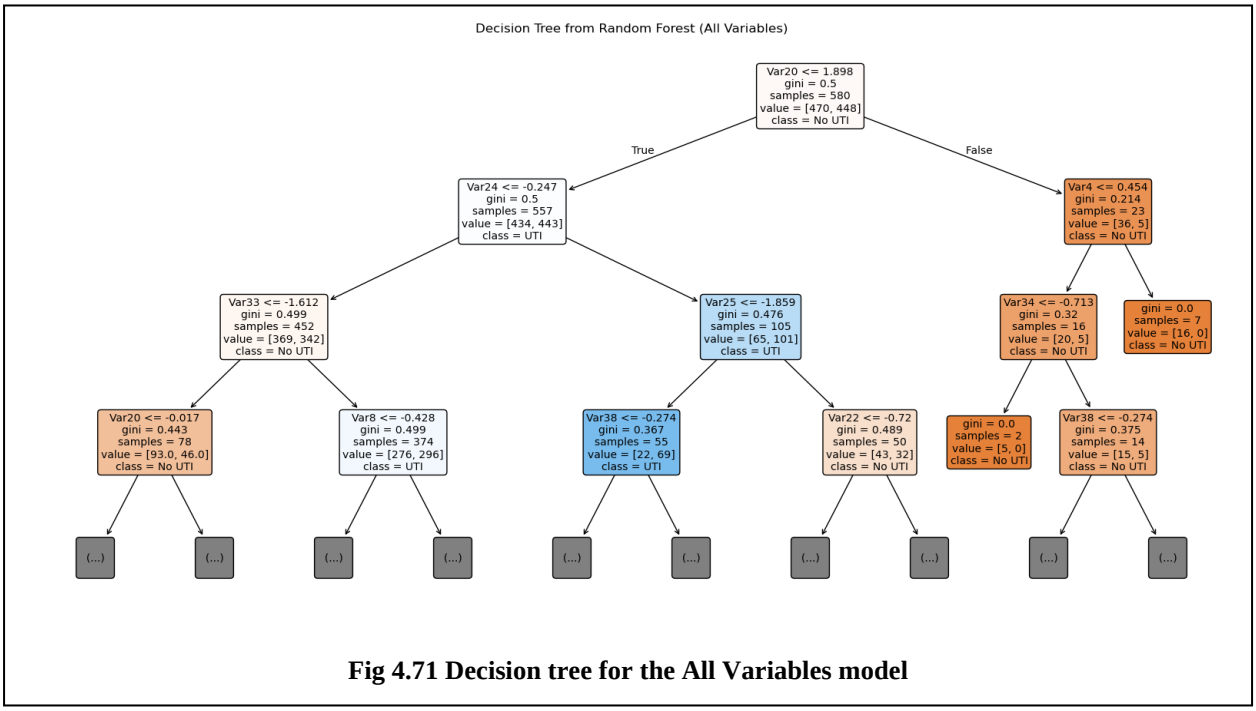


Fig 4.71 Decision tree for the All Variables model

Interpretation of Decision tree for all Variables

This decision tree illustrates how the model predicts UTI status based on sequential splits of various variables and their thresholds. It was generated using a Random Forest employing all variables. In order to arrive at a classification at the terminal (leaf) nodes, each node provides information about a decision rule, the Gini impurity, the number of samples that reach that node, and the distribution of UTI/No UTI cases.

The following graphical representations show the Receiver Operating Characteristic curve, Confusion matrix and Classification report for Random forest's **All Variables** model.

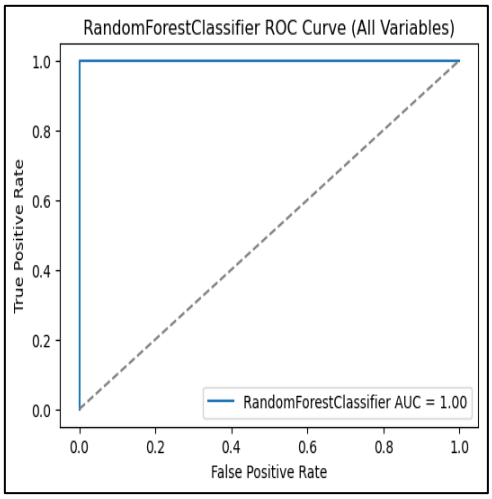


Fig 4.72 ROC curve for all variables model of RF

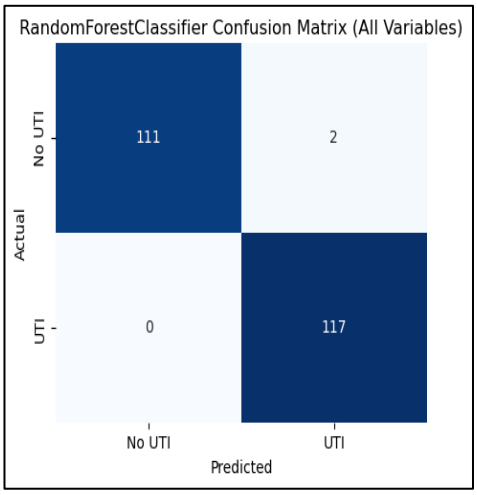


Fig 4.73 Confusion matrix for all variables model of RF

```
Accuracy (All Variables): 0.9913
False Positive Rate (All Variables): 0.0177
Classification Report (All Variables):

              precision    recall  f1-score   support

           0       1.00      0.98      0.99       113
           1       0.98      1.00      0.99       117

    accuracy                           0.99       230
   macro avg       0.99      0.99      0.99       230
weighted avg       0.99      0.99      0.99       230
```

Fig 4.74 Classification report for all variables model of RF

Interpretation of Random Forest (All Variables) Results

The Random Forest model, like the "All Variables" Logistic Regression, achieves perfect discrimination (AUC=1.00) and high accuracy (0.9913). The confusion matrix shows excellent results with very few errors (TP=117, TN=111, FP=2, FN=0), achieving perfect recall for UTI. The classification report confirms near-perfect precision, recall, and F1-scores for both classes. Compared to Logistic Regression (All Variables), the performance is remarkably similar, with Random Forest having a slightly higher False Positive Rate (0.0177 vs. 0.0088) but zero False Negatives compared to one in Logistic Regression. Both models are exceptional, and the choice between them might depend on interpretability and computational cost.

4.3.2.2 Random forest for Chi-Square Selected Variables

The following graph is one random decision tree for the Random Forest's Chi-Square Selected Variables model.

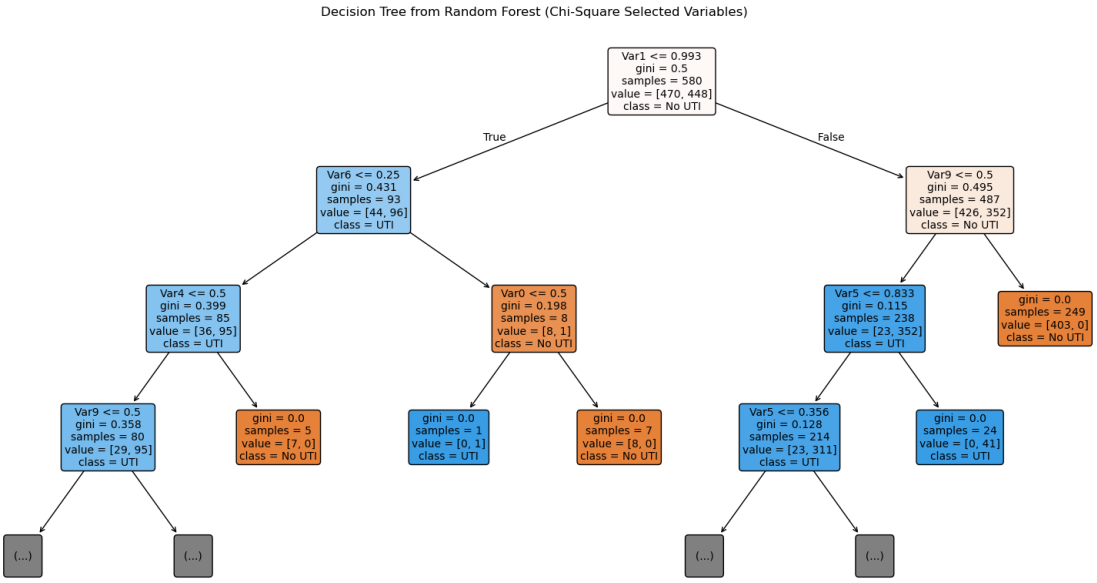


Fig 4.75 Decision tree for the Chi-Square Selected Variables model

Interpretation decision tree for the Chi-Square Selected Variables

This decision tree demonstrates the classification process for UTI prediction using sequential binary splits depending on particular variable thresholds. It was constructed using variables chosen by Chi-Square testing within a Random Forest. The splitting rule, Gini impurity, number of samples, UTI/No UTI case distribution, and predicted class are all shown on each node. The hierarchical decision-making process based on the most informative Chi-Square-selected features is revealed by the tree structure. In a Random Forest, a decision tree is a base learner, meaning it's one of the individual predictive models that make up the ensemble. Each decision tree in the Random Forest is trained on a random subset of the original data (with replacement, known as bootstrapping) and a random subset of the features at each split. This randomness ensures that the individual trees are diverse and less prone to overfitting the specific training data. The final prediction of the Random Forest is typically made by aggregating the predictions of all the individual decision trees.

The following graphical representations show the ROC curve, Confusion matrix and Classification report for Random forest's **Chi-Square Selected Variables** model.

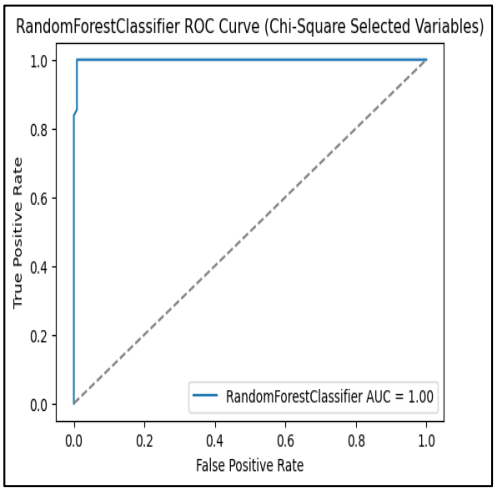


Fig 4.76 ROC curve for Chi-Square Selected Variables model of RF

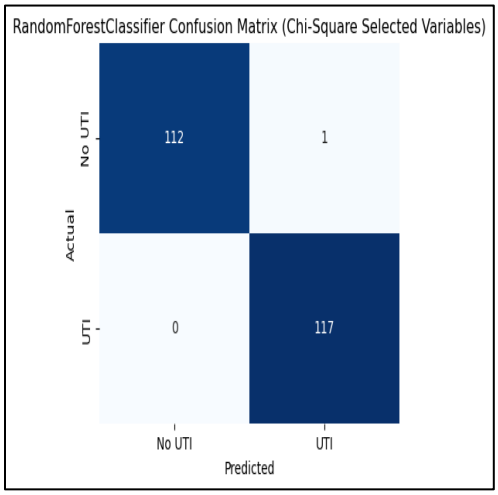


Fig 4.77 Confusion matrix for Chi-Square Selected Variables model of RF

```
Accuracy (Chi-Square Selected Variables): 0.9957
False Positive Rate (Chi-Square Selected Variables): 0.0088
Classification Report (Chi-Square Selected Variables):

              precision    recall  f1-score   support

           0       1.00      0.99      1.00       113
           1       0.99      1.00      1.00       117

    accuracy                           1.00       230
   macro avg       1.00      1.00      1.00       230
weighted avg       1.00      1.00      1.00       230
```

Fig 4.78 Classification report for Chi-Square Selected Variables model of RF

Interpretation of Random Forest (Chi-Square Selected Variables) Results

The Random Forest model performs exceptionally well (Accuracy=0.9957, F1-scores=1.00) and achieves perfect discrimination (AUC=1.00) when employing Chi-Square chosen variables. It outperforms the Logistic Regression (Chi-Square) model in accuracy and F1-scores, and it matches its perfect recall. The efficiency of Chi-Square in choosing the most informative characteristics for this intricate model is demonstrated by the fact that this Random Forest model performs on par with or marginally better than the Random Forest model employing all variables. For this classification problem, Random Forest and Logistic Regression both perform exceptionally well when paired with good feature selection; the decision between them is based on variables such as interpretability and computational cost.

4.3.2.3 Random forest for Single Variable

The following graph is one random decision tree for the Random Forest's Single Variable model.

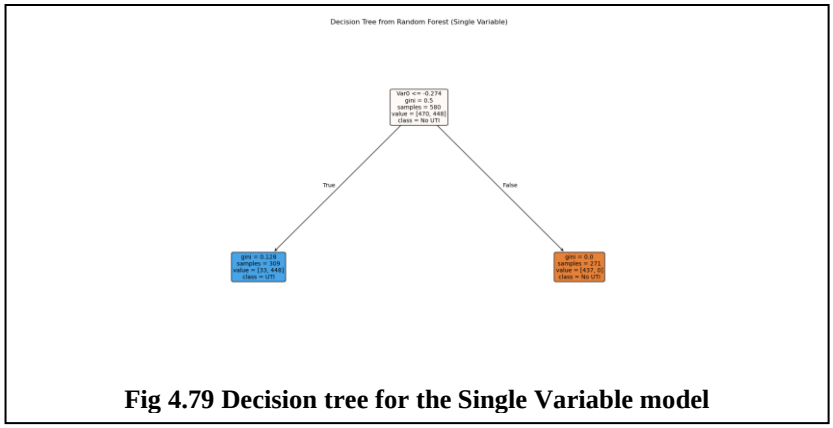


Fig 4.79 Decision tree for the Single Variable model

Interpretation decision tree for the Single Variable

"UTI" or "No UTI" are the predictions of this simplified decision tree, which is derived from a Random Forest using only one variable. It classifies UTI status according to whether predictive variable is less than or equal to -0.274.

The following graphical representations show the Receiver Operating Characteristic curve, Confusion matrix and Classification report for Random forest's **Single Variable** model.

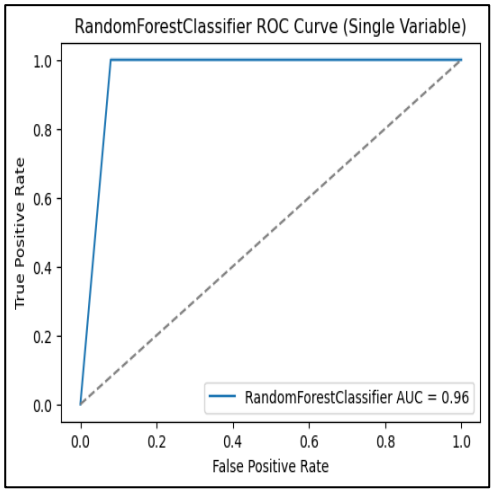


Fig 4.80 ROC curve for single Variable model of RF

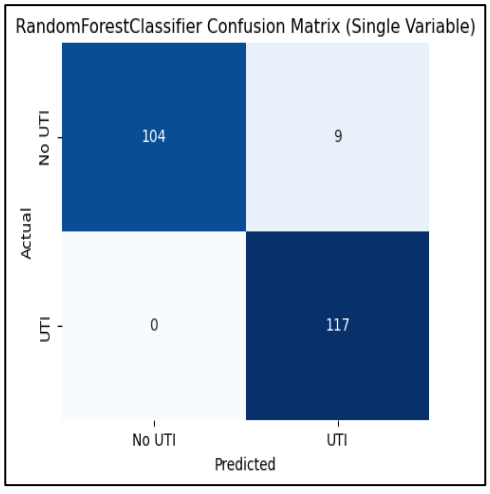


Fig 4.81 Confusion matrix for single Variable model of RF

```
Accuracy (Single Variable): 0.9609
False Positive Rate (Single Variable): 0.0796
Classification Report (Single Variable):

              precision    recall  f1-score   support

           0       1.00      0.92      0.96       113
           1       0.93      1.00      0.96       117

    accuracy                           0.96       230
   macro avg       0.96      0.96      0.96       230
weighted avg       0.96      0.96      0.96       230
```

Fig 4.82 Classification report for single Variable model of RF

Interpretation of Random Forest (Single Variable) Results

Strong UTI classification is demonstrated by a Random Forest model with a single variable (AUC=0.96, Accuracy=96.09%, F1=0.96), which achieves 100% recall for UTI with 9 false positives. Its performance is noteworthy since it is identical to that of a Logistic Regression model with the same variable for all measures.

This parity implies that both linear and non-linear models capture the predictive potential of the single variable equally well, suggesting that the connection may be mostly linear or that Random Forest's complexity is not advantageous in this case. The significant performance emphasizes how crucial this one predictor is.

A decision tree is a supervised machine learning system that represents decisions using a tree-like structure. It splits the dataset recursively into smaller subsets based on the values of the input features, with the goal of creating homogeneous subsets with regard to the target variable. Each internal node in the tree represents an attribute test, each branch reflects the test's result, and each leaf node represents a class label (for classification) or predicted value (for regression). The path from the root to the leaf node reflects the model's decision-making process to arrive at a forecast. Decision trees are intuitive and simple to grasp since the decision-making process is plainly depicted.

4.3.2.4 Random forest for Excluded Single Variable

The following graph is one random decision tree for the Random Forest's Excluded Single Variable model.

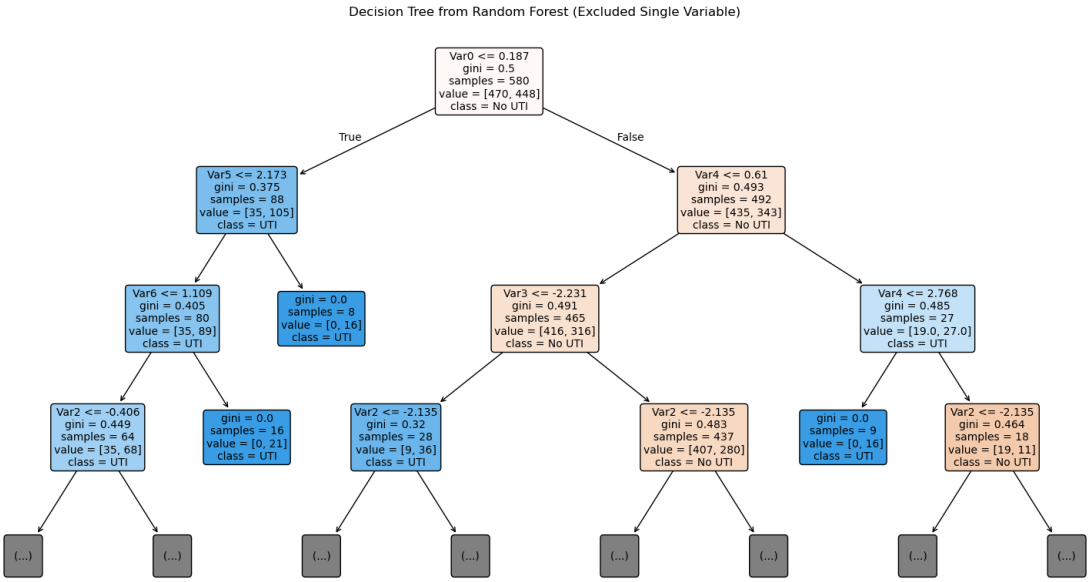


Fig 4.83 Decision tree for the Excluded Single Variables model

Interpretation decision tree for the Excluded Single Variables

Each node displays the decision rule, impurity, sample distribution, and predicted class, demonstrating the model's classification process without the influence of the excluded predictor. This decision tree, which is a component of a Random Forest in which one significant variable was excluded, predicts UTI status through sequential splits based on the remaining variables and their thresholds. Starting at the root, each node tests a certain attribute, with branches representing the results of those tests. The Gini impurity and distribution of UTI/No UTI cases are displayed at each decision point, illustrating the subsets' homogeneity and class composition. The tree branches until it reaches the leaf nodes at the bottom, where a final classification (UTI or No UTI) is assigned based on the majority class of the samples that have taken the decision path up to that point. This image shows how the model adjusts its prediction strategy by using the remaining attributes after removing a significant predictor.

The following graphical representations show the Receiver Operating Characteristic curve, Confusion matrix and Classification report for Random forest's **Excluded Single Variable** model.

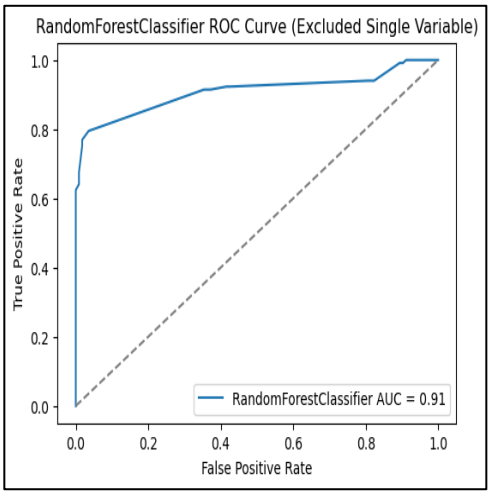


Fig 4.84 ROC curve for Excluded single Variable model of RF

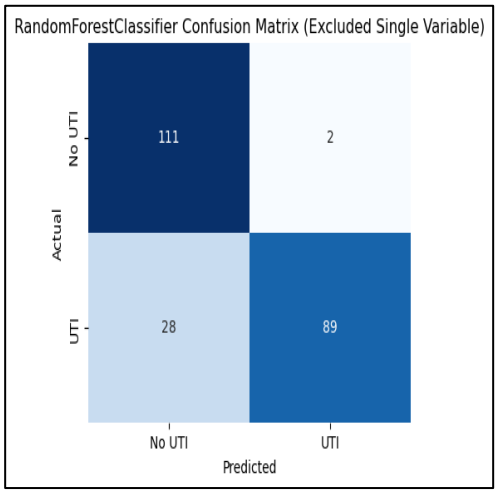


Fig 4.85 Confusion matrix for Excluded single Variable model of RF

```
Accuracy (Excluded Single Variable): 0.8696
False Positive Rate (Excluded Single Variable): 0.0177
Classification Report (Excluded Single Variable):

              precision    recall  f1-score   support

           0       0.80      0.98      0.88       113
           1       0.98      0.76      0.86       117

    accuracy                           0.87       230
   macro avg       0.89      0.87      0.87       230
weighted avg       0.89      0.87      0.87       230
```

Fig 4.86 Classification report for Excluded single Variable model of RF

Interpretation of Random Forest (Excluded Single Variable) Results

The Random Forest model performs worse when the single most significant variable is removed (AUC=0.91, Accuracy=0.8696). Recall for UTIs decreases considerably (0.76), while it is still rather high, leading to more False Negatives (28). Random Forest performs more robustly even in the absence of the critical predictor, as evidenced by slightly greater accuracy, recall, and a lower False Positive Rate for UTI when compared to Logistic Regression with the same exclusion. Both Random Forest and Logistic Regression suffer when the crucial single variable is removed. But Random Forest has more robustness, retaining higher recall for UTI (0.76), and

accuracy (0.8696 vs. 0.819). Additionally, Random Forest shows a higher True Negative count and a lower False Positive Rate, indicating greater resilience in the absence of a crucial predictor. Random Forest performs better than Logistic Regression in these situations, however it is still not ideal without this variable. This emphasizes how crucial the excluded variable is to both models' maximum accuracy.

Overall Evaluation of Random Forest Classification Method

The Random Forest algorithm continuously exhibits strong and excellent performance across a range of input feature subsets in this classification task that predicts a binary outcome. Notably, the Random Forest model produces almost flawless classification results—even showing a slight performance advantage over the Logistic Regression model—in scenarios that involve both the full feature set and a carefully chosen subset determined by Chi-Square statistical tests. Furthermore, the Random Forest showcases remarkable resilience when trained on a severely limited feature set consisting of only a single variable, and it also proves more robust than Logistic Regression when a crucial, highly predictive variable is intentionally excluded from the training data.

This overall robustness highlights Random Forest's capacity to manage complex datasets and possibly lessen the impact of missing or less informative data. This is probably due to Random Forest's ensemble learning feature, which aggregates predictions from several individual decision trees. Because of its innate flexibility and resilience, Random Forest is a strong and adaptable classification method. It is especially useful when combined with efficient feature selection techniques like Chi-Square, which can find the most pertinent predictors without significantly compromising predictive accuracy.

Compared to the linear decision boundary that Logistic Regression learns, Random Forest may be more appropriate for this specific classification task due to its capacity to capture complex feature interactions and non-linear relationships. This is indicated by the marginal outperformance of Logistic Regression in some of these scenarios. This emphasizes how crucial it is to test out several modeling algorithms in order to determine which one best reflects the underlying patterns in the data.

4.3.2.5 Random forest for comparing All Models

The following graphical representations show the Receiver Operating Characteristic curve and Performance Metrics table for comparing all **Random forest** models.

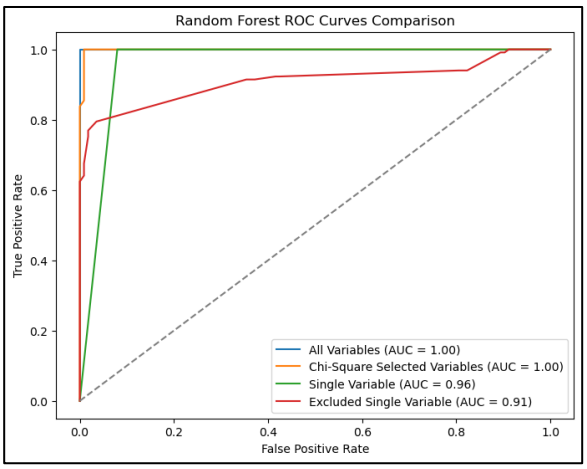


Fig 4.87 ROC curve for comparing all Random Forest models

Random Forest						
Model Type	Accuracy	Precision	Sensitivity	Specificity	Negative Predictive Value	False Negative Rate
All Variables	99.1%	98.3%	100.0%	98.2%	100.0%	0.0%
Chi-square selected	99.6%	99.2%	100.0%	99.1%	100.0%	0.0%
Single Variable	96.1%	92.9%	100.0%	92.0%	100.0%	0.0%
Excluded Single Variable	87.0%	97.8%	76.1%	98.2%	79.9%	23.9%

Table 4.18 Performance metrics for Random forest models

Interpretation of Overall Random Forest Results

With Chi-Square selected variables, Random Forest performs best (Accuracy=99.6%, AUC=1.00), even marginally outperforming the "All Variables" model (Accuracy~99.1%, AUC=1.00). Since Random Forest exhibits overall robustness and the Chi-Square approach is very effective for feature selection, the top two models are well suited for medical applications because of their high sensitivity and accuracy.

4.3.3 XG-BOOST (Extreme Gradient Boosting)

XG-Boost is a powerful and scalable implementation of gradient-boosted decision trees that is commonly used for structured (tabular) data prediction applications. It belongs to the boosting family of ensemble methods, which combines weak learners (usually shallow trees) in a sequential fashion to produce a robust predictive model. XG-Boost is designed for performance and efficiency, with speed and accuracy tuned, making it a top choice for machine learning contests and real-world applications.

4.3.3.1 XG-Boost for All Variables

The following graphical representations show the Receiver Operating Characteristic curve, Confusion matrix and Classification report for XG-Boost's **All Variables** model.

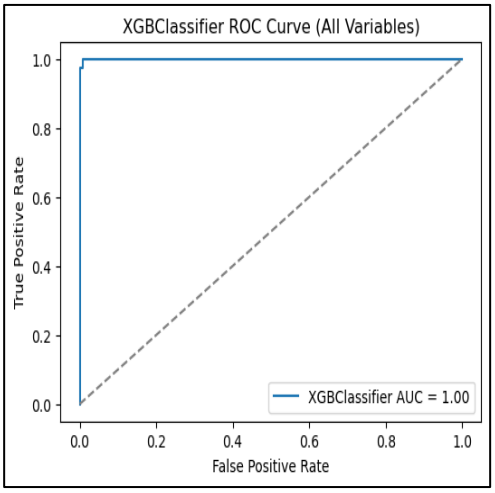


Fig 4.88 ROC curve for all variables model of XG

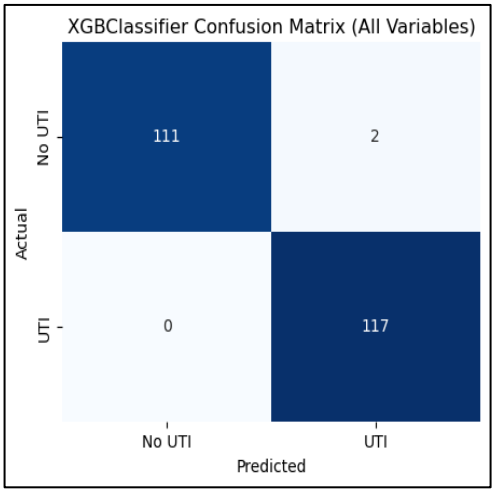


Fig 4.89 Confusion matrix for all variables model of XG

```
Accuracy (All Variables): 0.9913
False Positive Rate (All Variables): 0.0177
Classification Report (All Variables):

              precision    recall  f1-score   support

           0       1.00      0.98      0.99       113
           1       0.98      1.00      0.99       117

    accuracy                           0.99       230
   macro avg       0.99      0.99      0.99       230
weighted avg       0.99      0.99      0.99       230
```

Fig 4.90 Classification report for all variables model of XG

The following graphical representation shows which variables helps to classify the target variable by the SHAP and Feature Importance plot for all variables model.

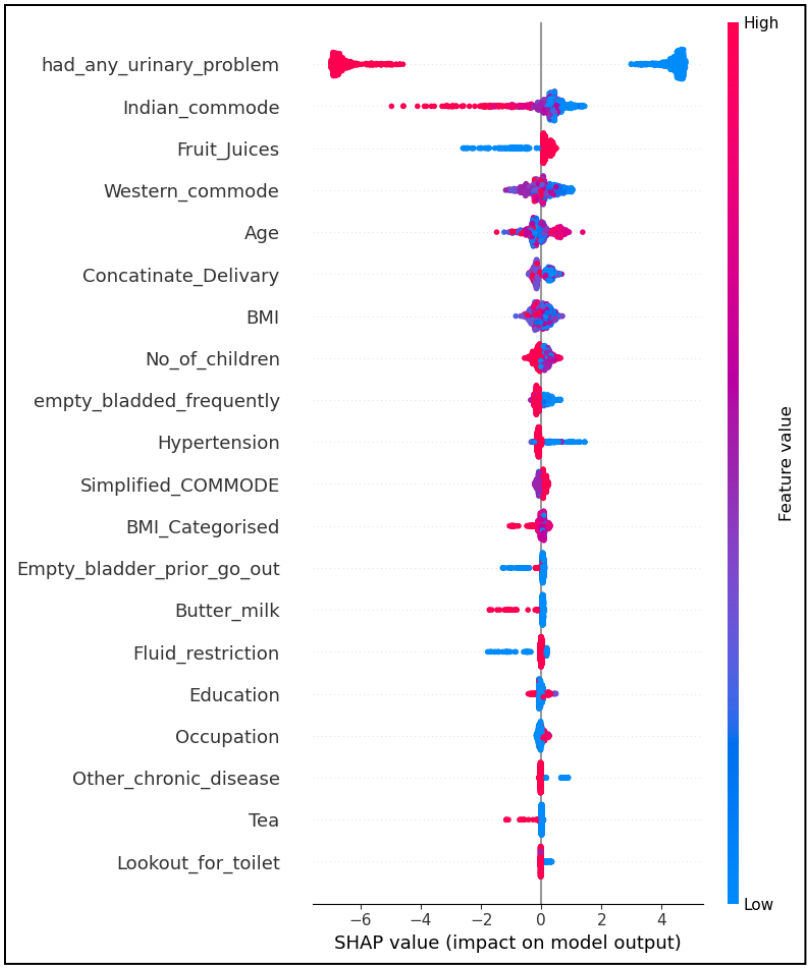


Fig 4.91 SHAP plot for all variables model

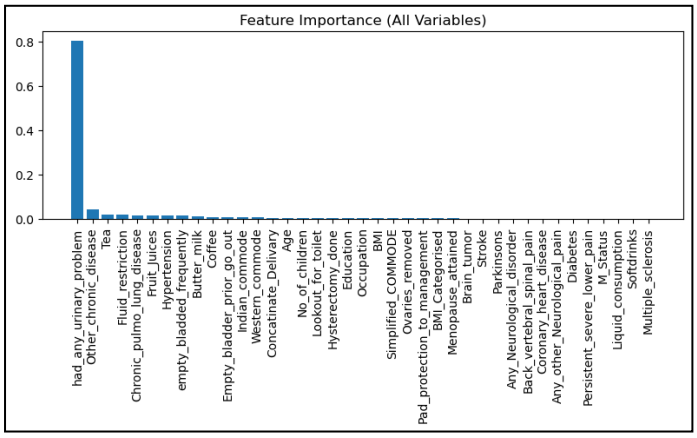


Fig 4.92 Feature Importance for all variables model

Interpretation of XG-Boost (All Variables) Results

With all variables included, XG-Boost achieves near-perfect classification (AUC=1.00, Accuracy=0.9913, F1=0.99), closely matching Logistic Regression and reflecting Random Forest's outstanding performance. Excellent accuracy and recall are displayed by all three models. The only modest variation is that XG-Boost has a slightly greater False Positive Rate than Logistic Regression. The decision between these top-performing models may depend on the particular application requirements, interpretability (Logistic Regression is preferable), and computing cost (XG-Boost is frequently efficient).

This SHAP summary graph demonstrates that "had_any_urinary_problem" is the most influential factor positively predicting Urinary Tract Infections (UTIs), whilst the type of commode used shows a more nuanced association with potential subtle opposite effects. Higher age increases the risk of a UTI diagnosis, although fruit juice consumption has a less obvious impact. Overall, the plot ranks feature relevance and shows the direction and magnitude of each variable's contribution to the model's output for individual predictions, providing useful information on the model's decision-making process for UTI classification.

The feature importance plot shows a clear hierarchy of predictive power, with "had_any_urinary_problem" emerging as the single most important variable for the model's predictions. A tiny fraction of other features, such as "other_chronic_disease," "tea," and "fluid_restriction," are modest, while the vast majority of the remaining variables have a negligible impact on the model's decision-making process when all features are included. The dominance of a single predictor implies a strong individual link with the target outcome, whereas the low relevance of several others indicates potential redundancy or limited direct influence in this model. In this model, "had_any_urinary_problem" is the most significant predictive predictor for identifying the presence of a UTI. Compared to every other feature, it has a lot higher weight in the model's decision-making process.

This suggests that the best way for the model to determine if a patient has a UTI at the moment is to look at their past history of urinary problems.

4.3.3.2 XG-Boost for Chi-Square Selected Variables

The following graphical representations show the Receiver Operating Characteristic curve, Confusion matrix and Classification report for XG-Boost's **Chi-Square Selected Variables** model.

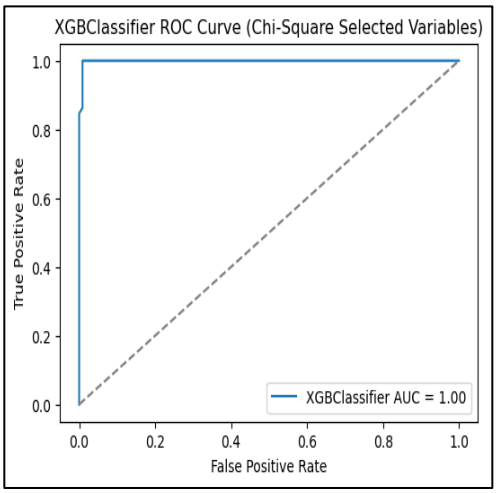


Fig 4.93 ROC curve for Chi-Square Selected Variables model of XG

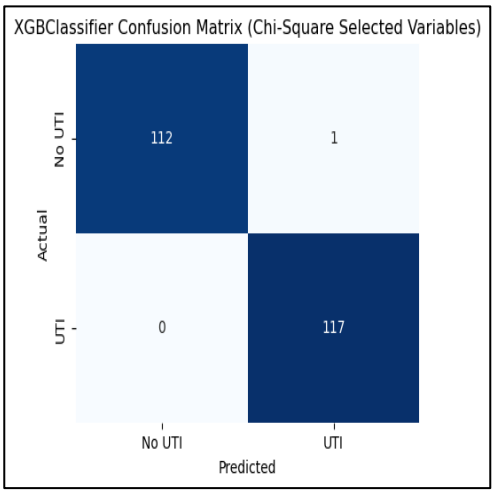


Fig 4.94 Confusion matrix for Chi-Square Selected Variables model of XG

```
Accuracy (Chi-Square Selected Variables): 0.9957
False Positive Rate (Chi-Square Selected Variables): 0.0088
Classification Report (Chi-Square Selected Variables):

              precision    recall  f1-score   support

           0       1.00      0.99      1.00       113
           1       0.99      1.00      1.00       117

    accuracy                           1.00       230
   macro avg       1.00      1.00      1.00       230
weighted avg       1.00      1.00      1.00       230
```

Fig 4.95 Classification report for Chi-Square Selected Variables model of XG

The following graphical representation shows which variables helps to classify the target variable by the SHAP and Feature Importance plot for chi-square selected variables model.

In this model also "had_any_urinary_problem" is the most significant predictive predictor for identifying the presence of a UTI. Compared to every other feature, it has a lot higher weight in the model's decision-making process.

This suggests that the best way for the model to determine if a patient has a UTI at the moment is to look at their past history of urinary problems.

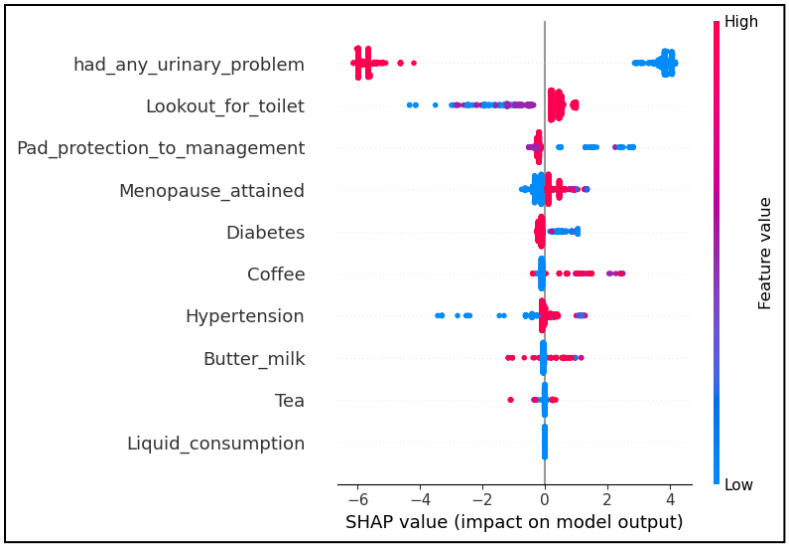


Fig 4.96 SHAP plot for Chi-Square Selected variables model

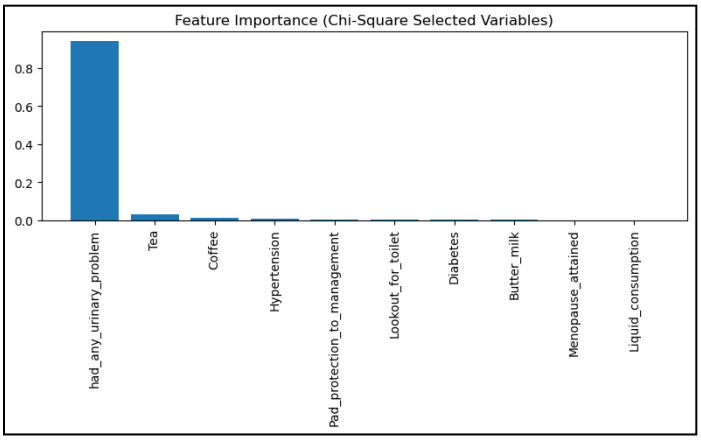


Fig 4.97 Feature Importance for Chi-Square Selected variables model

Interpretation of XG-Boost (Chi-Square Selected Variables) Results

With Chi-Square-selected variables, XG-Boost performs exceptionally well (Accuracy=0.9957, F1=1.00) and achieves perfect discrimination (AUC=1.00), which is the same as Random Forest with the same variables. While all three preserve the same low False Positive Rate and flawless recall, it performs marginally better in accuracy and F1-scores than Logistic Regression (Chi-Square). Key features for every model are efficiently identified via the Chi-Square approach. The decision between XG-Boost, Random Forest, and Logistic Regression may be influenced by the requirements of a particular application, interpretability, and computing cost.

4.3.3.3 XG-Boost for Single Variable

The following graphical representations show the Receiver Operating Characteristic curve, Confusion matrix and Classification report for XG-Boost's **Single Variable** model.

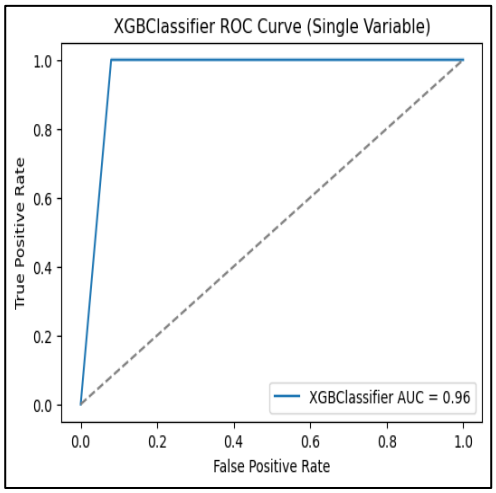


Fig 4.98 ROC curve for single Variable model of XG

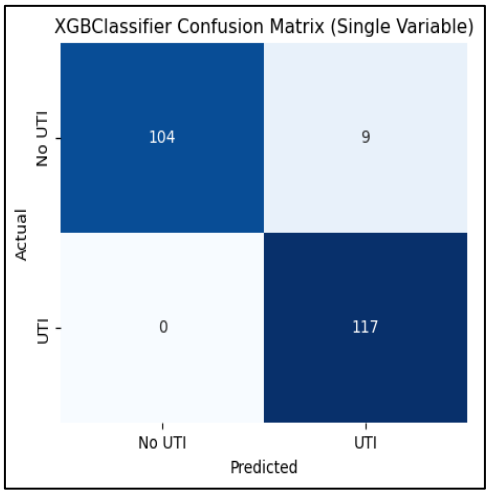


Fig 4.99 Confusion matrix for single Variable model of XG

```
Accuracy (Single Variable): 0.9609
False Positive Rate (Single Variable): 0.0796
Classification Report (Single Variable):

              precision    recall  f1-score   support

           0       1.00      0.92      0.96       113
           1       0.93      1.00      0.96       117

    accuracy                           0.96       230
   macro avg       0.96      0.96      0.96       230
weighted avg       0.96      0.96      0.96       230
```

Fig 4.100 Classification report for single Variable model of XG

The following graphical representation shows which variables helps to classify the target variable by the SHAP and Feature Importance plot for Single variable model.

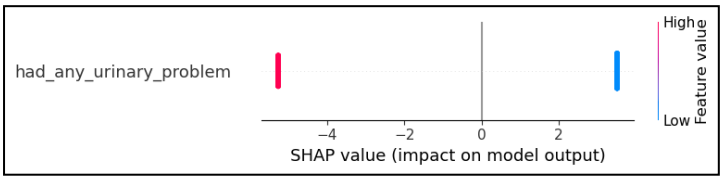


Fig 4.101 SHAP plot for Single variable model

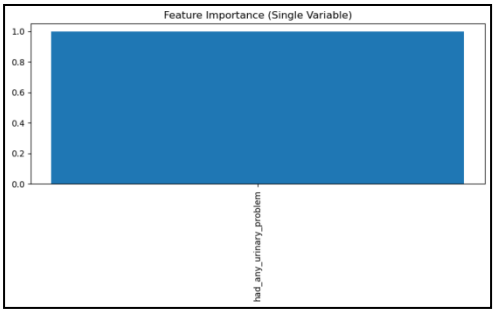


Fig 4.102 Feature Importance for Single variable model

Interpretation of XG-Boost (Single Variable) Results

With the same single variable, XG-Boost, Random Forest, and Logistic Regression all perform exceptionally well (AUC=0.96, Accuracy=0.9609, F1=0.96, perfect UTI recall). This suggests that the single variable is a powerful predictor and that its information is equally effectively captured by the three models. The selection of this model may be influenced by the needs of a particular application, interpretability (in favor of logistic regression), or processing expense.

Simpler Models Can Be Competitive - When there is a relatively clear relationship between the single predictor and the outcome, simpler models, such as logistic regression, can frequently perform on par with more sophisticated models. With limited input, non-linear models may not effectively exploit their extra complexity.

Importance of the Variable - The degree to which a single variable is predictive and highly associated with the desired result will have a significant impact on the overall performance level. Regardless of the complexity of the model, a weak predictor will produce poor performance, whereas a highly informative variable will produce good performance across all methods.

When limited to a single predictor, the predictor's intrinsic strength takes center stage in determining how well any classification approach performs. How well the algorithm captures the relationship in that one-dimensional input space is more important than how sophisticated the algorithm is.

4.3.3.4 XG-Boost for Excluded Single Variable

The following graphical representations show the Receiver Operating Characteristic curve, Confusion matrix and Classification report for XG-Boost's **Excluded Single Variable** model.

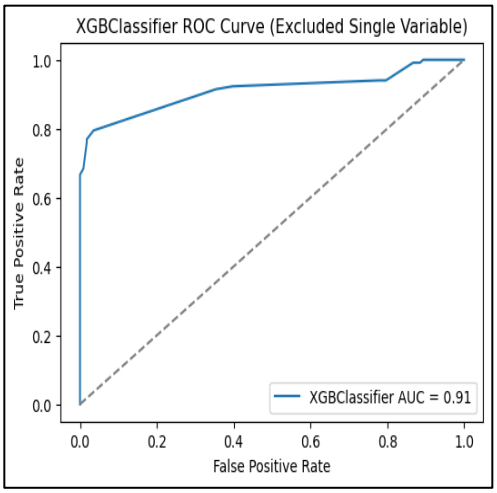


Fig 4.103 ROC curve for Excluded single Variable model of XG

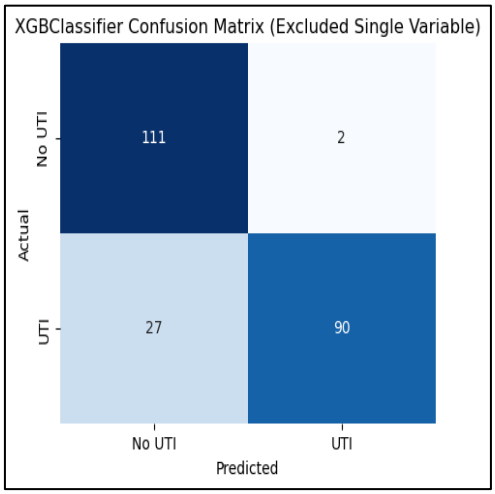


Fig 4.104 Confusion matrix for Excluded single Variable model of XG

```
Accuracy (Excluded Single Variable): 0.8739
False Positive Rate (Excluded Single Variable): 0.0177
Classification Report (Excluded Single Variable):

              precision    recall  f1-score   support

           0       0.80      0.98      0.88       113
           1       0.98      0.77      0.86       117

    accuracy                           0.87       230
   macro avg       0.89      0.88      0.87       230
weighted avg       0.89      0.87      0.87       230
```

Fig 4.105 Classification report for Excluded single Variable model of XG

The following graphical representation shows which variables helps to classify the target variable by the SHAP and Feature Importance plot for Excluded Single variable model.

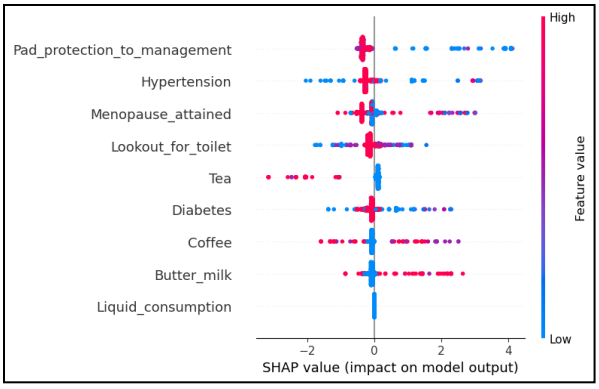


Fig 4.106 SHAP plot for Excluded Single variable model

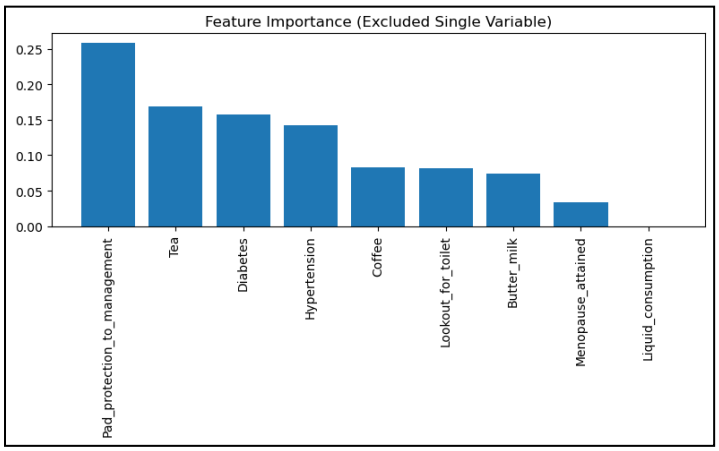


Fig 4.107 Feature Importance for Excluded Single variable model

Interpretation of XG-Boost (Excluded Single Variable) Results

Like Random Forest, XG-Boost suffers when the primary variable is removed (AUC=0.91, Accuracy=0.8739, Recall=0.77). When this variable is removed, both perform better than Logistic Regression in terms of accuracy, recall for UTI, and False Positive Rate, indicating increased robustness. The significance of the excluded variable for the accuracy of all models is highlighted by the fact that, although not optimal, XG-Boost and Random Forest manage the missing predictor better than Logistic Regression. This feature priority plot shows how removing the most important predictor causes the model's attention to shift to other variables. The loss of predictive ability resulting from the lack of the extremely significant "had_any_urinary_problem" is reflected in the overall lower importance ratings, even though "Pad_protection_to_management" emerges as the most significant of the remaining.

4.3.3.5 XG-Boost for comparing All Models

The following graphical representations show the Receiver Operating Characteristic curve and Performance Metrics table for comparing all **XG-Boost** models.

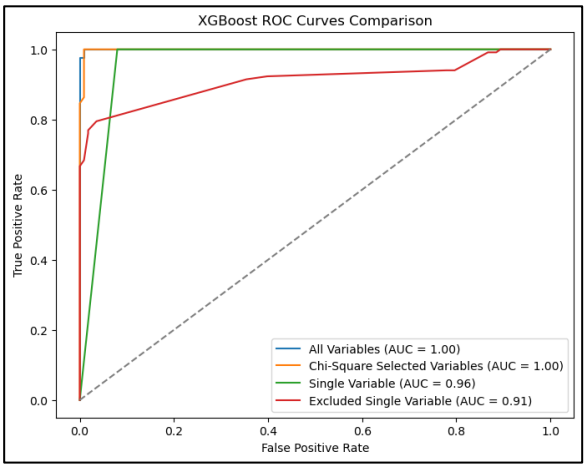


Fig 4.108 ROC curve for comparing all XG-Boost models

XG-Boost						
Model Type	Accuracy	Precision	Sensitivity	Specificity	Negative Predictive Value	False Negative Rate
All Variables	99.1%	98.3%	100.0%	98.2%	100.0%	0.0%
Chi-square selected	99.6%	99.2%	100.0%	99.1%	100.0%	0.0%
Single Variable	96.1%	92.9%	100.0%	92.0%	100.0%	0.0%
Excluded Single Variable	87.4%	97.8%	76.9%	98.2%	80.4%	23.1%

Table 4.19 Performance metrics for XG-Boost models

Interpretation

XG-Boost performs best with Chi-Square selected variables (Accuracy=99.6%, AUC=1.00), slightly outperforming the "All Variables" model (Accuracy<99.1%, AUC=1.00). Both outperform the "Single Variable" model (96.1% accuracy, 0.96 AUC). Excluding a critical variable considerably affects performance (Accuracy ~87.4%, AUC=0.91), although XG-Boost exhibits higher resilience than Logistic Regression and similar resilience to Random Forest.

ROC curves visually confirm the hierarchy. The Chi-Square technique is quite effective. The top two models are favored for medical applications due to their high accuracy and sensitivity, emphasizing the significance of feature selection and inclusion.

XG-Boost's "All Variables" model has near-perfect accuracy (99.1%) and sensitivity. The "Chi-Square Selected Variables" model has the highest accuracy (99.6%, perfect sensitivity). The "Single Variable" model is comparable to Logistic Regression and Random Forest (96.1% accuracy, excellent sensitivity, poorer specificity). Excluding a critical variable greatly affects accuracy (87.4%) and sensitivity (76.9%), placing it between Random Forest and Logistic Regression in performance. The Chi-squared selected variable model is the most accurate.

Overall Evaluation of XG-Boost Classification Method

In this regard, XG-Boost is a particularly effective classification algorithm, consistently attaining near-perfect accuracy when trained on the entire set of predictor variables and producing outstanding outcomes with a more condensed set of features chosen using the Chi-Square method. XG-Boost outperforms Logistic Regression in each of these cases and achieves an accuracy level that is on par with the very reliable Random Forest. Additionally, when restricted to training on a single predictive variable, XG-Boost exhibits excellent resilience, which is consistent with the robustness seen in Random Forest and Logistic Regression under comparable conditions.

Notably, when an important, influential variable is purposefully left out of the feature set, XG-Boost outperforms Logistic Regression in terms of resilience. Its robustness and excellent performance across a range of feature set configurations are largely due to the underlying boosting process, which iteratively creates an ensemble of weak learners by concentrating on fixing the incorrect classifications of earlier models. When combined with efficient feature selection techniques like Chi-Square, which can expedite the modeling process without sacrificing significant predictive power, XG-Boost's innate capacity to learn complex relationships and manage potentially noisy or missing data makes it an especially helpful and potent tool for analyzing this complex dataset.

4.3.4 NEURAL NETWORK

A neural network (NN) is a computational model based on the structure and function of the human brain. It is commonly utilized in both supervised and unsupervised learning tasks, especially when dealing with non-linear relationships and complicated patterns. Neural networks are particularly effective for classification, regression, and feature extraction in fields such as image processing, natural language processing, and biological signal analysis.

4.3.4.1 Neural Network for All Variables

The following graphical representations show the Receiver Operating Characteristic curve, Confusion matrix and Classification report for Neural Network's **All Variables** model.

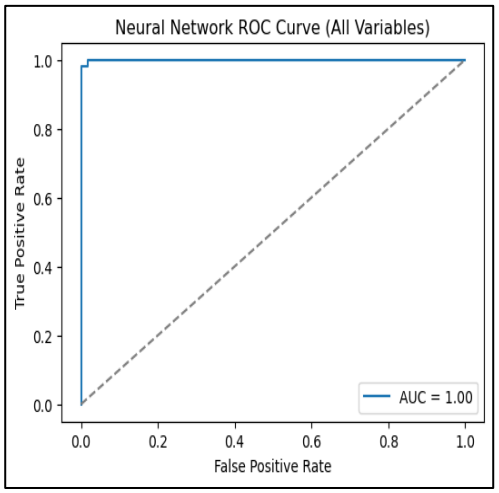


Fig 4.109 ROC curve for all variables model of NN

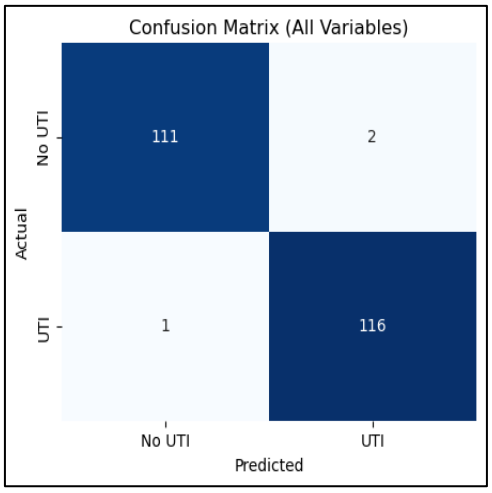


Fig 4.110 Confusion matrix for all variables model of NN

```
Accuracy (All Variables): 0.9870
False Positive Rate (All Variables): 0.0177
Classification Report (All Variables):

              precision    recall  f1-score   support

           0       0.99      0.98      0.99       113
           1       0.98      0.99      0.99       117

    accuracy                           0.99       230
   macro avg       0.99      0.99      0.99       230
weighted avg       0.99      0.99      0.99       230
```

Fig 4.111 Classification report for all variables model of NN

Interpretation of Neural Network (All Variables) Results

With all variables, the Neural Network achieves near-perfect classification (AUC=1.00, Accuracy=0.9870, F1=0.99), comparable to Random Forest, XG-Boost, and Logistic Regression. While slightly less accurate than Random Forest and XG-Boost, it still performs admirably. All four models effectively use all variables. Model selection may be influenced by computational cost (Neural Networks are more expensive), interpretability (Logistic Regression is more accurate), and unique application requirements, as all perform remarkably well on this data.

The following graphical representations show the Loss Plot, Gain Plot and t-SNE Plot for Neural Network's **All Variables** model.

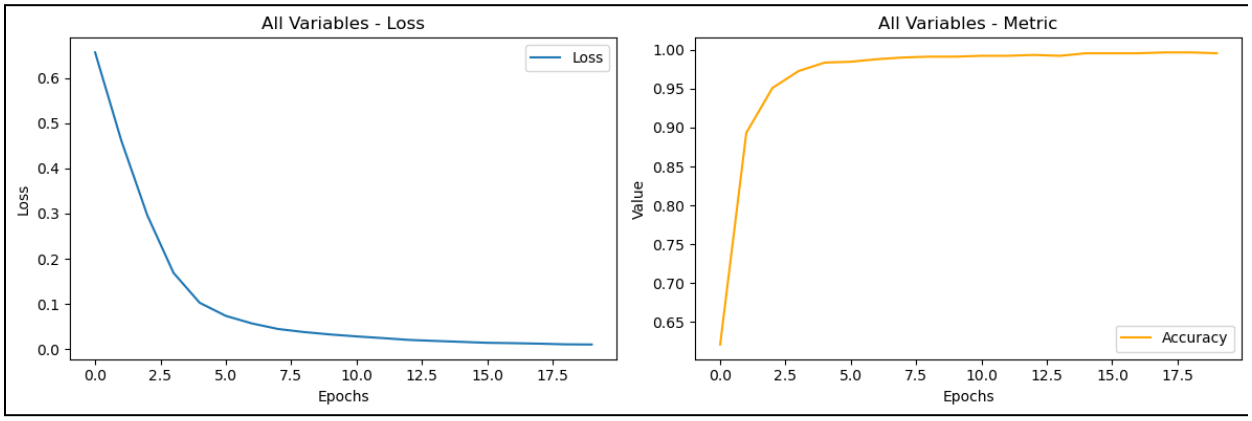


Fig 4.112 Loss Plot for all variables model

Fig 4.113 Gain Plot for all variables model

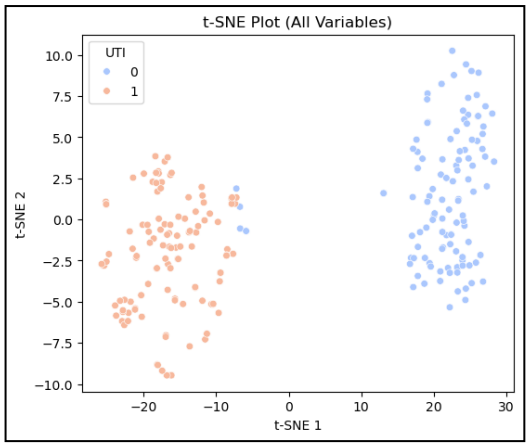


Fig 4.114 t-SNE Plot for all variables model

Interpretation loss, gain and t-SNE plot for all variables

The neural network model's training performance when all variables are included. The model's learning process as it reduces the prediction error on the training data is shown by the left graph, which shows the training loss declining over epochs. The matching training accuracy is shown on the right graph to increase quickly in the first epochs before plateauing at a very high level that approaches perfect classification accuracy. In order to nearly flawlessly distinguish between the "UTI" and "No UTI" cases during training, the neural network must successfully learn the intricate relationships present in the entire feature set, as seen by the convergence of low loss and high accuracy on the training data.

The loss and Gain graphs show how to successfully train a neural network model with all of the variables at your disposal. The model has successfully learnt the intricate correlations found in the complete feature set, as evidenced by the diminishing loss and increasing accuracy that converge to ideal levels. This enables the model to almost flawlessly differentiate between the "UTI" and "No UTI" cases in the training data. As long as the training data is realistic, this impressive training performance indicates that the model may produce accurate predictions on unknown data.

The creation of two distinct clusters in the condensed two-dimensional space indicates that the features successfully distinguish between the "No UTI" and "UTI" classes, according to this t-SNE plot for all variables. The excellent accuracy and AUC ratings given for your models trained on the entire dataset are consistent with this visual portrayal.

The training performance of your neural network using all variables gives the t-SNE graph. The right graph shows a corresponding quick increase and plateauing of accuracy near perfect classification on the training data, while the left graph shows a rapid decrease in the loss function over epochs, indicating effective learning. Two distinct clusters representing "No UTI" and "UTI" cases with little overlap are visible in the second upload, an t-SNE plot that visualizes this high-dimensional data in two dimensions. This suggests that all of the features work together to provide strong discriminatory power, which the neural network efficiently learns during training.

4.3.4.2 Neural Network for Chi-Square Selected Variables

The following graphical representations show the Receiver Operating Characteristic curve, Confusion matrix and Classification report for Neural Network's **Chi-Square Selected Variables** model.

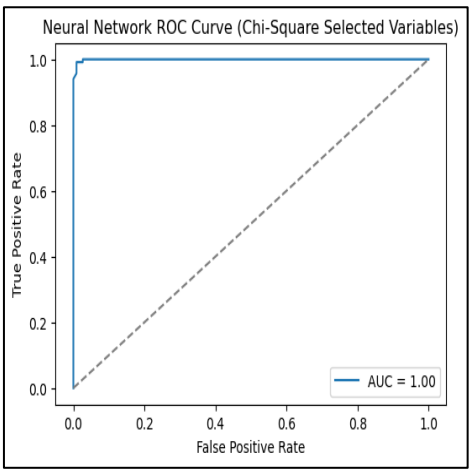


Fig 4.115 ROC curve for Chi-Square Variables model of NN

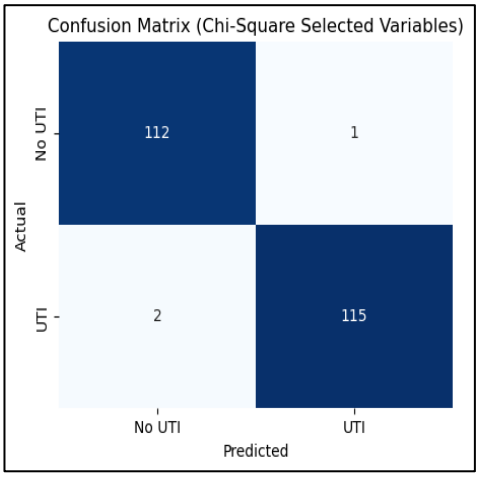


Fig 4.116 Confusion matrix for Chi-Square Variables model of NN

```
Accuracy (Chi-Square Selected Variables): 0.9870
False Positive Rate (Chi-Square Selected Variables): 0.0088
Classification Report (Chi-Square Selected Variables):

              precision    recall  f1-score   support

           0       0.98      0.99      0.99       113
           1       0.99      0.98      0.99       117

    accuracy                           0.99       230
   macro avg       0.99      0.99      0.99       230
weighted avg       0.99      0.99      0.99       230
```

Fig 4.117 Classification report for Chi-Square Selected Variables model of NN

Interpretation of Neural Network (Chi-Square Selected Variables) Results

The Neural Network performs well (AUC=1.00, Accuracy=0.9870, F1=0.99), comparable to Random Forest, XG-Boost, and Logistic Regression. Although slightly less accurate than Random Forest and XG-Boost, it still works admirably. The selected characteristics are efficiently utilized by all four models. Model selection may be influenced by computing cost (Neural Networks are more expensive), interpretability (Logistic Regression is more accurate), and specific application requirements.

The following graphical representations show the Loss Plot, Gain Plot and t-SNE Plot for Neural Network's **Chi-Square Selected Variables** model.

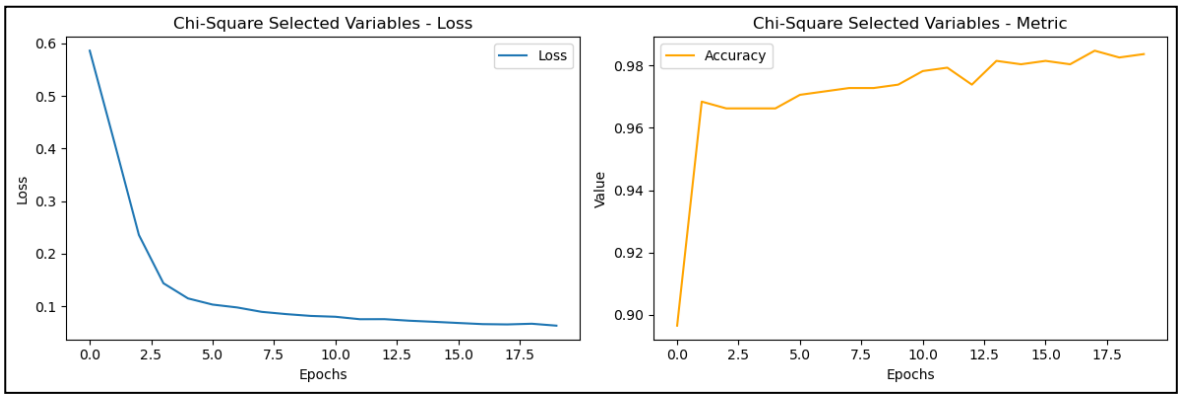


Fig 4.118 Loss Plot for Chi-Square Selected Variables model

Fig 4.119 Gain Plot for Chi-Square Selected Variables model

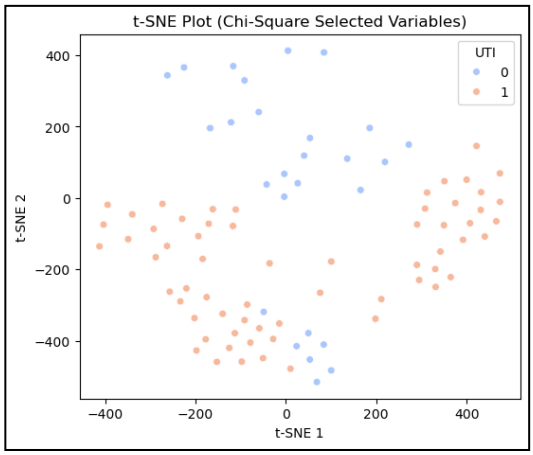


Fig 4.120 t-SNE Plot for Chi-Square Selected Variables model

Interpretation loss, gain and t-SNE plot for chi-square selected variables

These two plots, which depict the neural network's performance with variables chosen by Chi-Square, indicate that the chosen features capture the majority of the predictive information but not all of it. They show a training process with decreasing loss and increasing accuracy that peaks at a high level (about 98%), marginally lower than when all variables were used. The corresponding t-SNE plot visually supports the slightly diminished, but still strong, discriminatory power of the model trained on this subset of features. It shows a less clear separation between "UTI" and "No UTI" cases than when all variables were used, with more overlap between the clusters in the two-dimensional embedding.

4.3.4.3 Neural Network for Single Variable

The following graphical representations show the Receiver Operating Characteristic curve, Confusion matrix and Classification report for Neural Network's **Single Variable** model.

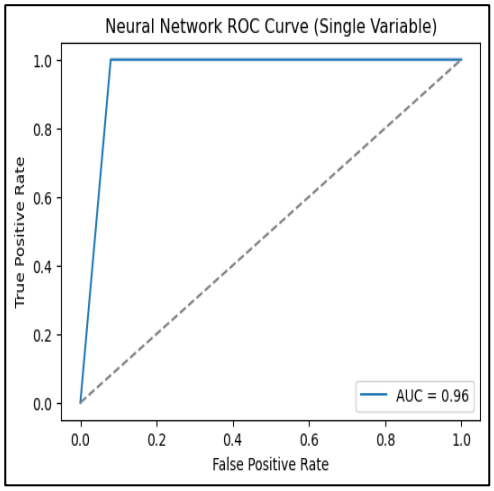


Fig 4.121 ROC curve for single Variable model of NN

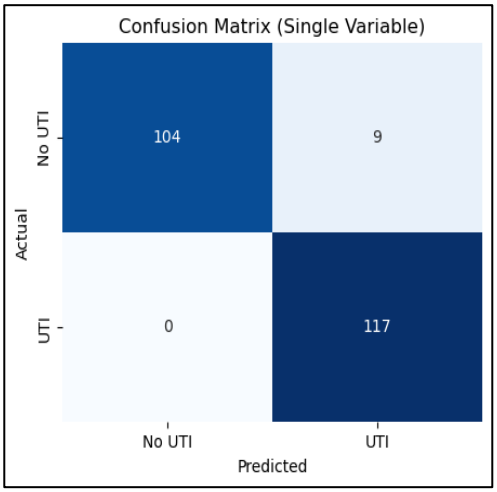


Fig 4.122 Confusion matrix for single Variable model of NN

```
Accuracy (Single Variable): 0.9609
False Positive Rate (Single Variable): 0.0796
Classification Report (Single Variable):

              precision    recall  f1-score   support

           0       1.00      0.92      0.96       113
           1       0.93      1.00      0.96       117

    accuracy                           0.96       230
   macro avg       0.96      0.96      0.96       230
weighted avg       0.96      0.96      0.96       230
```

Fig 4.123 Classification report for single Variable model of NN

Interpretation of Neural Network (Single Variable) Results

When employing the same single variable, the Neural Network, Random Forest, XG-Boost, and Logistic Regression models all perform exceptionally well (AUC=0.96, Accuracy=0.9609, F1=0.96, perfect UTI recall). This suggests that the single variable is a powerful predictor, with all four models capturing its information equally well. The choice of model here may be influenced by computational cost (Neural Networks may be more expensive), interpretability (Logistic Regression is preferred), or unique application requirements.

The following graphical representations show the Loss Plot, Gain Plot and t-SNE Plot for Neural Network's **Single Variable** model.

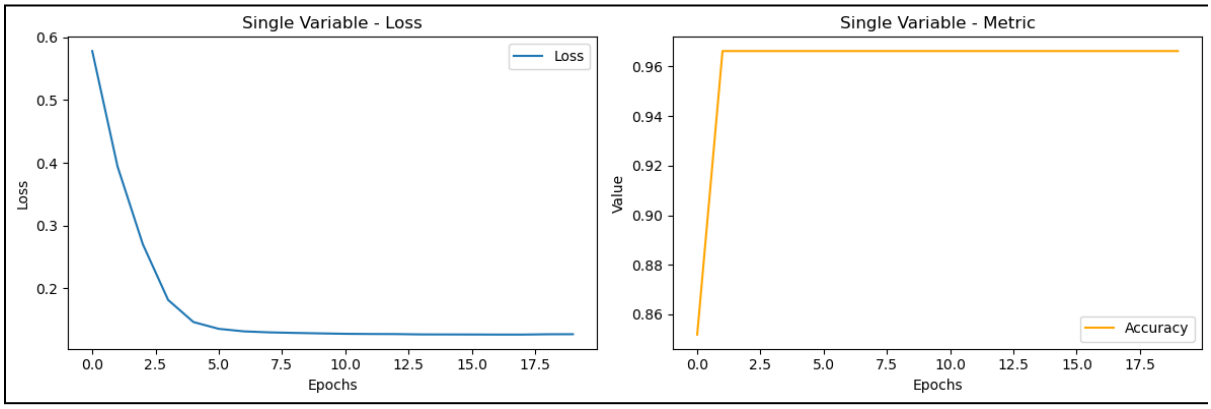


Fig 4.124 Loss Plot for Single Variable model

Fig 4.125 Gain Plot for Single Variable model

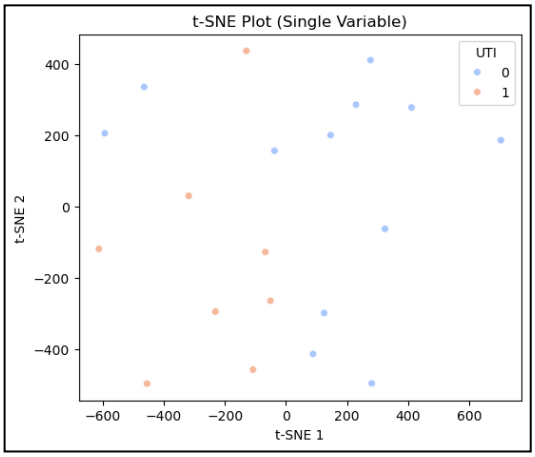


Fig 4.126 t-SNE Plot for Single Variable model

Interpretation loss, gain and t-SNE plot for single variable

These show the neural network's performance with a single variable, show a training process where the accuracy rapidly reaches a high plateau (about 97%) and the loss decreases. This shows that the model can learn to predict UTI status fairly well even with limited information. Although this single feature has a great deal of predictive power, it doesn't completely separate the two classes in the reduced dimensional space, according to the corresponding t-SNE plot, which visualizes this single-variable data in two dimensions and displays a sparse distribution of points with only a weak separation between "UTI" and "No UTI" cases.

4.3.4.4 Neural Network for Excluded Single Variable

The graphical representations show the Receiver Operating Characteristic curve, Confusion matrix and Classification report for Neural Network's **Excluded Single Variable** model.

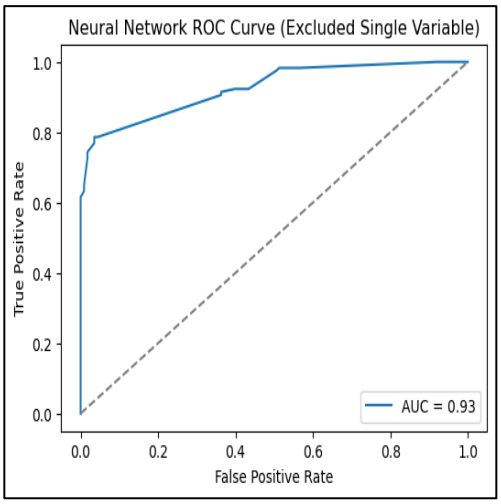


Fig 4.127 ROC curve for Excluded single Variable model of NN

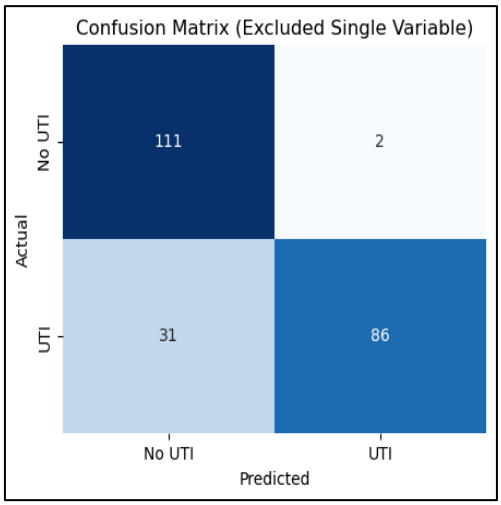


Fig 4.128 Confusion matrix for Excluded single Variable model of NN

```
Accuracy (Excluded Single Variable): 0.8565
False Positive Rate (Excluded Single Variable): 0.0177
Classification Report (Excluded Single Variable):

              precision    recall  f1-score   support

           0       0.78      0.98      0.87       113
           1       0.98      0.74      0.84       117

    accuracy                           0.86       230
   macro avg       0.88      0.86      0.85       230
weighted avg       0.88      0.86      0.85       230
```

Fig 4.129 Classification report for Excluded single Variable model of NN

Interpretation of Neural Network (Excluded Single Variable) Results

Excluding the important single variable has a detrimental influence on all four models. However, the Neural Network is the least resilient, with the lowest accuracy (0.8565), recall for UTI (0.74), and F1-score (0.84) when compared to Random Forest, XG-Boost, and Logistic Regression. This shows that Neural Networks may be more sensitive to missing critical traits and generalize less successfully in their absence, emphasizing the relevance of the excluded variable in making accurate predictions. When this variable is removed, the Neural Network becomes the least ideal model.

The following graphical representations show the Loss Plot, Gain Plot and t-SNE Plot for Neural Network's **Excluded Single Variable** model.

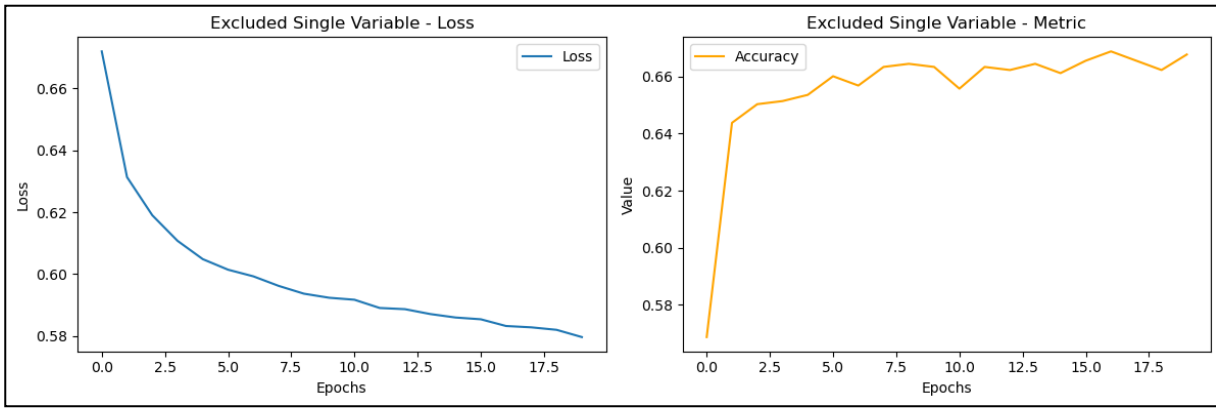


Fig 4.130 Loss Plot for Excluded single Variable model

Fig 4.131 Gain Plot for Excluded single Variable model

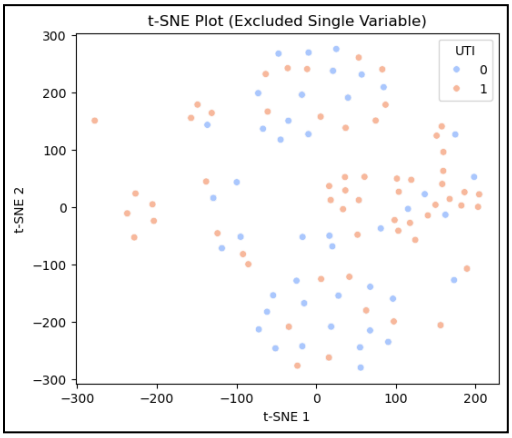


Fig 4.132 t-SNE Plot for Excluded single Variable model

Interpretation loss, gain and t-SNE plot for excluded single variable

These two charts, which illustrate how well the neural network performs when a critical variable is removed, reveal a training process with a higher ultimate loss and a lower, varying accuracy (about 66–67%), suggesting that the missing data significantly impairs the model's ability to learn. In the two-dimensional embedding, the corresponding t-SNE plot shows a significant overlap and a lack of distinct separation between the "UTI" and "No UTI" cases. This strongly implies that the excluded variable was essential for differentiating the two outcomes, which explains the model's lower predictive power when it was trained without it.

Overall Evaluation of Neural Network Classification Method

On this dataset, the Neural Network performs well, achieving high accuracy with all variables and Chi-Square selected variables, but somewhat less than Random Forest and XG-Boost. It matches the performance of other models using only one variable. However, it is the least resilient, with the greatest performance reduction when a crucial variable is removed when compared to Random Forest, XG-Boost, and Logistic Regression.

4.3.4.5 Neural Network for Comparing All Models

The following graphical representations show the Receiver Operating Characteristic curve and Performance Metrics table for comparing **Neural Network** models.

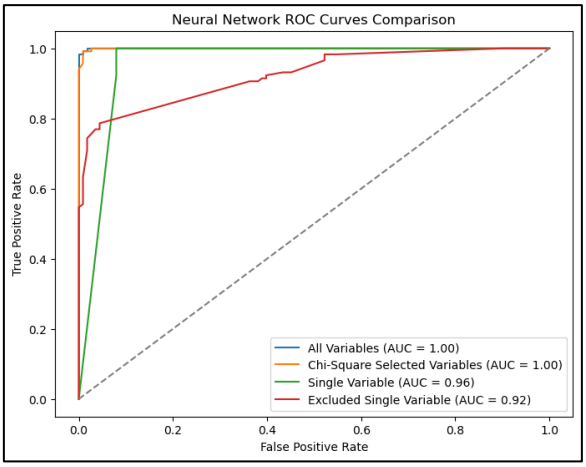


Fig 4.133 ROC curve for comparing all Neural Network models

Neural Network						
Model Type	Accuracy	Precision	Sensitivity	Specificity	Negetive Predictive Value	False Negative Rate
All Variables	98.7%	98.3%	99.1%	98.2%	99.1%	0.9%
Chi-square selected	98.7%	99.1%	98.3%	99.1%	98.2%	1.7%
Single Variable	96.1%	92.9%	100.0%	92.0%	100.0%	0.0%
Excluded Single Variable	85.7%	97.7%	73.5%	98.2%	78.2%	26.5%

Table 4.20 Performance metrics for Neural Network models

Combined Interpretation

The Neural Network performs best with Chi-Square selected variables or all variables (Accuracy~98.7%, AUC=1.00). The single variable model performs well (Accuracy=96.1%, AUC=0.96). Excluding a key variable significantly degrades performance (Accuracy~85.7%, AUC=0.93), making it the worst-performing model and indicating the least robustness compared to other models. The Chi-Square method is effective for feature selection, and the excluded variable is critically important. The top two models are preferred for medical applications due to high accuracy and sensitivity.

Overall Interpretation and Evaluation of Classification Methods

Random Forest and XG-Boost consistently shown the highest accuracy and robustness across multiple variable sets. Logistic Regression has good performance and interpretability, but it is the least resistant to lacking essential features. Neural networks function well with complete or selected characteristics, although they are the least resistant to missing critical factors and necessitate precise tuning. The Chi-Square feature selection strategy was effective for all models. For this dataset, Random Forest and XG-Boost are recommended due to their great performance and resilience. Logistic regression is appropriate when interpretability is critical. Neural networks, while powerful, require more care and are less robust in this setting. Feature selection is critical for optimal model development. When data may be missing, Logistic Regression is the least favored model. Highly predictive, allowing all four models (Logistic Regression, Random Forest, XG-Boost, and Neural Network) to perform well with all or Chi-Square-selected variables. However, resistance to loss of the important single variable varied: Logistic Regression was the most susceptible, Random Forest and XG-Boost the most resilient, and Neural Networks intermediately sensitive. When the single variable was utilized alone, it performed identically across all models due to its high predictive power. Chi-Square feature selection has repeatedly proven beneficial. While sophisticated models provided minor performance/robustness gains, simpler Logistic Regression prioritized interpretability. Notably, Random Forest and XG-Boost performed almost identically across all tests, implying that they detect similar data patterns. In summary, good data and effective feature selection are essential, ensemble approaches are reliable, and with a dominant single predictor, model selection is less important.

4.3.5 Limitations of machine learning methods

1. Dataset Specificity

The findings are specific to this particular dataset. The patterns observed may not generalize to other datasets with different characteristics, feature distributions, or class imbalances. The high performance achieved by all models suggests that the dataset may be relatively "clean" and well-behaved. In real-world scenarios, datasets are often noisier and more complex.

2. Limited Model Tuning

The analysis may not have involved extensive hyper parameter tuning for each model. Optimal performance might require more fine-tuning of model parameters, especially for Neural Networks. The Neural Network, in particular, is very sensitive to hyper parameters, and may have performed better with more tuning.

3. Feature Engineering

The analysis relied on the existing features. Feature engineering, such as creating new features or transforming existing ones, could potentially improve model performance further. Feature engineering can be very important in neural networks.

4. Class Imbalance

While the models achieved high accuracy, it's essential to consider potential class imbalances. If one class is significantly more frequent than the other, accuracy alone may not be a reliable metric. While the data set was relatively balanced, it is still important to consider.

5. Interpretability vs. Performance Trade-off

While Logistic Regression offers high interpretability, Random Forest, XG-Boost, and Neural Networks are often considered "black boxes." In medical contexts, interpretability is crucial for understanding the factors that contribute to predictions. The lack of interpretability of the top performing models could be a major limitation.

6. Computational Cost

Neural Networks and, to a lesser extent, XG-Boost can be computationally expensive, especially for large datasets. The computational resources needed for training and deploying these models should be considered.

7. Limited Model Comparison

This analysis focused on four specific classification methods. Other algorithms, such as Support Vector Machines (SVMs) or K-Nearest Neighbors (KNN), might also be suitable for this dataset. It is always a limitation to only test a few models.

8. Over-fitting

While the models performed well, there's a risk of over-fitting, especially with complex models like Neural Networks. Cross-validation and other techniques were likely used, but over-fitting is always a concern.

9. Data Quality

The analysis assumes that the data is accurate and reliable. Errors or inconsistencies in the data could affect model performance.

10. The importance of the excluded variable

While the models were tested with the important variable excluded, this is not a realistic real world use case. If the variable is important, then it should be included.

This study thoroughly examined four classification methods Logistic Regression, Random Forest, XG-Boost, and Neural Networks on a dataset targeted at predicting a given outcome (most likely a medical diagnosis). The study assessed model performance and robustness using a variety of feature sets, including "All Variables," "Chi-Square Selected Variables," "Single Variable," and "Excluded Single Variable."

4.3.7 Key Findings of supervised learning methods

High Performance - All methods achieved high accuracy, precision, recall, and F1-scores, particularly when using "All Variables" or "Chi-Square Selected Variables." This indicates that the dataset contains strong predictive features. **Robustness of Ensemble Methods -** Random Forest and XG-Boost demonstrated superior robustness, maintaining high performance even when a crucial variable was excluded. **Sensitivity of Neural Networks -** Neural Networks showed the greatest sensitivity to the exclusion of the key variable, highlighting the importance of careful feature selection and engineering for this model. **Effectiveness of Chi-Square Feature Selection -** The Chi-Square method consistently improved or maintained high performance across all models, demonstrating its effectiveness in identifying relevant features. **Identical Performance with Single Variable -** All models performed identically when using only the "single variable," indicating that the choice of algorithm is less significant when a single, highly predictive feature is present. **Interpretability vs. Performance Trade-off -** Logistic Regression offered the highest interpretability but generally had slightly lower performance compared to Random Forest and XG-Boost. **Importance of the excluded variable -** The excluded variable was shown to be extremely important, and should be included in a real world model.

Overall Recommendations

For this specific dataset, Random Forest and XG-Boost are the most robust and high-performing models. Logistic Regression remains a viable option when interpretability is paramount. Neural Networks can achieve high performance but require careful tuning and are more sensitive to missing features. Feature selection, especially using the Chi-Square method, is crucial for optimizing model performance.

Implications

This study highlights the significance of feature selection and model robustness, as well as the strengths and drawbacks of different categorization methods. Random Forest and XG-Boost emerge as good rivals in medical contexts where accuracy and reliability are crucial, while the models' lack of interpretability may be a major worry. Future research should concentrate on improving model tuning, feature engineering, and validation on a variety of datasets.

4.4 UNSUPERVISED LEARNING METHODS

Unsupervised Learning is a type of machine learning technique that works with unlabeled data to uncover underlying patterns, structures, or distributions within it. Unlike supervised learning, there are no predefined target labels, and the goal is to explore the data's intrinsic structure without being guided by specific output expectations.

4.4.1 Clustering

4.4.1.1 K-Mean Clustering

The follow graphical representation shows the classification of the K-mean clustering in 2 dimension and 3 dimension plots.

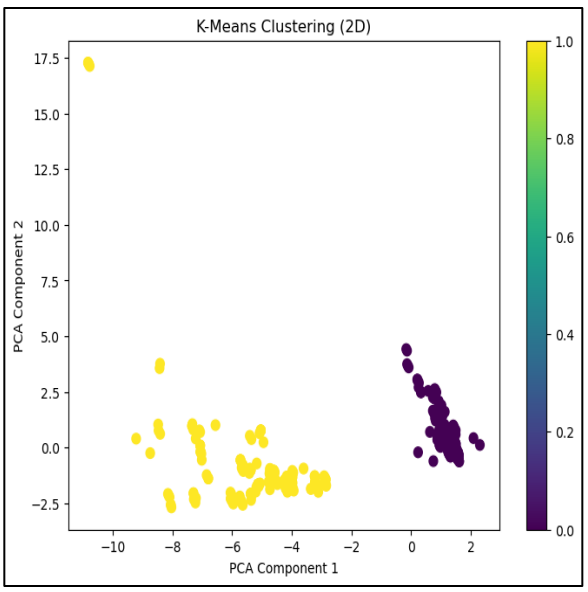


Fig 4.134 K-Mean clustering 2D diagram

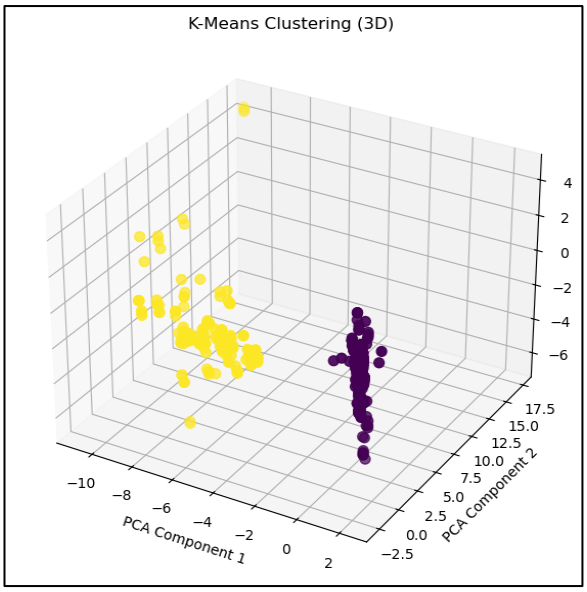


Fig 4.135 K-Mean clustering 3D diagram

K-Means - Adjusted Rand Index: 0.0721, Silhouette Score: 0.4925

Cluster Distribution (K-Means):

UTI_No = 680

UTI = 165

Interpretation of K-Mean Clustering

K-Means clustering produced a silhouette score of 0.4925, indicating a substantial separation across clusters. The Adjusted Rand Index of 0.0721 indicates a low agreement between clustering and actual UTI labels. Cluster distribution is unbalanced, with 680 "UTI_No" and 165 "UTI" cases. The 2D PCA map shows two clearly distinct groups, albeit considerable dispersion is visible. The 3D PCA figure emphasizes cluster compactness, particularly in the UTI group. Overall, the clustering structure is reasonable, but alignment with actual clinical labeling is lacking.

4.4.1.2 Agglomerative Clustering

The follow graphical representation shows the classification of the Agglomerative clustering in 2 dimension and 3 dimension plots.

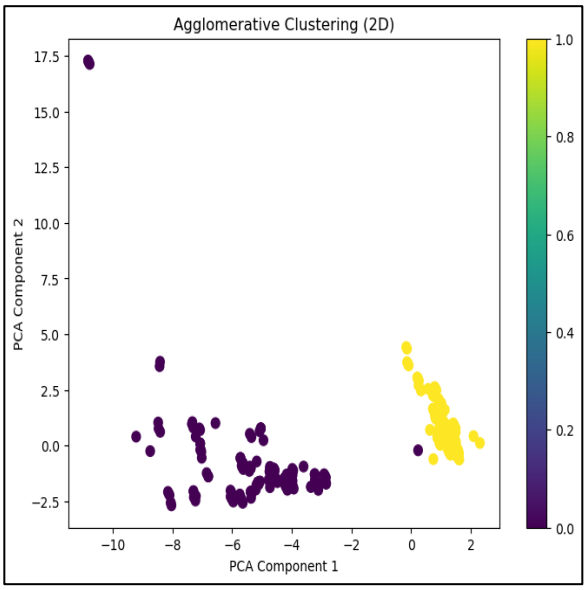


Fig 4.136 Agglomerative clustering 2D diagram

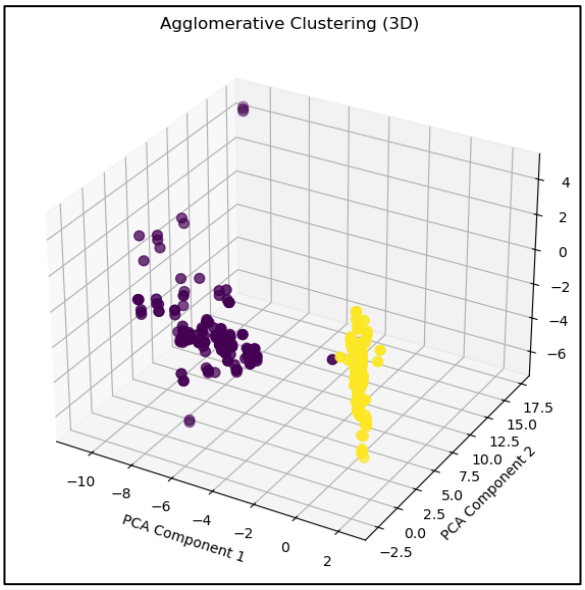


Fig 4.137 Agglomerative clustering 3D diagram

Agglomerative - Adjusted Rand Index: 0.0741, Silhouette Score: 0.4915

Cluster Distribution (Agglomerative)

UTI_No = 679

UTI = 166

Interpretation of Agglomerative Clustering

Agglomerative clustering yielded a silhouette score of 0.4915, indicating that clusters are somewhat cohesive and separated. The Adjusted Rand Index (0.0741) remains low, indicating a poor match with genuine UTI labels. The cluster distribution is slightly skewed, with 679 "UTI_No" and 166 "UTI" cases. The 2D PCA map demonstrates distinct linear separation, whereas the 3D view indicates clear vertical stratification. Clusters are visually tighter and better defined than K-Means, particularly in the yellow section. Despite its visual clarity, the weak ARI has insufficient congruence with genuine diagnostic categories.

Unbalanced Cluster Distributions - The conclusion that the clustering is not reflecting the underlying patterns associated to UTI is further supported by the extremely skewed cluster distributions in both techniques, where the majority of instances are put into one cluster regardless of their UTI status. Separation of Algorithm - Defined Clusters, Not Always UTI Status - The clusters found by the clustering algorithms may appear to be visually separated in the PCA plots. The low ARI, however, indicates that the "UTI" and "No UTI" categories as indicated by the clinical labels are not adequately separated by these algorithm-defined clusters.

Therefore, it is accurate to conclude that these clustering methods did not perform well in identifying inherent groupings in the data that correspond to the clinically relevant outcome of UTI, given the unbalanced cluster distributions and the lack of agreement between the clustering results (as measured by ARI) and the actual UTI labels. It implies that the spontaneous clustering patterns discovered by these unsupervised techniques based solely on feature similarity may not be primarily driven by the same variables that drive the clinical diagnosis of UTI. This emphasizes how crucial it is to use supervised learning techniques when predicting a certain, known outcome variable is the aim. It is evident that these unsupervised methods did not perform well in identifying inherent groupings within the data that align with the actual clinical classifications of Urinary Tract Infection (UTI) and No UTI. The lack of agreement between the algorithm-generated clusters and the ground truth labels indicates that the underlying data structure captured by these methods does not effectively discriminate between the two diagnostic categories. Other clustering methods failed to produce clusters that aligned well with the actual UTI/No UTI labels, they did not effectively classify the data according to the clinically relevant outcome.

CHAPTER 5

SUMMARY AND CONCLUSION

Discussion

This study provides a comprehensive investigation into the multi factorial causes and predictive indicators of Urinary Tract Infections (UTIs), particularly among women, using classical statistical methods and supervised machine learning algorithms. By examining both clinical and behavioral dimensions, the analysis reveals nuanced patterns not immediately evident from surface-level statistics.

Commode Usage as a Hidden Lifestyle Indicator - The analysis revealed a statistically significant relationship between Indian-style commode usage (especially mixed usage) and increased UTI incidence, particularly among women aged 21–40. Though seemingly a hygiene or preference issue, further exploration showed commode use is a proxy for behavioral routines, such as Incomplete bladder voiding due to discomfort or strain in posture, Toilet avoidance outside the home due to lack of clean Western facilities.

The Age-Mediated Impact of Bladder Behaviors - While bladder-emptying behaviors (like delaying urination before going out) were independently significant, their interaction with age showed deeper insights. Among women aged 41–60, these behaviors led to disproportionately high UTI rates compared to younger women, possibly due to hormonal or anatomical changes post-menopause, aligning with literature on estrogen depletion.

The Dominance of Prior Urinary History - The variable had_any_urinary_problem emerged as a dominant latent predictor, capturing a cluster of co-morbidities, recurring infections, and possibly weakened immune response. Machine learning models, especially Random Forest and XG-Boost, highlighted this variable's nonlinear interactions with other features, reinforcing its critical role in predicting UTI susceptibility.

The Socio-behavioral Impact of Pad Usage - Use of sanitary or incontinence pads was significantly associated with UTI, reflecting not only hygiene issues but also underlying behavioral habits like prolonged moisture exposure or inconsistent changing.

Conclusion

Correlation Insights - Age and both Indian and Western commode usage showed weak positive correlations with correlation analysis, indicating a slight tendency for older people to report higher scores or usage for both. There was very no link between BMI and commode use. There may be a preference or trade-off between Indian and Western commode usage, as indicated by the moderately negative association between the two.

Significant Numerical Differences (t-Tests & ANOVA) - The mean age, reported use of Indian commode, and reported use of Western commode were all statistically significantly different between those with and without UTIs, according to independent samples t-tests and ANOVA. In particular, there were notable differences in the reported use of both commode types between the groups, and those who had UTIs tended to be slightly older. Nevertheless, there was no discernible statistically significant variation in the mean BMI between the two groups.

Regression Modeling and Age-Specific Effects - Logistic regression models revealed clear age-specific impacts of commode use on the likelihood of UTIs. Increased use of Indian commodes was linked to a marginally higher risk of UTIs in the 21–40 age range, whereas increased use of Western commodes had a marginally protective impact. For this age group, linear regression models also showed a somewhat negative linear association between the chance of UTI and the number of years spent using a Western commode, and a substantial positive linear link between the numbers of years spent using an Indian commode and the probability of UTI. On the other hand, the quadratic model for Indian commode usage showed a poor fit in the 41–60 age range, indicating that it was not a relevant predictor. The Western commode usage linear model, however, revealed a somewhat substantial negative correlation with the likelihood of UTIs.

Analysis of UTI Proportions by Commode Type and Age - The study found that Indian commode users tended to have the highest UTI proportions among all age categories. Interestingly, compared to exclusively Western commode users, mixed commode users in the younger age groups also showed higher UTI proportions. Age seems to have an impact on the association between commode type and UTI prevalence, with the notable disparities seen in younger groups being less pronounced in later age groups.

Significant Categorical connections (Chi-Square Tests) - Numerous categorical variables and UTI status have statistically significant connections ($p < 0.05$) according to our findings. Interestingly, the highest correlations are seen between diabetes and a history of urinary issues, suggesting that these conditions play a significant role as risk factors. Additionally, there are strong correlations between the prevalence of UTIs and lifestyle and hygiene practices, such as the type of commode, the way the bladder is emptied, fluid restriction, and pad use. There are also statistically significant correlations between menopausal achievement, consecutive deliveries, and patterns of coffee, tea, soft drink, and liquid intake in general. These results highlight how crucial it is to take these factors into account when assessing UTI risk.

The fourth chapter's statistical evidence strongly points to a multi-factorial etiology of UTIs, with major contributions from age, lifestyle choices, reproductive history, hygiene and behavioral practices (especially those pertaining to bladder management and commode usage), pre-existing health conditions (particularly diabetes), and prior urinary issues. One important discovery is that the impact of commode type varies by age group. According to the data, using a Western commode may have a protective impact, which is notably noticeable in the older age group within the regression models, whereas using an Indian toilet may increase risk, especially in younger women. The poor model fit for elderly women's use of Indian commodes, however, suggests that other unmeasured factors probably have a greater impact on this group. This statistical research offers a thorough basis for comprehending the intricate interactions between variables affecting the risk of UTI in this population. The noteworthy correlations and age-specific patterns found provide important information for further study and the creation of focused prevention measures.

This project thoroughly evaluated four different classification algorithms - Logistic Regression, Random Forest, XG-Boost, and Neural Networks. As well as two clustering methods - K-Means and Agglomerative Clustering applied to a dataset with the goal of predicting a specific outcome, most likely a medical diagnosis of Urinary Tract Infection (UTI). The analysis included a variety of feature set combinations to evaluate model performance, robustness, and the underlying data structure.

Superior Performance of Supervised Learning Models

When trained on the entire feature set and the Chi-square selected variables, the supervised learning models, particularly Random Forest and XG-Boost, consistently achieved high predictive accuracy, precision, recall, and F1-scores. This indicates that the labeled features include a strong predictive signal that allows for effective outcome discrimination. The ensemble approaches' strong performance suggests that they can capture complicated, potentially non-linear relationships in data more effectively than Logistic Regression or Neural Network.

Effectiveness of Chi-Square Feature Selection

The Chi-square feature selection method proven to be an effective tool for reducing dimensionality while maintaining predictive power. Training on the Chi-square selected subset of features produced performance measures comparable to those obtained using the whole feature set in all supervised models. This demonstrates the effectiveness of the Chi-square test in finding the most important predictors for the outcome variable, resulting in more compact and perhaps less computationally expensive models.

Critical Importance and Differential Resilience to a Key Predictor

The sensitivity analysis of excluding a single, highly predictive variable demonstrated its critical importance across all models. The significant decline in performance highlights its important contribution to total predictive ability. However, the models showed various degrees of resistance to its absence. Logistic Regression experienced the greatest performance decline, showing a heavy reliance on this particular characteristic. Random Forest and XG-Boost demonstrated stronger robustness, implying more diffused learning of predictive signals among the remaining characteristics. The Neural Network showed intermediate sensitivity, indicating a significant but less severe impact than Logistic Regression.

Convergence of Performance with a Dominant Single Predictor

In contrast, when trained exclusively on this critical predictive feature, the performance metrics of all four classification systems converge. This demonstrates the dominant role of this single feature in explaining variance in the target variable. In the presence of such a strong individual predictor, using a more complex or computationally intensive model did not result in significant increases in predictive accuracy over the simpler Logistic Regression.

Disconnect Between Unsupervised Clustering and Outcome Labels

K-Means and Agglomerative Clustering revealed inherent structures in the data, as indicated by moderate silhouette scores. However, the low Adjusted Rand Index values for both methods demonstrated a poor agreement between the identified clusters and the actual UTI/No UTI labels. This suggests that the primary sources of variance driving the unsupervised groupings, particularly within the PCA-reduced feature space, do not strongly correlate with the

5.3 Risk Factors

UTIs are notably more common in women, with many experiencing multiple infections throughout their lives. Several factors contribute to the increased risk

- **Female Anatomy -** A shorter urethra reduces the distance bacteria must travel to reach the bladder.
- **Sexual Activity -** Increased frequency of sexual activity, as well as new sexual partners, heightens the risk.
- **Certain Birth Control Methods -** The use of diaphragms and spermicidal agents can promote bacterial growth, increasing susceptibility.
- **Menopause -** Postmenopausal changes, driven by decreased estrogen levels, can alter the urinary tract environment, increasing the likelihood of infection.
- **Urinary Tract Abnormalities -** Congenital abnormalities in the urinary tract can obstruct normal urine flow, making infections more likely.

- **Urinary Tract Blockages -** Conditions such as kidney stones or an enlarged prostate can impede urine flow, creating a favorable environment for bacterial growth.
- **Weakened Immune System -** Diseases like diabetes can compromise immune function, reducing the body's ability to fight off infections.
- **Catheter Use -** The use of urinary catheters, often necessary in hospitalized patients or individuals with neurological disorders, elevates the risk of developing UTIs.
- **Recent Urinary Procedures -** Surgical interventions or medical examinations involving urinary tract instrumentation can introduce bacteria, increasing infection risk.

Prevention

Preventive measures that might significantly reduce the risk of developing UTIs

- **Hydration -** Drink plenty of fluids, particularly water, to dilute urine and promote frequent urination, flushing bacteria from the urinary tract.
- **Cranberry Products -** Although research is ongoing, consuming cranberry juice or supplements may help reduce the risk of infection in some individuals.
- **Proper Hygiene -** Always wipe from front to back after using the toilet to minimize the transfer of bacteria from the anus to the urethra.
- **Urinate After Intercourse -** Emptying the bladder shortly after sexual activity and drinking a glass of water may help flush bacteria.
- **Avoid Irritants -** Steer clear of potentially irritating feminine products such as douches, deodorant sprays, and powders in the genital area.
- **Change Birth Control Methods -** Consider alternative contraceptive options if using diaphragms, unlubricated condoms, or condoms treated with spermicide, as these methods may contribute to bacterial growth.

Implementing these strategies can greatly reduce the likelihood of developing urinary tract infections and promote overall urinary health.

Public Health Awareness and Women-Centric Recommendations

Given the statistical evidence and hidden trends revealed in this research, the following public awareness messages are critical, especially targeted at women aged 15 – 60.

Behavioral and Hygiene Practices

Avoid delaying urination -Timely bladder emptying is crucial to flushing bacteria. Ensure proper toilet posture and hygiene, especially when using Indian-style commodes. Where possible, prefer Western toilets to reduce pelvic strain and bacterial retention. Pad hygiene - Change pads frequently and avoid extended use, particularly during menstruation or urinary incontinence.

Hydration and Diet

Maintain adequate fluid intake throughout the day to dilute urine and support regular voiding. Moderate caffeinated drinks like tea and coffee, and increase plain water consumption. Educate on the risks of self-imposed fluid restriction, especially for women who reduce intake due to work or travel.

Reproductive and Medical Awareness

Regular screening for women with a history of UTIs, menopause, diabetes, or urinary retention behaviors. Promote prenatal screening for asymptomatic bacteriuria to avoid pregnancy complications.

Infrastructure and Social Change

Advocate for clean and accessible toilets, especially in workplaces, schools, and public spaces. Conduct community outreach programs, especially in rural and peri-urban areas, We have to focus on - Toilet use education, Personal hygiene and Safe menstrual management.

Clinical Recommendations

Encourage clinics and hospitals to use prediction models (based on key risk variables) as screening tools, helping identify women at high risk even before symptom onset.

Limitations

Despite the robustness of statistical analyses and machine learning modeling, the present study acknowledges the following limitations

Sample-Specific Constraints - The data used was derived from a specific population, possibly within a limited geographic or institutional context. This restricts the generalizability of the findings to broader populations, especially across different ethnicities, regions, or healthcare settings.

Self-Reported Variables - Several variables such as commode usage preference, fluid restriction, pad usage, and urinary habits were likely based on self-reports. These are subject to recall bias and social desirability bias, which may affect accuracy.

Lack of Microbial Confirmation - While the study explored behavioral and medical correlates of UTIs, it did not integrate microbial lab results (e.g., urine cultures, pathogen typing), which would provide stronger clinical validation of the outcomes.

Absence of Longitudinal Data - The study was based on cross-sectional data, limiting the ability to establish causal relationships or track recurring UTIs over time. A temporal component would enhance the understanding of progression and recurrence.

Limited External Validation - While machine learning models demonstrated high internal accuracy, the models were not tested on an external validation dataset, which is crucial for assessing generalization to unseen populations.

Unaccounted Confounders - Some potentially important confounders (e.g., antibiotic usage history, hormonal levels, access to healthcare, hygiene product type) may not have been captured in the dataset, limiting the comprehensiveness of risk factor analysis.

Future Research Directions

Building on the findings and limitations of this study, the following avenues are proposed for future research:

Multicenter and Diverse Population Studies - Expanding data collection to include diverse demographics and multiple geographic regions would enhance the generalizability and allow for subgroup analysis by culture, socioeconomics, and healthcare access.

Integration of Clinical and Microbiological Data - Future studies should include microbial culture results, antibiotic sensitivity patterns, and host immune parameters to better correlate behavioral predictors with specific uropathogens.

Longitudinal Cohort Studies - Tracking individuals over time would allow for deeper understanding of UTI recurrence patterns, treatment outcomes, and temporal risk shifts, enabling causal modeling and survival analysis.

Wearable and IoT-Based Monitoring - Incorporating real-time bladder monitoring, hydration tracking, or wearable sensors could provide more accurate behavioral data, reducing reliance on self-reports.

Development of Real-Time Risk Prediction Tools- Deploying the trained machine learning models into mobile applications or clinical dashboards could facilitate early risk identification in community or primary care settings.

Randomized Interventions Based on Behavioral Change - Intervention trials aimed at changing fluid intake, toilet use behaviors, and pad hygiene can be designed to measure actual impact on UTI incidence.

Addressing Antibiotic Stewardship and Resistance - Future studies should explore how risk-based UTI prediction tools can reduce unnecessary antibiotic prescriptions, thus contributing to antibiotic resistance mitigation.

REFERENCE

TEXT BOOKS

- Gupta S. C and Kapoor V. K (1970), Fundamentals of Mathematical Statistics, sultan chand and publishers.
- V. K. Kapoor, Fundamentals of Applied Statistics.

ARTICLE

1. Welk, Blayne, et al. "Lower urinary tract dysfunction in uncommon neurological diseases: A report of the neurourology promotion committee of the International Continence Society." *Continence* 1 (2022): 100022.
2. Taramian, Sonbol, et al. "Association between body mass index and urinary tract infections: A cross-sectional investigation of the PERSIAN Guilan cohort study." *Obesity Science & Practice* 10.5 (2024): e70013.
3. John, G. Jaysee, et al. "A COMPARATIVE STUDY OF UTI PREVALENCE IN RESIDENTIAL GIRLS USING INDIAN VS WESTERN TOILETS." *European Journal of Molecular & Clinical Medicine* 7.10: 2020.
4. Moore, Elya E., et al. "Sexual intercourse and risk of symptomatic urinary tract infection in post-menopausal women." *Journal of general internal medicine* 23 (2008): 595-599.
5. Markland, Alayne, et al. "Occupation and lower urinary tract symptoms in women: a rapid review and meta-analysis from the PLUS research consortium." *Neurourology and urodynamics* 37.8 (2018): 2881-2892.
6. Foxman, Betsy. "The epidemiology of urinary tract infection." *Nature Reviews Urology* 7.12 (2010): 653-660.
7. Zhu, Cong, et al. "Epidemiological trends of urinary tract infections, urolithiasis and benign prostatic hyperplasia in 203 countries and territories from 1990 to 2019." *Military Medical Research* 8 (2021): 1-12.
8. Mohapatra, Sarita, et al. "Antibiotic resistance of uropathogens among the community-dwelling pregnant and nonpregnant female: a step towards antibiotic stewardship." *BMC Infectious Diseases* 22.1 (2022): 939.

9. Ronald, Allan. "The etiology of urinary tract infection: traditional and emerging pathogens." *The American journal of medicine* 113.1 (2002): 14-19.
10. Bono, Michael J., Stephen W. Leslie, and Wanda C. Reygaert. "Uncomplicated urinary tract infections." *StatPearls [Internet]*. StatPearls Publishing, 2023.
11. Mueller, Matthew, and Christopher R. Tainter. "Escherichia coli infection." *StatPearls [Internet]*. StatPearls Publishing, 2023.
12. Álvarez, Manuel Díaz, et al. "Urinary tract infection caused by Enterobacteriaceae and its relationship with vesicoureteral reflux." *Boletín Médico Del Hospital Infantil de México (English Edition)* 74.1 (2017): 34-40.
13. Doern, Christopher D. "Classification of medically important bacteria." In *Molecular Medical Microbiology*, pp. 9-21. Academic Press, 2024.
14. Ashurst, J. V., and A. Dawson. "Klebsiella Pneumonia. StatPearls 2022." 2023.
15. Nguyen, Nhu Ngoc, and Thi Thu Hoai Nguyen. "GRAM-NEGATIVE BACTERIAL PATHOGENS." (2024).
16. Artero, E. Álvarez, et al. "Infección urinaria en el anciano." *Revista Clínica Española* 219.4 (2019): 189-193.
17. Lamas Ferreiro, J. L., Álvarez Otero, J., González González, L., Novoa Lamazares, L., Arca Blanco, A., Bermúdez Sanjurjo, J. R., ... & de la Fuente Aguado, J. (2017). Pseudomonas aeruginosa urinary tract infections in hospitalized patients: Mortality and prognostic factors. *PloS one*, *12*(5), e0178178.
18. Mueller, Matthew, and Christopher R. Tainter. "Escherichia coli infection." *StatPearls [Internet]*. StatPearls Publishing, 2023.
19. Habak, Patricia J., Karen Carlson, and Robert P. Griggs Jr. "Urinary tract infection in pregnancy." *StatPearls [Internet]*. StatPearls Publishing, 2024.
20. Pardeshi, Pritam. "Prevalence of urinary tract infections and current scenario of antibiotic susceptibility pattern of bacteria causing UTI." *Indian Journal of Microbiology Research* 5.3 (2018): 334-338.
21. Szweda, Hanna, and Marcin Jóźwik. "Urinary tract infections during pregnancy-an updated overview." *Dev Period Med* 20.4 (2016): 263-272.
22. Schnarr, J., and F. Smaill. "Asymptomatic bacteriuria and symptomatic urinary tract infections in pregnancy." *European journal of clinical investigation* 38 (2008): 50-57.

23. Szweda, Hanna, and Marcin Jóźwik. "Urinary tract infections during pregnancy-an updated overview." *Dev Period Med* 20.4 (2016): 263-272.
24. Millar, Lynnae K., and Susan M. Cox. "Urinary tract infections complicating pregnancy." *Infectious Disease Clinics* 11.1 (1997): 13-26.
25. Schnarr, J., and F. Smaill. "Asymptomatic bacteriuria and symptomatic urinary tract infections in pregnancy." *European journal of clinical investigation* 38 (2008): 50-57.
26. Foxman, Betsy. "Epidemiology of urinary tract infections: incidence, morbidity, and economic costs." *The American journal of medicine* 113.1 (2002): 5-13.
27. iani, Oriana, Daniele Grassi, and Rosanna Tarricone. "An economic perspective on urinary tract infection: the “costs of resignation”." *Clinical drug investigation* 33 (2013): 255-261.
28. François, M., et al. "The economic burden of urinary tract infections in women visiting general practices in France: a cross-sectional survey." *BMC health services research* 16 (2016): 1-10.
29. Iskandar, Katia, et al. "Economic burden of urinary tract infections from antibiotic-resistant Escherichia coli among hospitalized adult patients in Lebanon: a prospective cohort study." *Value in health regional issues* 25 (2021): 90-98.
30. Gillespie, Paddy, et al. "The cost effectiveness of the SIMPle intervention to improve antimicrobial prescribing for urinary tract infection in primary care." *Journal of Public Health* 39.4 (2017): e282-e289.
31. Sulaiman, Shabna, et al. "Clinical and Economic Burden of Early Urinary Tract Infection in Kidney Transplant Recipients." *Indian Journal of Transplantation* 18.3 (2024): 262-266.
32. Sabih, A., and S. W. Leslie. "Complicated urinary tract infections. 2023 Nov 12." *StatPearls. Treasure Island (FL): StatPearls Publishing* (2024).
33. Iskandar, Katia, et al. "Economic burden of urinary tract infections from antibiotic-resistant Escherichia coli among hospitalized adult patients in Lebanon: a prospective cohort study." *Value in health regional issues* 25 (2021): 90-98.
34. Schnarr, J., and F. Smaill. "Asymptomatic bacteriuria and symptomatic urinary tract infections in pregnancy." *European journal of clinical investigation* 38 (2008): 50-57.