THE USE OF PREDICTIVE ANALYTICS IN ASSISTING THE REDUCTION OF PATIENT WAITING TIMES: A STUDY OF COMMON RESPIRATORY AILMENTS.

ANNE AWELE NWAOKOLO
P2743914
Msc Data Analytics, De Montfort University, Leicester
September 1st 2023

Acknowledgements

I would want to use this opportunity to offer my sincere gratitude and appreciation to everyone who has helped my project be completed successfully. Throughout this process, their advice, assistance, and encouragement have been priceless.

I want to start by expressing my gratitude to my spouse Nnamdi Obi-Ijeh for his consistent support and trust in my skills. His unwavering support and encouragement have given me the drive to work on this project relentlessly.

I'm grateful to my dearest children Michelle and Michael Obi-Ijeh for their understanding and patience as I worked on this project for endless hours. Your steadfast assistance and comprehension of my obligations have been simply amazing.

Additionally, I want to thank my project supervisor, Sarah Greenfield, from the bottom of my heart. The direction and breadth of my research have been greatly influenced by her astute criticism, vast expertise, and invaluable counsel. Its success has been greatly aided by her constant commitment and support.

I also want to express my thanks to all the lecturers and module leaders who shared their knowledge and skills with me during my academic career. My comprehension of the topic of data analytics has significantly increased because of their dedication to quality and enthusiasm for teaching.

Last but not least, I want to express my sincere gratitude to the staff at The Health Informatics Service (THIS) for their tremendous help in supplying the data I needed for my study. Their knowledge and desire to work together have greatly aided in the success of this project as a whole.

I sincerely appreciate all of the support and guidance I got from each of these people. Their inputs significantly improved the quality and outcome of my project. I feel honoured to have been supported throughout this journey by such a remarkable set of people.

My spouse, kids, supervisor, module leaders, lecturers, and staff from The Health Informatics Service (THIS) are all once again acknowledged. I genuinely appreciate all of your help and support.

Regards, [Anne Awele Nwaokolo]

Abstract

This project assesses the use of artificial intelligence in healthcare and the role that predictive analytics can play in improving the healthcare system through the reduction of patient waiting time. The study reviews the data of patients with respiratory illness from two NHS hospitals, Calderdale Royal(CRH) and Huddersfield Royal(HRI) and accesses ailment prevalence and hospital waiting time(Length of Stay). According to the study's analysis of demographic data, older people between the ages of 71 and 90 were the group most commonly impacted by respiratory ailments. A minor prevalence of female patients was shown by the gender study. Time series analysis demonstrated greater patient visits in December and a consistent increase in the overall number of patients from 2019 to 2022, showing the effect of seasonal fluctuations. Asthma and pneumonia were found to be the most common respiratory illnesses, and high amount of influenza cases were also found. Blood tests, ECGs, and chest X-rays were some of the methods utilised in diagnostic testing to evaluate organ and respiratory functioning. Based on results from test of normality of the data, the length of hospital stay for respiratory patients had a large rightward skew, indicating that a significant number of patients had longer stays. The most prevalent patient complaints, such as asthma, shortness of breath, feeling unwell, sepsis, coughing, and chest pain, were also examined in the study. The study also looked at the relationships between factors including gender, age group, diagnosis, and length of stay through non-parametric testings, results from analysis of variance, median scores, Van Da Waerden and Wilcoxon were used to access the impact of the variables on Length of Stay (LOS). The findings demonstrated that while gender had no significant effect on length of stay, age group and diagnosis did.

Being that asthma and pneumonia had the highest number of cases for respiratory ailments, four machine learning models that can predict and pre-diagnose the ailments were developed . The models were, K-Nearest Neighbours (KNN), Random Forest (RF), Logistic Regression (LR), and Decision Tree (DT). The assessment's findings suggest that the Random Forest model scored highly for both illness prediction in accuracy, precision, recall, and F1 score. Based on these performance metrics, the Random Forest model may be considered to be the best model for predicting both pneumonia and asthma. To keep the models effective over time, they must be evaluated and improved continuously. The study emphasises how crucial feature selection and data preparation are to producing reliable prediction models. Overall, the research reveals that there is great promise for using predictive analytics to enhance patient care and solve inefficiencies in healthcare systems. Healthcare institutions may improve patient experiences by lowering wait times by using data-driven insights to make predictions ahead, diagnose respiratory illnesses, simplify operations, and distribute resources wisely.

Contents

List of Figures

Acronyms

A&E Accident and Emergency. 4, 6, 8

AI Artificial Intelligence. 4, 6

CRH Calderdale Royal. 4, 13

DT Decision Tree. 1, 4, 11

ED Emergency Departments. 4, 9

HES Hospital Episode Statistics. 4, 10

HRI Huddersfield Royal. 4, 13

ICU Intensive Care Unit. 4, 8

IT Information Technology. 4, 6

KNN K-Nearest Neighbours. 1, 4, 11

LOS Length of Stay. 1, 4, 12

LR Logistic Regression. 1, 4, 11

ML Machine Learning. 4, 6

NHS National Health Service. 4, 6, 9

PEDW Patient Episode Database for Wales. 4, 10

RF Random Forest. 1, 4

SMR Standard Mortality Ratio. 4, 6

SVM Support Vector Machine. 4, 11

THIS The Health Informatics Service. 4, 8

CHAPTER 1

INTRODUCTION

Several studies indicate a strong inverse relationship between patient satisfaction and waiting time. Patients should not have to wait too long for appointments and consultations in a well-designed healthcare system (Li *et al*, 2021). Longer wait times in emergency rooms and longer visits have a negative impact on care quality and raise the risk of bad outcomes (Horwitz *et al*, 2010). The Handbook to the National Health Service (NHS) Constitution has a commitment to the four-hour Accident and Emergency (A&E) waiting time objective. At least 95% of patients who visit A&E should be hospitalised, transferred, or released within four hours, according to the operational criteria established in 2010. A 76% intermediate threshold objective was adopted in December 2022 with a projected improvement of 2% in 2024/25(NHS, 2023). According to NHS. England, 347,703 patients had to wait 12 hours between the decision to admit them and their admission in 2022. The Royal College of Emergency Medicine (Beedle, 2023) suggests that there is one patient mortality for every 72 patients who wait 8 to 12 hours after arriving at an emergency room. By using the Standard Mortality Ratio Standard Mortality Ratio (SMR). to determine the total number of 12-hour time of arrival delays for 2022, we can estimate that there were 23,003 patient fatalities in England because of long wait times.

In England, respiratory illness is the third leading cause of mortality and affects one in five people (Clayton, *et al*). The mortality rate from respiratory disease in 2020 was 89 deaths per 100,000 women and 130 deaths per 100,000 men in England. In the winter, older adults are more vulnerable to respiratory illnesses. In England and Wales during the winter of 2019–20, underlying respiratory disorders were the likely cause of almost 8.3 thousand winter deaths among people over 75. About a sixth of people between the ages of 75 and 79 (Stewart, 2022).

Due to the introduction of Information Technology (IT), the healthcare industry has advanced more recently. IT. is used in healthcare with the goal of improving patient care while lowering costs and enhancing comfort (Alanazi, 2022). Predictive analytics is a subset of advanced analytics that makes predictions about the unknown. Predictive analytics examines current discoveries to create predictions about the future by utilizing a variety of approaches from data mining, statistics, modelling,Machine Learning (ML) and Artificial Intelligence (AI).
Various healthcare provider areas are supported by predictive analytics. It attempts to improve clinical outcomes, improve patient care, optimise resource use, and properly diagnose illnesses. By maximising the cost, predictive analytics aids organisations in planning for health care. The application of predictive analytics in this sector is anticipated to produce effective results by raising the level of Patient Satisfaction. The future of the health care sector will be transformed by predictive analytics (Nithya,2017).

This study aims to critically analyse the current wait time in NHS. hospitals, examine the demographics and other activities of patients with common respiratory illnesses with a view of introducing machine learning predictions that can be used to perform quick diagnosis of patients conditions prior to them seeing a doctor in order for quick checks and tests to be carried out as a way of reducing the amount of time spent in the hospital and improve the patients outcome.

1.1 Aim

The main aim of this study is to evaluate the waiting times(Length of Stay) that patients with common respiratory diseases experience at NHS.hospitals and to access the relationships between demographics and type of diagnosis how they relate to patient wait times or length of stay and to develop machine learning prediction models that medical personnel may use to pre-diagnose common respiratory ailments and reduce patient wait times.

CHAPTER 2

LITERATURE REVIEW

2.1 Predictive Analytics in Healthcare

Predictive Analytics is the specialty of making predictions about unknown future events. To explore recent findings and generate predictions about the future, predictive analytics employs a variety of techniques from data mining, statistics, modelling, machine learning, and artificial intelligence. Predictive analytics' fundamental component, the predictor, is described as a variable used to estimate future behaviour. Future probabilities are predicted using predictors, with incredibly accurate outcomes. Machine learning techniques and regression techniques are two categories of approaches used to undertake predictive analytics (Nithya, 2017). Prediction has several uses in the healthcare industry. To administer the emergency department more efficiently, machine learning techniques including Support Vector Machine, Linear Regression, Decision Tree, and Random Forest are utilised in hospital administration. Many diseases are predicted using machine learning algorithms(Rasjid, 2021).

In India, the use of cutting-edge technologies to perform operations on patient data in healthcare has greatly improved decision-making in the healthcare industry. The goal of health prediction is to foresee patients' health problems in advance. Patient records are now kept electronically rather than as hard copies, which has resulted in the availability of a vast amount of historical data in the healthcare industry. Healthcare analytics research is increasingly employing predictive models to identify patients' ailments using electronic health information gathered from various sources. Predictive models can be used for several tasks, including feature generation, feature selection, cross-validation, and data classification. To create an effective model, it is crucial to evaluate and enhance models created from a range of associations, patient-related information, and mathematical outlines (Mangat & Saini, 2021).
Health care organisations in India are diagnosing more ailments on a yearly basis. Anyone suffering from a disease would benefit from receiving the greatest care if the ailment could be predicted or detected early on. As a result, disease prediction has taken on more importance to help medical professionals provide patients with effective treatment (Mall *et al*, 2022)

In an article titled "Predictive Analytics: An Emerging Asset in the Healthcare Industry", The Health Informatics Service (THIS) an NHS. organisation which provides IT. and digital services through innovation and collaboration to health and care providers across the United Kingdom identifies the potential of predictive analytics in healthcare. According to the report, predictive analytics helps health institutions improve patient care, increase outcomes, and reduce costs. Predictive models can be used to decide the best times, places, and methods for providing care. Predicting the need for emergency care, including identifying and customising the optimal treatment plan for ailments when patients arrive at A&E, is feasible through thorough analysis of historical and statistical data. During the peak of Covid-19 pandemic in 2020 and 2021, predictive analytics solutions which were developed by THIS. were effectively used in several extremely difficult operating situations. The solutions aided NHS. hospital teams in preparing for and responding to significant spikes in demand for intensive care, which called for the addition of clinical personnel, beds, and equipment to the Intensive Care Unit (ICU) (THIS, 2022).

2.2 Patient wait time and its effect on patient outcomes and satisfaction.

Chu *et al* (2019) in a study at the United States defines wait time as the amount of time spent in the waiting and exam rooms while awaiting a clinician. The study accessed several factors which affect the perception of patient wait time and ways through which it can be improved. The research demonstrated that patient "willingness to wait" was a function of individual characteristics, including the perceived value of the visit and the cost of a prolonged wait, as well as clinic and provider factors. Some actions were identified by analyses which clinics and clinicians can take to improve the wait time, they include being proactive, giving patients opportunities for diversion and notifying patients of delays. In publicly funded healthcare systems like the England NHS, waiting times are a significant phenomenon. Waiting lists can be used to meet demand as it arises and to make the most use of the limited supply of resources, such as qualified personnel and medical facilities. However, issues occur in situations where waiting time may limit the patient's ability to benefit from the treatment (Reichert & Jacobs, 2018)

Longer stays in Emergency Departments (ED) are linked to worse outcomes and a higher patient fatality rate. However, patient demand for emergency treatment has been increasing globally due to ageing populations. Short throughput times become more challenging due to the increased load on Emergency Department. The length of Emergency Department visits has been measured and waiting time guidelines have been put in place in several nations and areas, including England, Australia, Stockholm, Scotland, and Northern Ireland. The NHS constitution in England mandates that 95% of ED patients be seen, treated, and either hospitalised or discharged within 4 hours. The percentage of emergency patients visiting English EDs who were there for less than 4 hours fell to a record low of 77% in the winter of 2018 after more than 5 years of deterioration. Reducing patient waiting times in emergency departments and getting back to the 95% constitutional requirement have received a lot of attention. (Paling *et al*, 2020).

The four-hour A&E waiting time aim is stated in the Handbook to the NHS Constitution. The operational requirements set in 2010 state that at least 95% of patients who enter A&E should be admitted, transferred, or released within four hours. In December 2022, a 76% intermediate threshold aim was set, with a 2% improvement expected in 2024–2025 (NHS, 2023). 347,703 patients had to wait 12 hours between the decision to admit them and their admittance in 2022, according to NHS. England.

2.3 The use of Artificial Intelligence in improving patient waiting time.

Some researchers have explored the possibility of using Artificial Intelligence to improve patient waiting time. In China, people wait for a long period while receiving diagnosis and treatment takes comparably little time. In a study by Li *et al* (2021), using Artificial Intelligence, a model was created which was called XIAO YI, a personalised inquiry and automatic diagnostic system to simulate a doctor's visit based on patient's primary complaints. It was suggested that it may assist outpatients in automatically ordering for tests. This meant that before visiting a doctor, patients might be evaluated or inspected. The period between registration and getting ready for lab tests or imaging examinations was also classified as waiting time. Electronic medical records were organised using natural language processing, which enables the automatic analysis of clinical data. The system utilized AI to recommend low-cost, non-invasive testing and shorten patient wait times. It was based on logistic regression classifiers. From the study, it was established that undergoing a laboratory test or imaging examination before seeing a doctor might greatly reduce patients' waiting time.

In another research by Li *et al* (2022), a programme called Smart-doctor that uses artificial intelligence (AI) to aid users was created. The research was done at Shanghai Children's Medical Centre in China. Participants were split into two groups, one with AI assistance and the other without. In the AI-assisted group, Smart-doctor was utilised as a medical assistant. An electronic medical satisfaction survey was requested to be completed following the visit. The queue time was the main result, while the consultation time, test time, overall time, and satisfaction score were the secondary objectives. Multiple linear regression, the Wilcoxon rank sum test, and ordinal regression were used. Results from the study revealed that patients' wait times can be reduced and visit satisfaction increased by using AI to streamline the outpatient care process.

In recent research by Benevento, *et al* (2023), Using Machine Learning methods, real data from two Italian hospitals were used to estimate wait times in emergency. Finding the most precise learning method for estimating patient wait times in emergency departments in real time was the objective. A simulated experiment was suggested to demonstrate the real-time use of predictive models in emergency departments by estimating waiting times in real-time for arriving patients using data on the system's present status. Due to newly built queue-based predictors that successfully capture the current ED state, the generated models were able to deliver precise waiting time predictions.

2.4 The need for urgent care for patients with respiratory ailments.

According to Naser *et al* (2021), due to exposure of the human lungs to airborne pollutants, irritants and infectious agents, respiratory ailments are among the world's major causes of morbidity and mortality. Respiratory illnesses cause 24% of all fatalities and 6.5% of hospital admissions in the UK. These illnesses lower the patients' quality of life and have a detrimental effect on societies. Respiratory illnesses, particularly respiratory tract infections, continue to have a significant negative impact on global health today. Numerous causes have impacted respiratory illnesses' startlingly high prevalence rates throughout the years. In the research, data on hospital admissions were gathered from the two main medical databases in England and Wales, the Patient Episode Database for Wales Patient Episode Database for Wales (PEDW) and the Hospital Episode Statistics Hospital Episode Statistics (HES) database in England. Both the PEDW. and the HES. database offered comprehensive data on hospital admissions for all illnesses linked to the respiratory system for individuals of all ages in England and Wales. The period in review was between April 1999 and March 2019. For the two decades, lung diseases saw the largest increase in hospitalisation rates, followed by influenza and pneumonia.

One in five persons in England are affected by respiratory illnesses, which are the third biggest cause of death (Clayton, *et al.*). In England, the mortality rate from respiratory illness was 89 per 100,000 women and 130 per 100,000 men in 2020. Older persons are especially susceptible to respiratory diseases in the winter. Underlying respiratory problems were the most likely cause of about 8.3 thousand winter fatalities among adults over 75 in England and Wales during the winter of 2019–20 (Stewart, 2022). In an analysis of data obtained from NHS digital, pneumonia and influenza, which are common respiratory illnesses, totalled 414,802 cases and are among top 20 ailments for consultancy episodes in England, according to data from NHS Digital's summary of primary diagnosis for finished consultancy episodes for 2021 to 2022. (NHS, 2021).

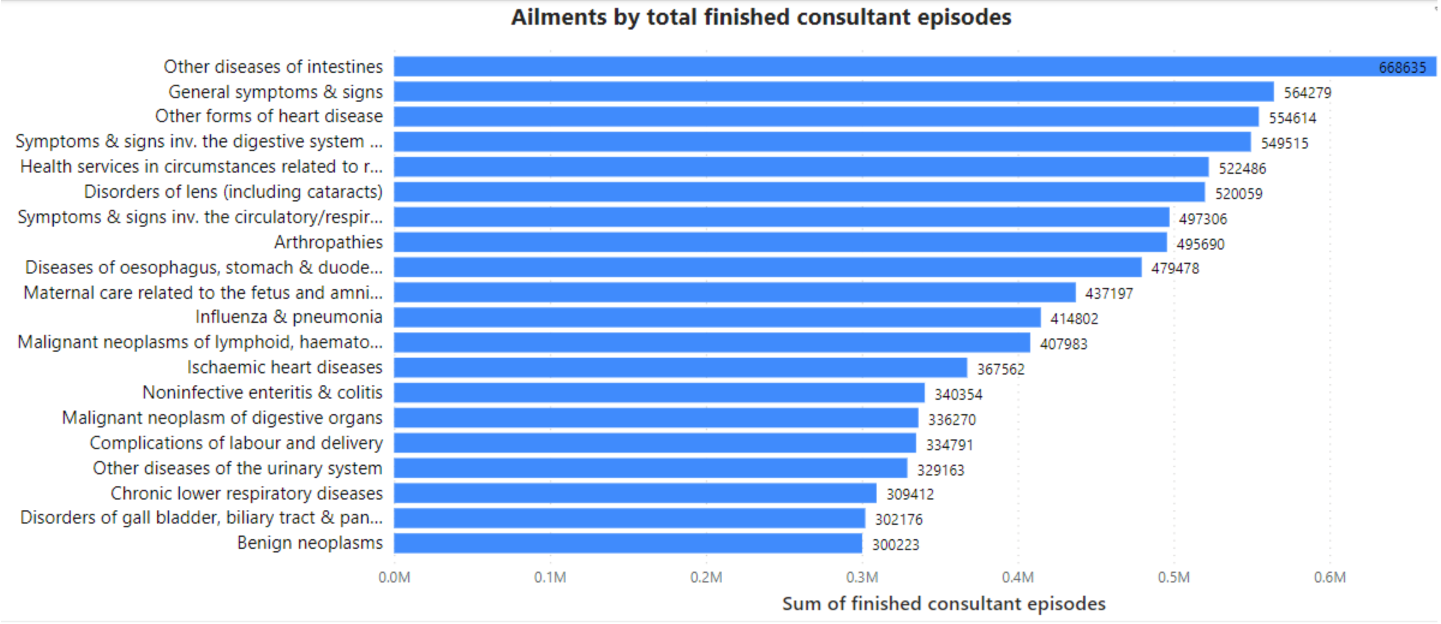


Figure 2.1: Ailments by total finished consultant episodes 2021 to 2022

2.5 Predictive Analytics for respiratory ailment detection

Li *et al* (2022) developed a model that successfully predicted and diagnosed bronchiectasis, pulmonary embolism, pulmonary TB, and chronic obstructive pulmonary disease based on the respiratory illness big data platform in southern Xinjiang China and assisted primary care providers. To predict and diagnose respiratory illnesses, the approach used combined long-short-term memory network (LSTM) with convolutional neural network (CNN) which are both deep learning methods that are used to make predictions. The major complaints, past medical history, and chest computed tomography were all taken from the medical records of inpatients in the respiratory unit and used to build the models.

In India, Study by Prasad *et al* (2011) suggests employing expert systems and machine learning to diagnose asthma, including auto-associative memory neural networks, bayesian networks. Using patient information and clinical asthma signs and symptoms acquired from multiple sources, the study assesses the effectiveness of these algorithms. According to the findings, auto-associative memory neural networks focus on associative learning and information retrieval. These networks are among the artificial neural network types that have been studied the most and are among the top algorithms for detecting asthma among those that were put to the test.

In a study by Poreva *et al* (2017) in Ukraine, many classifiers were looked at and analysed for the diagnosis of lung illnesses including K-Nearest Neighbours (KNN), Decision Tree (DT),Support Vector Machine (SVM), naïve bayesian classifiers, and Logistic Regression (LR) methods. Based on the set of lungs sounds, several signal characteristics were obtained. Using the five distinct machine learning techniques, the study's objective was to categorise sounds. The greatest precision signal parameters had to be found among a variety of signal parameters. Thus, the seven lung sound characteristics that are most useful for diagnosis were identified. The reference vectors and decision tree techniques of machine learning was found to have the highest levels of accuracy. As a result, a pulmonary physician may use this categorization method as a diagnostic tool.

Summarily, predictive analytics, which employs methods from data mining, statistics, and machine learning to create precise predictions about the future, has become a useful tool in the healthcare industry. Predictive analytics have been used in the healthcare sector to boost operational effectiveness, optimise resource allocation, and improve patient care. Disease prediction and hospital management have both benefited from the use of machine learning techniques. Identification and prediction of patient ailments have been facilitated using algorithms and electronic health data. In healthcare settings, patient wait times have been highlighted as a significant factor, and the use of artificial intelligence has showed prospects in resolving this problem. Additionally, respiratory disease detection and diagnosis have benefited from the use of predictive analytics, which has led to better healthcare outcomes.

Predictive analytics has the potential to change the way healthcare is provided because it allows for data-driven decision-making and improves patient experiences. Given the current poor wait time within the NHS. and the number of patients with episodes for respiratory ailments such as pneumonia and influeza, a study which investigates the use of machine learning to predict common respiratory illness with a view to reducing patient wait time will be a great idea and the current healthcare system can benefit from this.

CHAPTER 3

RESEARCH METHODOLOGY

The philosophical frameworks used in this study include positivism and phenomenology. Positivism relies heavily on quantitative observations and statistical analysis (Collins 2018). Microsoft power BI is used for data visualisation and exploration. SAS programming is used for statistical testing to explore the association between variables and patients' Length of Stay (LOS). Python is used to implement machine learning modelling.

3.1 Methods

- Data analysis and visualisation are done in the first stage using Power BI. With the use of this application, interactive dashboards can be made, allowing for a thorough knowledge of the information and the identification of any potential patterns or trends that could affect the length of time patients must wait to be seen for respiratory conditions.
- SAS programming is then used to do statistical tests. To determine the importance and impact of patients' primary complaint,demographics, final diagnosis, and other factors on waiting times,hypothesis testing are used. These statistical analyses provide light on the connections between various factors and patient length of stay.
- Python is used to create machine learning models that can provide a better option for diagnosis through predictive analytics. The K-Nearest Neighbour(KNN), Decision Tree(DT), Random Forest(RF) and Logistic Regression(LR) algorithms are implemented and assessed. To accurately predict patient outcomes, these models are trained and evaluated using a portion of the dataset to predict respiratory disease.

This study intends to give a thorough knowledge of patient waiting times for respiratory diseases by merging the tools of Power BI, SAS programming, and Python for data analysis, statistical testing, and machine learning modelling. The study's prediction skills are improved via the use of several machine learning algorithms, allowing healthcare practitioners to make wise judgements and efficiently allocate resources.

CHAPTER 4

RESEARCH ANALYSIS AND FINDINGS

4.1 Data exploration and analysis

4.1.1 Dataset Description.

The dataset used in this project was provided by The Health Informatics Services (THIS) The data contains a list of 6774 patients with respiratory diagnoses who visited emergency departments of either Calderdale Royal (CRH) or Huddersfield Royal (HRI) hospitals in England between April 2022 and March 2023.

The excel file contains three worksheets, in the main one, called Resp Patients, has all the patient information; the other two, called Events and Orders, contain additional information about the visit. Events include everything from NEWS scores to blood pressure readings to temperature readings and more. Orders includes all requests for medication, X-rays, ultrasounds, and other procedures.

Description of variables

Resp patients' sheet

- **EncounterSK**- Unique ID of each patient which in numerical format
- **MonthCommence**- Details of Month and Year
- **DaysofWeek**- The day of the week when the visit was made.
- **TimeofArrival**- The time when the patient arrived the hospital.
- **TimeofDeparture**- Time when the patient left the hospital or was admitted.
- **LOS(Length of stay)**- the amount of time the patient stayed in the hospital in minutes which is the same as the patient wait time.
- **PresentingComplaint**- The patient complaint or symptoms experienced.
- **DiagnosisDescription**- The final diagnosis of the patient.
- **ArrivalMode**- How the patient arrived the hospital either through ambulance or other means.
- **TriagePriority**-
- **AgeOnArrival**- The age of patient which has been categorised into groups, 0-5, 6-12, 13-19, 20-24, 25-34, 25-44, 45-54, 55-64, 65-70,71-80, 81-90, 91-99, 100 and above.
- **Gender**- Either Male or Female
- **Facility**- either Calderdale Royal (CRH) or Huddersfield Royal (HRI)
- **Disposal**- The way the visit ended, either through admission, discharge with follow up, death of the patient, patient leaving before being attended to, or transferred to another facility.

Events sheet

- **EncounterSK** - Unique numeric identification patient number
- **EventTime**- The time the event occurred.
- **EventType**- The type of event eg BP check, Pulse rate, heart rate etc
- **EventSK**- unique ID of each order

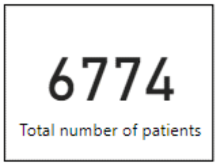
The Orders sheet

- **EncounterSK**- Unique ID of each patient in numerical format
- **OrderPlacedTime**- The time the order was placed.
- **CatalogDescription**- The type of test that was ordered.
- **OrderStatus**- The status of the order, it was ordered, completed or in process.
- **OrderSK**- Unique ID of each other

4.1.2 Creating measures

For Initial analysis and visualisation, the data is imported into power BI and measures are created by calculating the total number of patients and total number of tests ordered.

- **Total number of patients**- This is gotten using a distinct count of the EncounterSK of the respiratory patients' table.



6774

Total number of patients

Figure 4.1: total number of patients

- **Total number of tests ordered** – This is gotten using a distinct count of the Order SK of the test ordered table.

4.1.3 Demographic analysis

Patients by gender

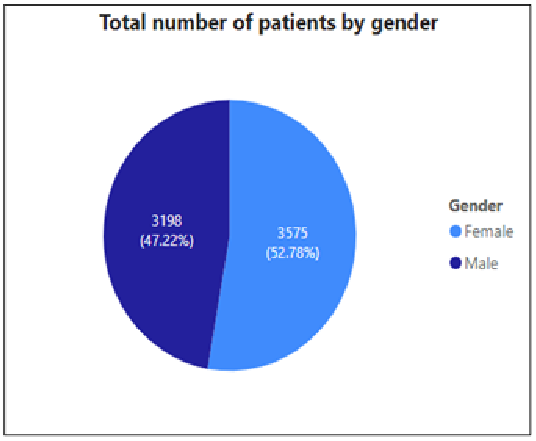


Figure 4.2: Patients by gender

Analysis show that more females than males visited the hospitals for respiratory related illness with a total of 3,575 patients for females and 3,198 for males which shows that females were 377 more than males.

Patients by Age group

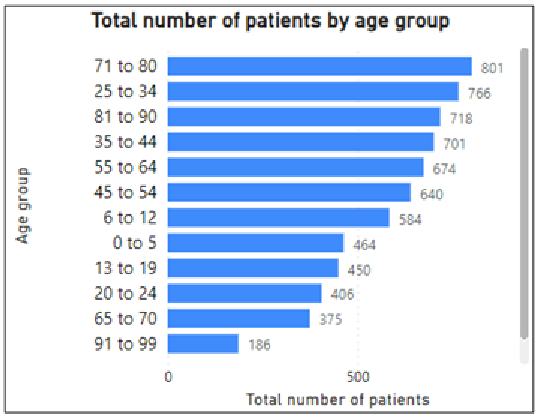


Figure 4.3: Patients by Age group

A review of the statistics of patients reveals that patients within the ages of 71-80 had the most hospital visits due to respiratory ailments, this was closely followed by ages 25 to 34 and 81 to 90, this means that elderly people from 71 to 90 are the most affected.

4.1.4 Analysis by Day, Month and Year

Number Of patients by day of week

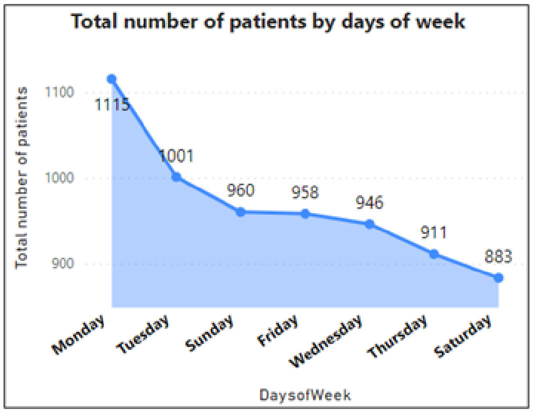


Figure 4.4: Patients by day of the week

Observation from the data shows that the most visits were experienced on Mondays while Saturdays had the least frequent visits for patients with respiratory illnesses.

Number Of patients by Month

A review of the data by month shows that the highest number of respiratory patient visit occurred in the month of December with a total of 957.

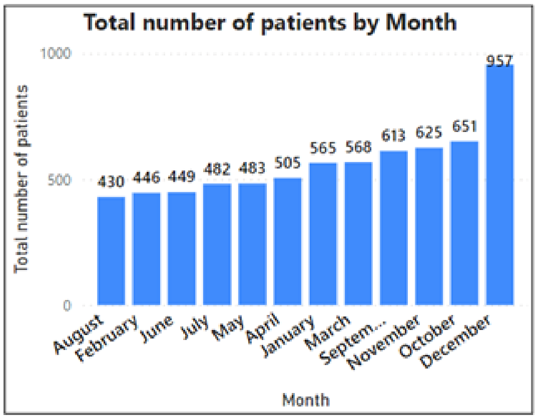


Figure 4.5: Patients by Month

Number of Patients by Year

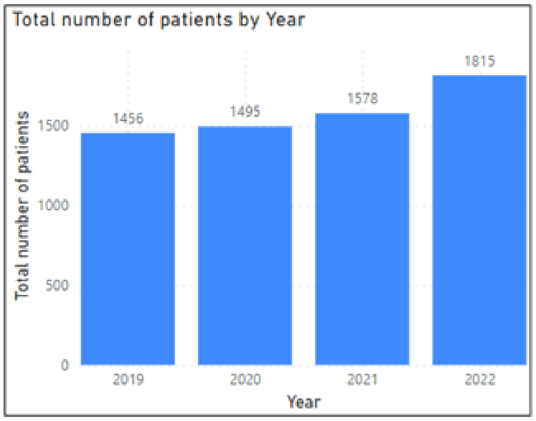


Figure 4.6: Patients by Year

A review of the data shows a continuous increase in the number of patients with respiratory illness from 1456 patients in 2019 to 1815 patients in 2022.

4.1.5 Analysis by patient complaint and diagnosis

Top 10 patient complaint

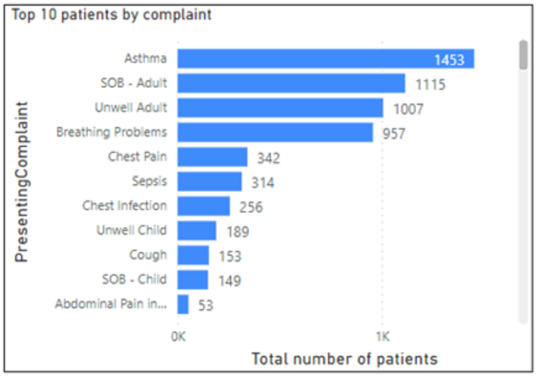


Figure 4.7: Top 10 patient complaint

In the dataset patients had several complaints or symptoms which led them to the hospital, but for the purpose of this visualisation analysis, only top 10 complaints are identified, they include asthma which is also a diagnosis, Shortness of breath (SOB) in adult, unwell adult, breathing problems, chest pain, sepsis, chest infection, unwell child, cough, SOB child and abdominal pain.

Number of patients by diagnosis

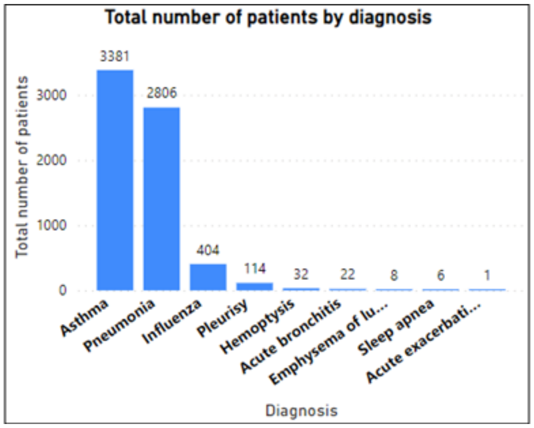


Figure 4.8: Number of patients by diagnosis

Analysis of type of respiratory ailments shows that the ailments with the greatest number of hospital visits are patients diagnosed with Asthma and Pneumonia with 3,381 and 2,806 respectively. Influenza also has a large number as compared to the rest of the respiratory ailments.

4.1.6 Analysis of diagnostic testing

Top 10 diagnostic testing

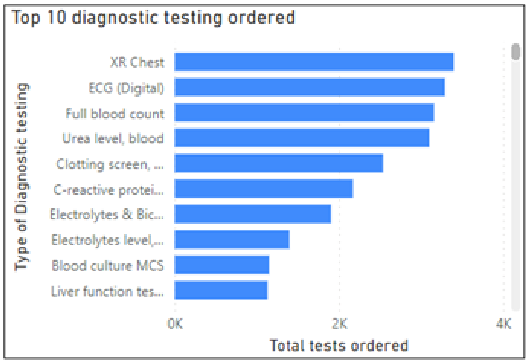
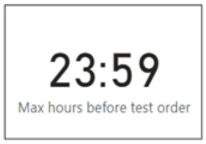


Figure 4.9: Top ten diagnostic testing

To get the details of diagnostic testing, other orders such as medications were filtered off to focus more on medical examinations and testing such as blood tests X-rays etc. Top 10 tests include: XR chest, ECG, full blood count, urea level, clotting screen, C-reactive protein, electrolytes and bicarbonate level, electrolyte level, blood culture MCS and liver function test.

Maximum number of hours spent before test order

Using the difference in the time of arrival and time of test order the total hours spent before test order was calculate and a review of hours shows that the highest time spent before test order is 23 hours.



23:59

Max hours before test order

Figure 4.10: Maximum hours spent before test order

4.2 Statistical Analysis with SAS

To access the relationship between variables such as gender, sex, age group and disgnosis and how they affect patients' Length of Stay(LOS) in the hospital, the data is imported to SAS for statistical analysis and testing.

4.2.1 Creating the Library name File name and importing the data

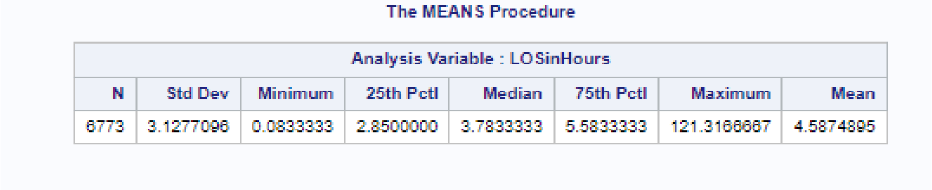
Library with name 'PROJECT' is created while file name resppat is created, the file which is named as resppat.csv is imported into the sheet, the data is named as 'PROJECT.resppat' for the file to be recognised.

4.2.2 Creating a new variable for Length of Stay(LOS) in hours

A new variable for Length of stay in hours is created for easy understanding and then columns that are not necessary for the analyses are removed using the "keep" statement. The LOS in hours is gotten by dividing it by 60 and then rounding it up and removing the decimals. The important variables are Presenting Complaint, Diagnosis, TriagePriority, Age, Gender and LOSinHours.

4.2.3 Statistical summary of LOS in hours

Using the Proc means statement, we get the summary of length of stay in hours



The MEANS Procedure

Analysis Variable : LOSinHours							
N	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum	Mean
6773	3.1277096	0.0833333	2.8500000	3.7833333	5.5833333	121.3166667	4.5874895

Figure 4.11: Statistical summary of length of stay in hours

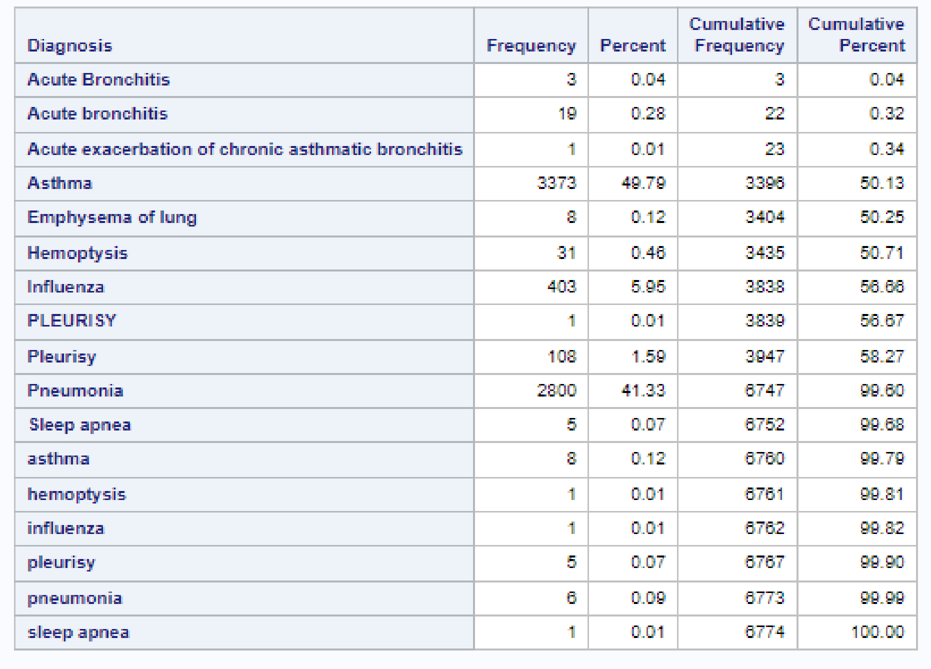
There have been 6773 observations of the variable.

- **Standard deviation**: The variable's standard deviation is around 3.1277096.
- **Minimum**: The variable's minimum value, 0.083333, represents the shortest period of stay.
- **25th Percentile**: The data's 25th percentile value, 2.8500000, is below this figure in 25% of the cases.
- **Median**:The dataset's median value, 3.7833333, is the point where 50% of the values fall below it and 50% rise beyond it.
- **75th Percentile**: The data falls within 75% of this value at a 75th percentile of 5.5833333.
- **Maximum**: The variable's highest value of 121.3166667 represents the longest period of stay.
- **Mean** The average value for the variable is approximately 4.5874895.

4.2.4 Frequency of the variables

Using the Proc freq statement we generate the frequencies of the class variables for analysis

Frequency of diagnosis



Diagnosis	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Acute Bronchitis	3	0.04	3	0.04
Acute bronchitis	19	0.28	22	0.32
Acute exacerbation of chronic asthmatic bronchitis	1	0.01	23	0.34
Asthma	3373	49.79	3396	50.13
Emphysema of lung	8	0.12	3404	50.25
Hemoptysis	31	0.46	3435	50.71
Influenza	403	5.95	3838	56.66
PLEURISY	1	0.01	3839	56.67
Pleurisy	108	1.59	3947	58.27
Pneumonia	2800	41.33	6747	99.60
Sleep apnea	5	0.07	6752	99.68
asthma	8	0.12	6760	99.79
hemoptysis	1	0.01	6761	99.81
influenza	1	0.01	6762	99.82
pleurisy	5	0.07	6767	99.90
pneumonia	6	0.09	6773	99.99
sleep apnea	1	0.01	6774	100.00

Figure 4.12: Frequency of diagnosis

1. **Acute bronchitis**-There were a total 22 instances of acute bronchitis, or 32% of all cases.
2. **Acute exacerbation of chronic asthmatic bronchitis** - Only one case which is 0.01% of all cases was reported. There are 23 cumulative frequencies, or 0.34% of all frequencies.
3. **Asthma**-3,373 instances of asthma, or 49.79% of the total, were reported.
4. **Lung emphysema** - There were 8 instances, or 0.12% of the total, with this condition.
5. **Haemoptysis** -31 instances, or 0.46% of the total, were reported.
6. **Influenza**-403 cases of influenza, or 5.95% of the total, were reported.
7. **Pleurisy** -109 instances, or 1.59% of the total, were reported.
8. **Pneumonia**-2,800 cases, or 41.33% of the total, were reported. 6,747 cumulative frequencies, or 99.60% of the total, make up the frequency.
9. **Sleep apnea** - There were 6 cases of sleep apnea, or 0.09% of all patients.

Frequency of top 10 presenting complaints

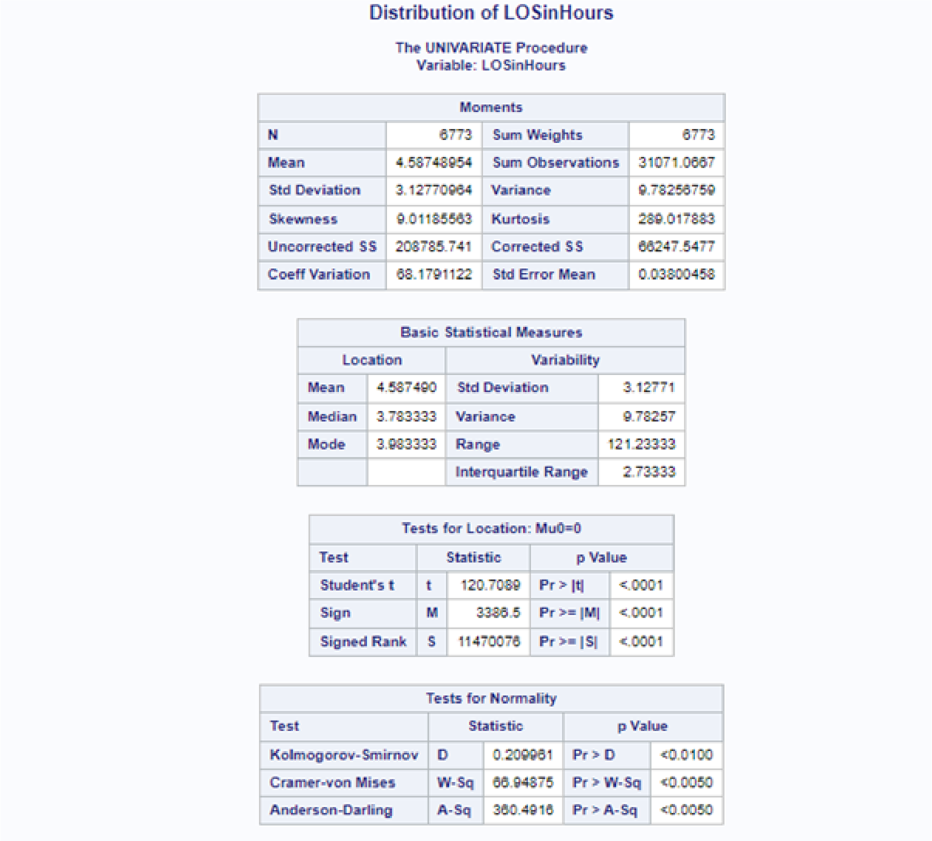
For this research, due to the high number of presenting complaints, only the frequency of the top 10 complaints are analysed.

1. **Asthma**: The most common presenting complaints, with a total of 1453 occurrences accounting for 21.45% of all patients, is asthma. This shows that many people with asthma-related problems seek medical attention.
2. **SOB Adult (Shortness of Breath in Adult)**: With a frequency of 1115 and 16.46% of patients reporting it, shortness of breath is the second most frequent complaint in adults. It implies that a sizable percentage of people require medical treatment for respiratory issues.
3. **Unwell Adult**: Adults who complained of feeling unwell accounted for 14.87% of visits with a frequency of 1007 making them a sizable category of patients. People who feel "unwell" may experience a variety of symptoms or a general ailment, which may prompt them to seek medical attention.

4. **Breathing Problems**: The frequency of 957 instances, and 14.13 percent of visits, indicates that breathing issues are a substantial worry for patients. In addition to asthma, this area includes other respiratory conditions.
5. **Sepsis**: -Sepsis, which is a severe systemic illness, accounted for frequency of 314 and 4.64% of visits. This emphasises the need of early diagnosis and treatment of this potentially fatal illness.
6. **Chest Pain**: With 338 instances (4.99%), chest pain is a frequent complaint for patients with respiratory ailments. Due to its possible connection to cardiac or other severe problems, it demands care.
7. **Cough**:There were 153 instances and 2.26% of patients with respiratory ailments complained of cough.
8. **Unwell child**: Children who visited the hospital with "Unwell" complaints accounted for 2.79% of visits, or 189 instances. This group includes a variety of paediatric illnesses that need timely medical intervention.
9. **Chest infection**: 256 cases or 3.78% of visits report chest infection as presenting complaints. This group comprises several chest-related respiratory illnesses including pneumonia or bronchitis.
10. **Falls**: With 52 recorded instances (0.77%), falls are a frequent cause for medical evaluation. This category includes fall-related injuries, which can range in severity from mild to severe and need medical care.

4.2.5 Testing the data for normality of the distribution

The distribution of the data is tested using proc univariate as this will help determine the best type of statistical model to adopt.



Distribution of LOSinHours

The UNIVARIATE Procedure
Variable: LOSinHours

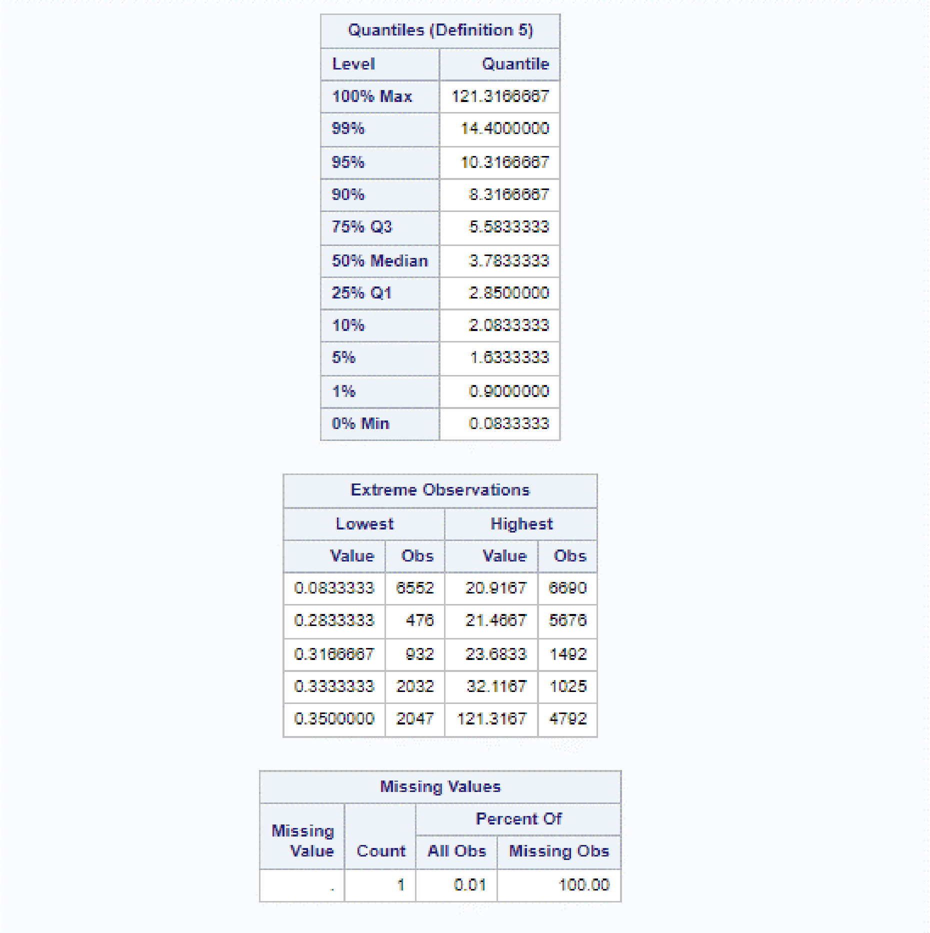
Moments			
N	6773	Sum Weights	6773
Mean	4.58748954	Sum Observations	31071.0667
Std Deviation	3.12770964	Variance	9.78256759
Skewness	9.01185563	Kurtosis	289.017883
Uncorrected SS	208785.741	Corrected SS	66247.5477
Coeff Variation	68.1791122	Std Error Mean	0.03800458

Basic Statistical Measures			
Location		Variability	
Mean	4.587490	Std Deviation	3.12771
Median	3.783333	Variance	9.78257
Mode	3.983333	Range	121.23333
		Interquartile Range	2.73333

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	120.7089	Pr > |t|	<.0001
Sign	M	3386.5	Pr >= |M|	<.0001
Signed Rank	S	11470076	Pr >= |S|	<.0001

Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.209961	Pr > D	<0.0100
Cramer-von Mises	W-Sq	66.94875	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	360.4916	Pr > A-Sq	<0.0050

Figure 4.13: Summary of moments, basic statistical measures,test for location,tests for normality



Quantiles (Definition 5)	
Level	Quantile
100% Max	121.3166667
99%	14.4000000
95%	10.3166667
90%	8.3166667
75% Q3	5.5833333
50% Median	3.7833333
25% Q1	2.8500000
10%	2.0833333
5%	1.6333333
1%	0.9000000
0% Min	0.0833333

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0.0833333	6552	20.9167	6690
0.2833333	476	21.4667	5676
0.3166667	932	23.6833	1492
0.3333333	2032	32.1167	1025
0.3500000	2047	121.3167	4792

Missing Values			
		Percent Of	
Missing Value	Count	All Obs	Missing Obs
.	1	0.01	100.00

Figure 4.14: Summary of quantiles, extreme observations and missing values

Understanding the distribution of LOSinHours is made possible by analysing the SAS findings.

Basic statistical measurements including the mean, standard deviation, skewness, and kurtosis are briefly discussed in the moments section. The average length of stay in this instance appears to be about 4.6 hours, according to the mean value of 4.58749. The data have a considerable level of variability, as shown by the standard deviation of 3.12771. The distribution appears to be strongly skewed to the right, according to the 9.01186 skewness, which denotes a large amount of asymmetry. A non-normal distribution is further supported by the fact that the kurtosis of 289.01788 shows a significant degree of peakness.

Additional details regarding the location and variability of the data are provided in the section on basic statistical measures. The range shows the variation between the minimum and greatest values, which are respectively 0.08333 hours and 121.31667 hours. The median, which represents the middle of the data, is 3.78333 hours. The interval between the 25% percentile (Q1) and the 75% percentile (Q3) is known as the interquartile range, and it is 2.73333 hours. The dispersion and central tendency of the data are better understood thanks to these metrics.

The tests for location section looks at how considerably the data deviates from a certain mean or median. Both the Student's t-test and the sign test in this situation provide p-values of less than 0.0001, suggesting a significant departure from the corresponding median and mean values of 0. Additionally, the signed rank test yields a p-value of less than 0.0001, which denotes a significant deviation from the median value of 0. These evaluations demonstrate that the data are not centred around 0. If the data has a normal distribution, it will be examined in the tests for normality section. A considerable deviation from a normal distribution is shown by the Kolmogorov-Smirnov test, the Cramer-von Mises test, and the Anderson-Darling test, which all have p-values of less than 0.05. These results indicate that the data is not normally distributed.

A summary of the different percentiles in the data distribution is provided in the quantiles section. For instance, if the

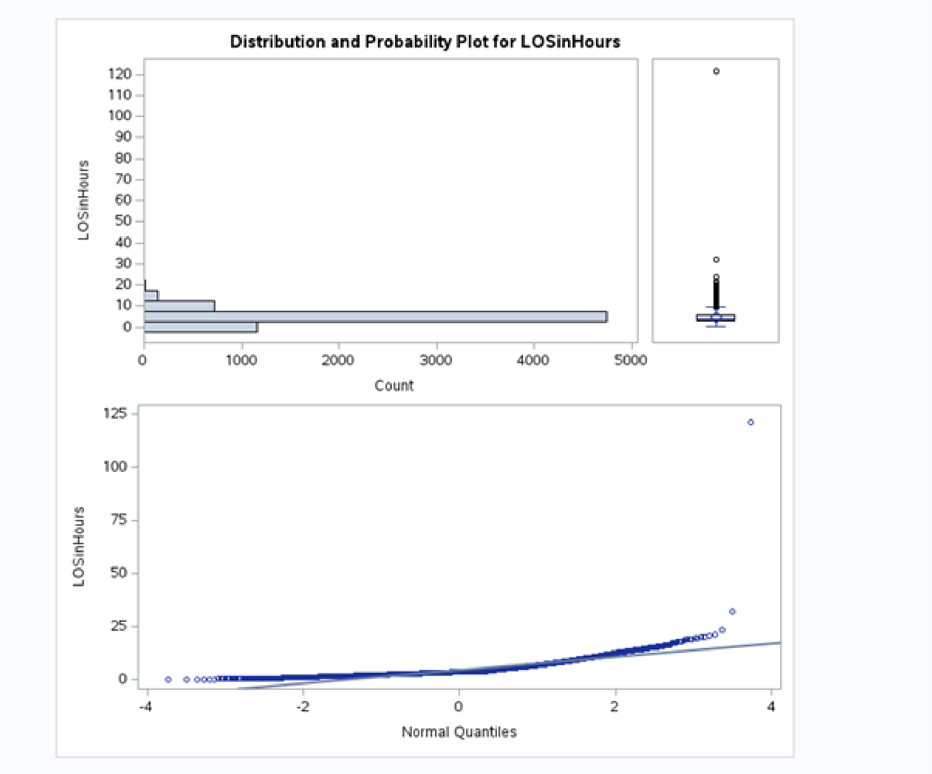


Figure 4.15: Distribution and probablity plots

median (50th percentile) is 3.78333, then 50% of the observations fall below this mark. Higher numbers are shown as we proceed up the percentile scale, with the 99th percentile being at 14.4.

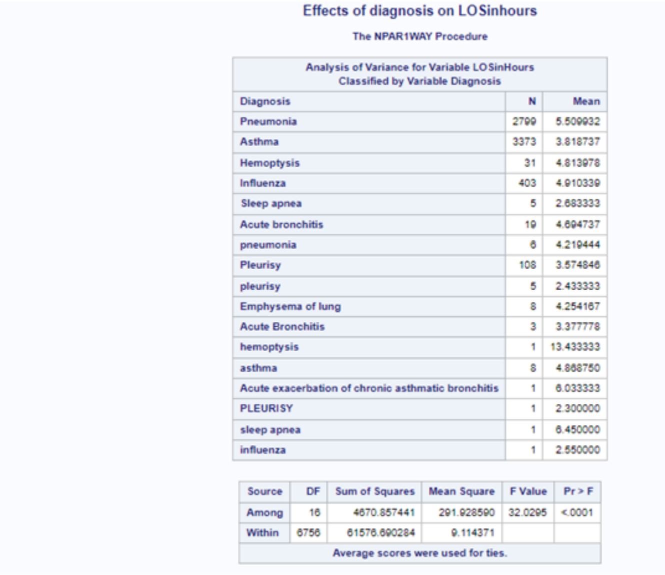
According to the results, the LOSinHours data distribution shows a significant amount of skewness and kurtosis, which point to a non-normal distribution. As a result, conventional parametric tests that rely on normalcy might not be suitable. Instead, when analysing this dataset, non-parametric tests or robust approaches may produce more accurate findings.

When certain normality assumptions are not satisfied, commonly used statistical tests usually perform poorly and have a higher error rate. Non-parametric tests are developed to have the appropriate statistical properties when just a few assumptions about the underlying distribution of the data may be made. That is to say, when data are taken from a non-normal distribution or one that contains outliers, a non-parametric test is typically a more useful statistical tool than a parametric test(Pappas and DePuy, 2004)

In summary, there is a large rightward skew in the distribution of LOSinHours throughout the whole range of values in this dataset which is an indication that more patients had longer stays than shorter stays. In comparison to a normal distribution, the data dramatically deviates. To secure reliable and correct inferences from the analysis of this dataset, non-parametric tests or robust approaches would be preferable.

4.2.6 Accessing the relationship between the variables and LOS

Nonparametric tests are run using the NPAR1WAY technique to look for scale and location variations across a one-way categorization. Additionally, PROC NPAR1WAY offers tests based on the empirical distribution function and a common analysis of variance on the raw data. Based on the following scores for a response variable, PROC NPAR1WAY runs tests for location and scale differences: Wilcoxon, median, Van der Waerden (normal), Savage, Siegel-Tukey, Ansari-Bradley, Klotz, Mood, and Conover. PROC NPAR1WAY also offers exams that assign grades based on the supplied data's raw form. Tests based on simple linear rank statistics are used when the data are divided into two samples(SAS, 2013). In this project, results from Analysis of Variance, Median scores, Van Da Waerden and Wilcoxon scores will be used to access the effect of the variables on Length of Stay.



Effects of diagnosis on LOSinhours

The NPAR1WAY Procedure

Analysis of Variance for Variable LOSinHours
Classified by Variable Diagnosis

Diagnosis	N	Mean
Pneumonia	2799	5.509932
Asthma	3373	3.818737
Hemoptysis	31	4.813978
Influenza	403	4.910339
Sleep apnea	5	2.683333
Acute bronchitis	19	4.694737
pneumonia	6	4.219444
Pleurisy	108	3.574846
pleurisy	5	2.433333
Emphysema of lung	8	4.254167
Acute Bronchitis	3	3.377778
hemoptysis	1	13.433333
asthma	8	4.868750
Acute exacerbation of chronic asthmatic bronchitis	1	6.033333
PLEURISY	1	2.300000
sleep apnea	1	6.450000
influenza	1	2.550000

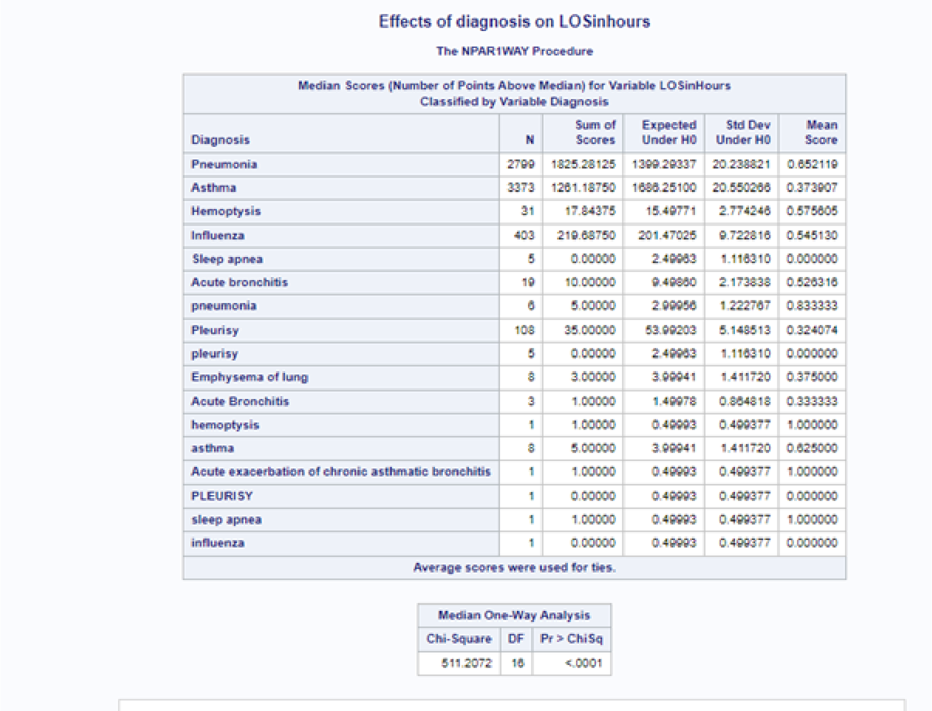
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Among	16	4670.857441	291.928590	32.0295	<.0001
Within	6756	61576.690284	9.114371		

Average scores were used for ties.

Figure 4.16: Analysis of Variance for LOSinHours by diagnosis

Effect of diagnosis on Length of stay(LOS)

- **Analysis of Variance**- This test checks to see if the mean LOSinHours differs significantly across various diagnosis. There are substantial variations in LOSinHours between diagnoses, as shown by the F-value of 32.0295 and the corresponding p-value of 0.0001.



Effects of diagnosis on LOSinhours

The NPAR1WAY Procedure

Median Scores (Number of Points Above Median) for Variable LOSinHours
Classified by Variable Diagnosis

Diagnosis	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Pneumonia	2799	1825.28125	1399.29337	20.238821	0.652119
Asthma	3373	1261.18750	1686.25100	20.550266	0.373907
Hemoptysis	31	17.84375	15.49771	2.774246	0.575605
Influenza	403	219.68750	201.47025	9.722816	0.545130
Sleep apnea	5	0.00000	2.49963	1.116310	0.000000
Acute bronchitis	19	10.00000	9.49860	2.173838	0.526316
pneumonia	6	5.00000	2.99956	1.222767	0.833333
Pleurisy	108	35.00000	53.99203	5.148513	0.324074
pleurisy	5	0.00000	2.49963	1.116310	0.000000
Emphysema of lung	8	3.00000	3.99941	1.411720	0.375000
Acute Bronchitis	3	1.00000	1.49978	0.864818	0.333333
hemoptysis	1	1.00000	0.49993	0.499377	1.000000
asthma	8	5.00000	3.99941	1.411720	0.625000
Acute exacerbation of chronic asthmatic bronchitis	1	1.00000	0.49993	0.499377	1.000000
PLEURISY	1	0.00000	0.49993	0.499377	0.000000
sleep apnea	1	1.00000	0.49993	0.499377	1.000000
influenza	1	0.00000	0.49993	0.499377	0.000000

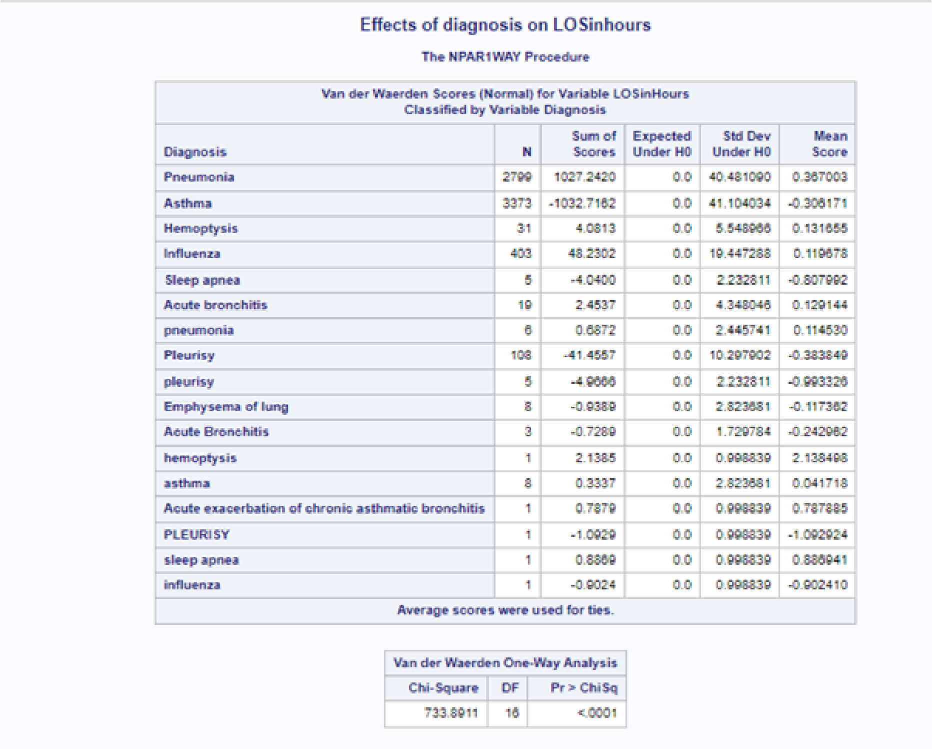
Average scores were used for ties.

Median One-Way Analysis

Chi-Square	DF	Pr > ChiSq
511.2072	16	<.0001

Figure 4.17: Median scores for LOSinhours by diagnosis

- **Median Scores**- This test determines if there is a statistically significant difference in the median LOSinHours values between various diagnosis. The Chi-square value of 511.2072 and the p-value of 0.0001 show that the medians of LOSinHours across diagnosis differ significantly from one another.



Effects of diagnosis on LOSinhours

The NPAR1WAY Procedure

Van der Waerden Scores (Normal) for Variable LOSinHours
Classified by Variable Diagnosis

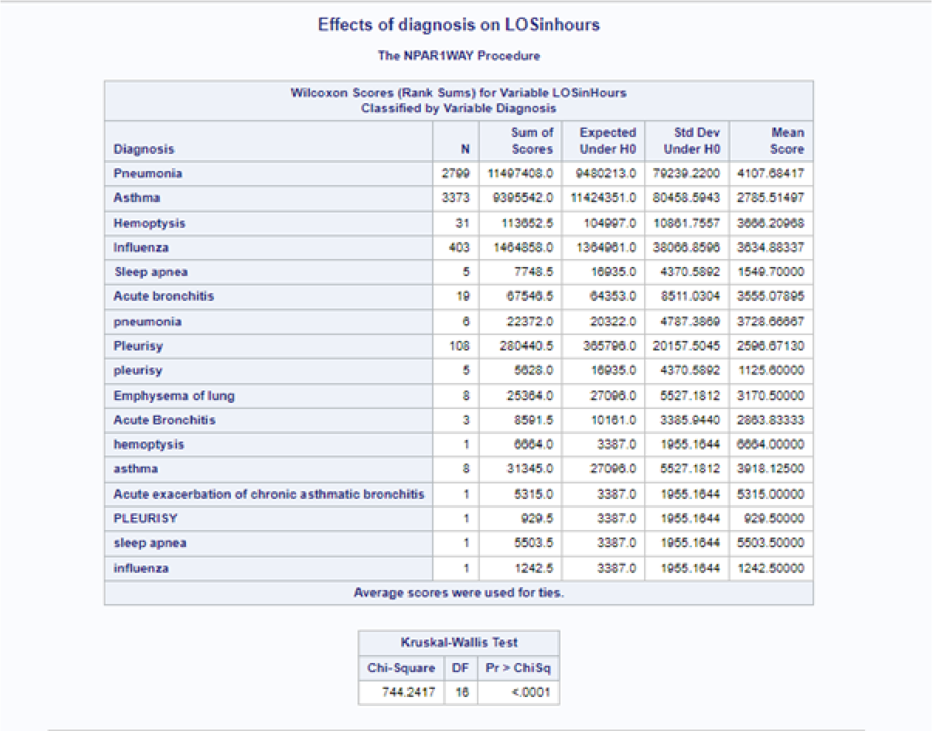
Diagnosis	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Pneumonia	2799	1027.2420	0.0	40.481090	0.367003
Asthma	3373	-1032.7162	0.0	41.104034	-0.306171
Hemoptysis	31	4.0813	0.0	5.548966	0.131655
Influenza	403	48.2302	0.0	19.447288	0.119678
Sleep apnea	5	-4.0400	0.0	2.232811	-0.807992
Acute bronchitis	19	2.4537	0.0	4.348046	0.129144
pneumonia	6	0.6872	0.0	2.445741	0.114530
Pleurisy	108	-41.4557	0.0	10.297902	-0.383849
pleurisy	5	-4.9666	0.0	2.232811	-0.993326
Emphysema of lung	8	-0.9389	0.0	2.823681	-0.117362
Acute Bronchitis	3	-0.7289	0.0	1.729784	-0.242962
hemoptysis	1	2.1385	0.0	0.998839	2.138498
asthma	8	0.3337	0.0	2.823681	0.041718
Acute exacerbation of chronic asthmatic bronchitis	1	0.7879	0.0	0.998839	0.787885
PLEURISY	1	-1.0929	0.0	0.998839	-1.092924
sleep apnea	1	0.8869	0.0	0.998839	0.886941
influenza	1	-0.9024	0.0	0.998839	-0.902410

Average scores were used for ties.

Van der Waerden One-Way Analysis		
Chi-Square	DF	Pr > ChiSq
733.8911	16	<.0001

Figure 4.18: Van der Waerden scores for LOSinhours by diagnosis

- **Van der Waerden Scores (Normal)**- This test examines the hypothesis that the reported LOSinHours scores originate from populations that are normally distributed across a range of diagnoses. Significant deviations from normality are shown by the p-value of 0.0001 and the Chi-square value of 733.8911 in this case.



Effects of diagnosis on LOSinhours

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable LOSinHours
Classified by Variable Diagnosis

Diagnosis	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Pneumonia	2799	11497408.0	9480213.0	79239.2200	4107.68417
Asthma	3373	9395542.0	11424351.0	80458.5943	2785.51497
Hemoptysis	31	113652.5	104997.0	10861.7557	3666.20968
Influenza	403	1464858.0	1364961.0	38066.8596	3634.88337
Sleep apnea	5	7748.5	16935.0	4370.5892	1549.70000
Acute bronchitis	19	67546.5	64353.0	8511.0304	3555.07895
pneumonia	6	22372.0	20322.0	4787.3869	3728.66667
Pleurisy	108	280440.5	365796.0	20157.5045	2596.67130
pleurisy	5	5628.0	16935.0	4370.5892	1125.60000
Emphysema of lung	8	25364.0	27096.0	5527.1812	3170.50000
Acute Bronchitis	3	8591.5	10161.0	3385.9440	2863.83333
hemoptysis	1	6664.0	3387.0	1955.1644	6664.00000
asthma	8	31345.0	27096.0	5527.1812	3918.12500
Acute exacerbation of chronic asthmatic bronchitis	1	5315.0	3387.0	1955.1644	5315.00000
PLEURISY	1	929.5	3387.0	1955.1644	929.50000
sleep apnea	1	5503.5	3387.0	1955.1644	5503.50000
influenza	1	1242.5	3387.0	1955.1644	1242.50000

Average scores were used for ties.

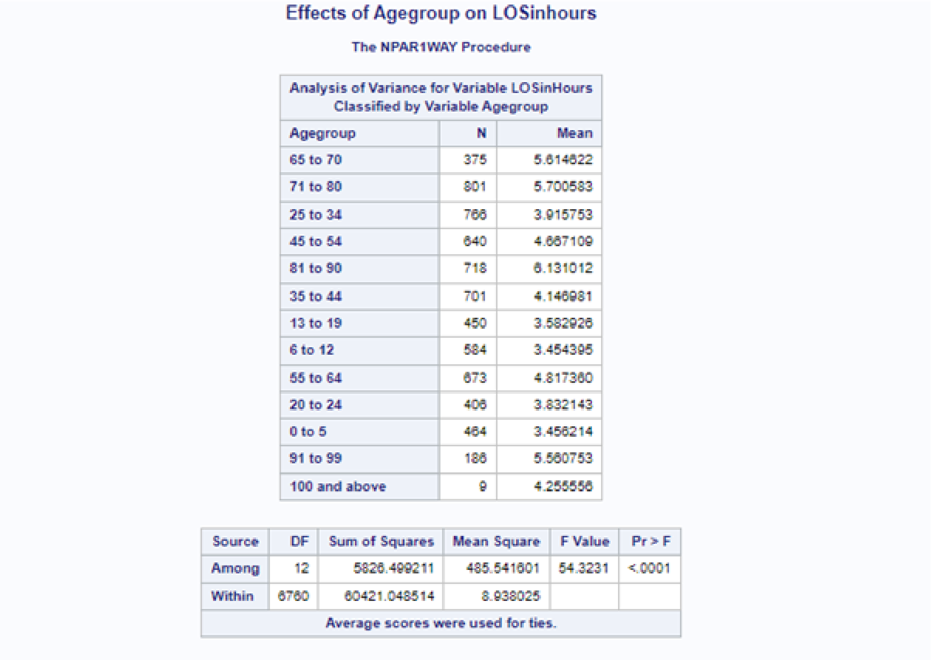
Kruskal-Wallis Test		
Chi-Square	DF	Pr > ChiSq
744.2417	16	<.0001

Figure 4.19: Wilcoxon scores(Rank Sums) for LOSinhours by diagnosis

- **Wilcoxon Scores (Rank Sums)**- The non-parametric alternative to the two-sample t-test is the Wilcoxon Rank Sum (which is mathematically identical to the Mann-Whitney U test). If just ranks (or ordinal data) are given, this test can still be run.

It contrasts the alternative hypothesis, that the two distributions differ solely with regard to the median, with the null hypothesis, that the two distributions are identical. For the purposes of this test, it is assumed that each sample contains independently dispersed observations that are similar in terms of both shape and spread. As long as the data are drawn at random from the same underlying population, it doesn't matter what distribution they come from(Pappas and DePuy, 2004). This test determines if there is a statistically significant difference between the rankings of the LOSinHours values for various diagnoses. The p-value of 0.0001 and the Chi-square value of 744.2417 both point to substantial variations in the LOSinHours rankings among diagnosis.

Effect of Age group on Length of Stay(LOS)



Effects of Agegroup on LOSinhours

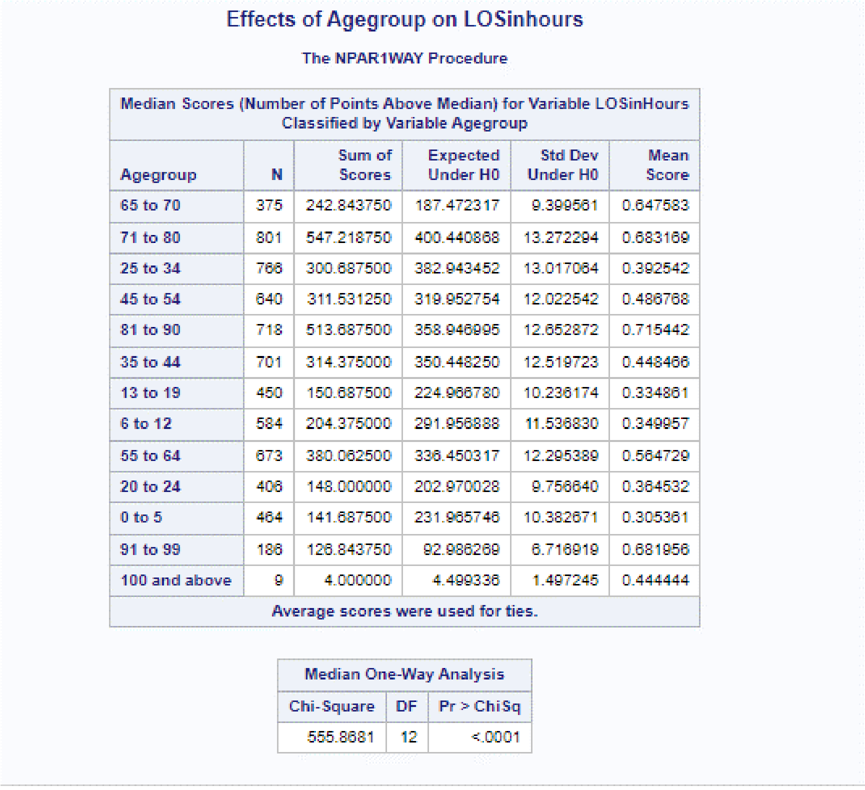
The NPAR1WAY Procedure

Analysis of Variance for Variable LOSinHours Classified by Variable Agegroup		
Agegroup	N	Mean
65 to 70	375	5.614622
71 to 80	801	5.700583
25 to 34	766	3.915753
45 to 54	640	4.667109
81 to 90	718	6.131012
35 to 44	701	4.146981
13 to 19	450	3.582926
6 to 12	584	3.454395
55 to 64	673	4.817360
20 to 24	406	3.832143
0 to 5	464	3.456214
91 to 99	186	5.560753
100 and above	9	4.255556

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Among	12	5826.499211	485.541601	54.3231	<.0001
Within	6760	60421.048514	8.938025		
Average scores were used for ties.					

Figure 4.20: Analysis of Variance for LOSinHours by Age group

- **Analysis of Variance** According to the findings of the analysis of variance (ANOVA), the age group has a significant impact on the length of stay in hours (LOSinhours) variable. The F-value of 54.3231 and the p-value of 0.0001 show that the mean LOSinhours differs significantly between age groups with ages 81-90 having the highest mean LOS of 6.13 hours.
- **Median Scores** The findings of the median scores demonstrate significant differences across age groups. The median one-way analysis's chi-square value is 555.8681, and its p-value is 0.0001, showing that there are significant variations in the median scores between age groups.
- **Van der Waerden Scores (Normal)** Significant differences across age groups are also shown by the Van der Waerden scores (normal) and Savage scores (exponential). The chi-square values for the Van der Waerden with p-values 0.0001 showing significant differences among age groups.
- **Wilcoxon Scores (Rank Sums)** The mean scores among age groups differ significantly, according to the Wilcoxon scores (rank sums). There are substantial disparities in the mean scores across age groups, as indicated by the chi-square value of 848.0952 and the p-value of 0.0001.



Effects of Agegroup on LOSinhours

The NPAR1WAY Procedure

Median Scores (Number of Points Above Median) for Variable LOSinHours
Classified by Variable Agegroup

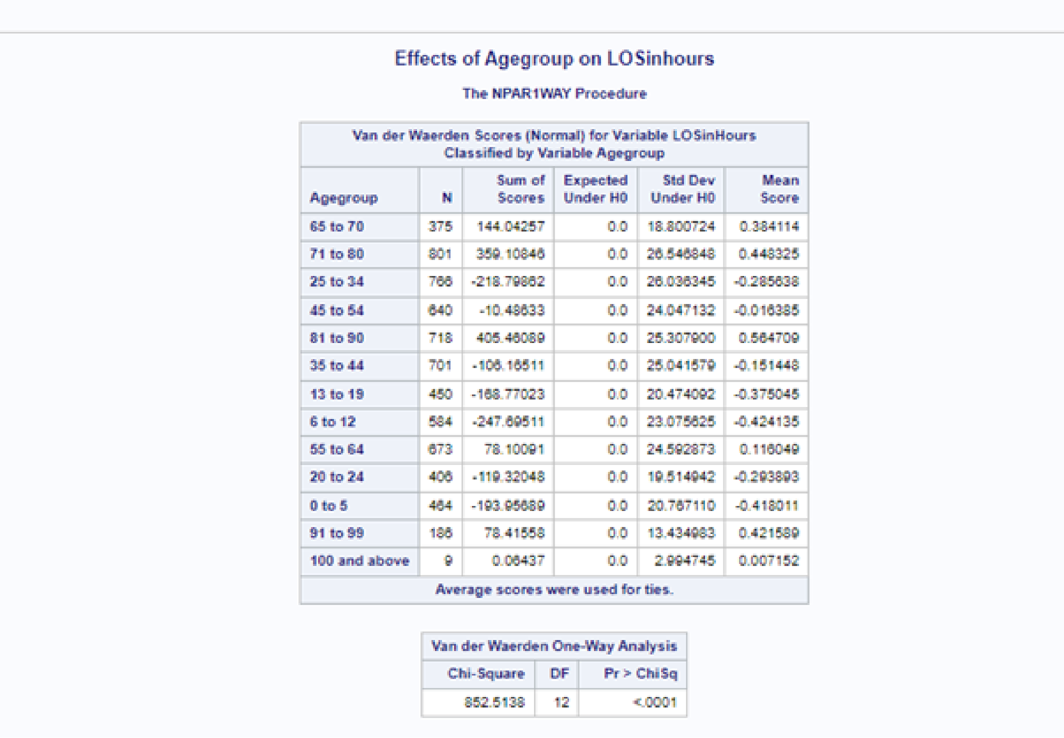
Agegroup	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
65 to 70	375	242.843750	187.472317	9.399561	0.647583
71 to 80	801	547.218750	400.440868	13.272294	0.683169
25 to 34	766	300.687500	382.943452	13.017064	0.392542
45 to 54	640	311.531250	319.952754	12.022542	0.486768
81 to 90	718	513.687500	358.946995	12.652872	0.715442
35 to 44	701	314.375000	350.448250	12.519723	0.448466
13 to 19	450	150.687500	224.966780	10.236174	0.334861
6 to 12	584	204.375000	291.956888	11.536830	0.349957
55 to 64	673	380.062500	336.450317	12.295389	0.564729
20 to 24	406	148.000000	202.970028	9.756640	0.364532
0 to 5	464	141.687500	231.965746	10.382671	0.305361
91 to 99	186	126.843750	92.986269	6.716919	0.681956
100 and above	9	4.000000	4.499336	1.497245	0.444444

Average scores were used for ties.

Median One-Way Analysis

Chi-Square	DF	Pr > ChiSq
555.8681	12	<.0001

Figure 4.21: Median scores for LOSinhours by Age group



Effects of Agegroup on LOSinhours

The NPAR1WAY Procedure

Van der Waerden Scores (Normal) for Variable LOSinHours
Classified by Variable Agegroup

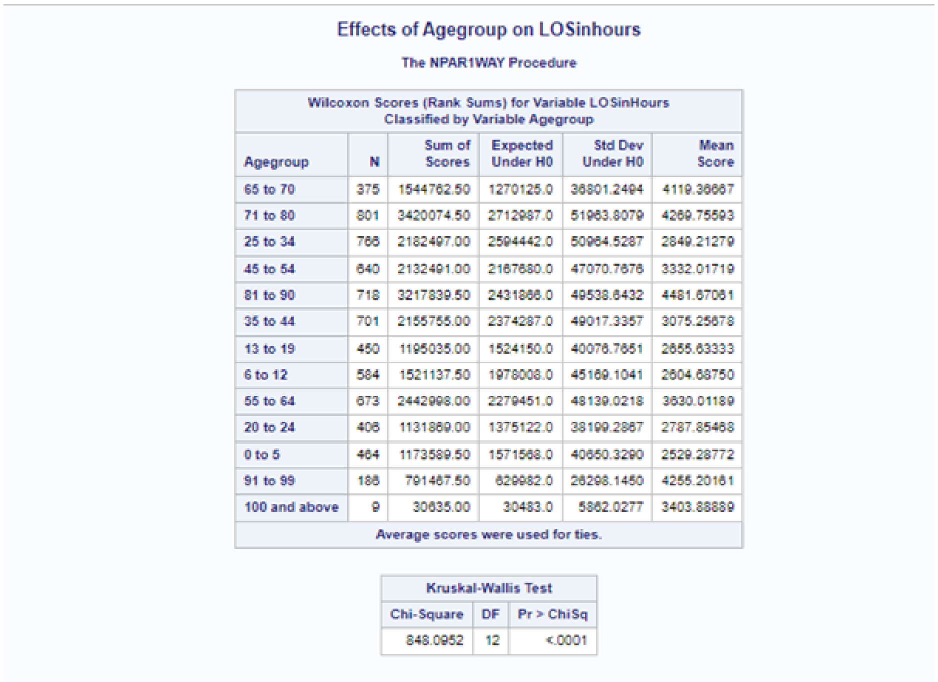
Agegroup	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
65 to 70	375	144.04257	0.0	18.800724	0.384114
71 to 80	801	359.10846	0.0	26.546848	0.448325
25 to 34	766	-218.79862	0.0	26.036345	-0.285638
45 to 54	640	-10.48633	0.0	24.047132	-0.016385
81 to 90	718	405.46089	0.0	25.307900	0.564709
35 to 44	701	-106.16511	0.0	25.041579	-0.151448
13 to 19	450	-168.77023	0.0	20.474092	-0.375045
6 to 12	584	-247.69511	0.0	23.075625	-0.424135
55 to 64	673	78.10091	0.0	24.592873	0.116049
20 to 24	406	-119.32048	0.0	19.514942	-0.293893
0 to 5	464	-193.95689	0.0	20.767110	-0.418011
91 to 99	186	78.41558	0.0	13.434963	0.421589
100 and above	9	0.06437	0.0	2.994745	0.007152

Average scores were used for ties.

Van der Waerden One-Way Analysis

Chi-Square	DF	Pr > ChiSq
852.5138	12	<.0001

Figure 4.22: Van der Waerden scores for LOSinhours by Age group



Effects of Agegroup on LOSinhours

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable LOSinHours
Classified by Variable Agegroup

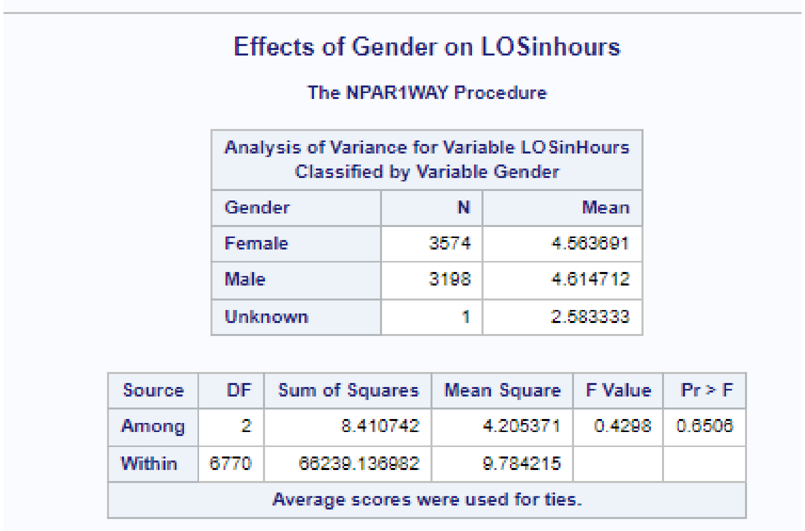
Agegroup	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
65 to 70	375	1544762.50	1270125.0	36801.2494	4119.36667
71 to 80	801	3420074.50	2712987.0	51963.8079	4269.75593
25 to 34	766	2182497.00	2594442.0	50964.5287	2849.21279
45 to 54	640	2132491.00	2167680.0	47070.7676	3332.01719
81 to 90	718	3217839.50	2431866.0	49538.6432	4481.67061
35 to 44	701	2155755.00	2374287.0	49017.3357	3075.25678
13 to 19	450	1195035.00	1524150.0	40076.7651	2655.63333
6 to 12	584	1521137.50	1978008.0	45169.1041	2604.68750
55 to 64	673	2442998.00	2279451.0	48139.0218	3630.01189
20 to 24	406	1131869.00	1375122.0	38199.2867	2787.85468
0 to 5	464	1173589.50	1571568.0	40650.3290	2529.28772
91 to 99	186	791467.50	629982.0	26298.1450	4255.20161
100 and above	9	30635.00	30483.0	5862.0277	3403.88889

Average scores were used for ties.

Kruskal-Wallis Test

Chi-Square	DF	Pr > ChiSq
848.0952	12	<.0001

Figure 4.23: Wilcoxon scores(Rank Sums) for LOSinhours by Age group



Effects of Gender on LOSinhours

The NPAR1WAY Procedure

Analysis of Variance for Variable LOSinHours
Classified by Variable Gender

Gender	N	Mean
Female	3574	4.563691
Male	3198	4.614712
Unknown	1	2.583333

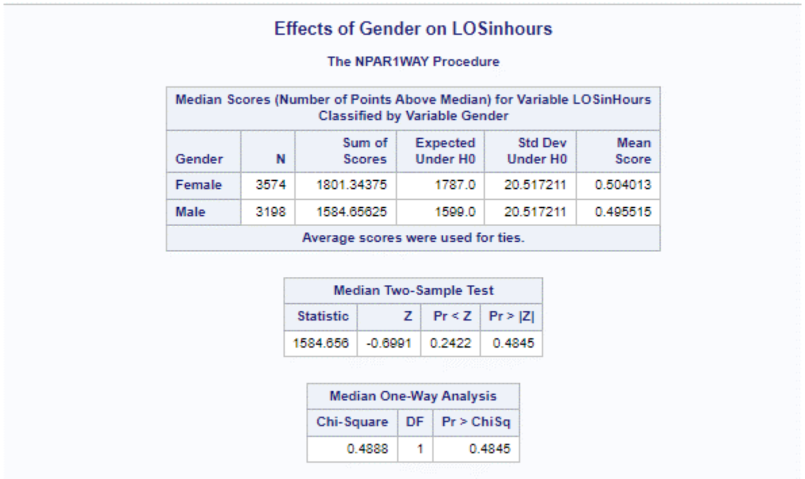
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Among	2	8.410742	4.205371	0.4298	0.6506
Within	6770	66239.136982	9.784215		

Average scores were used for ties.

Figure 4.24: Analysis of Variance for LOSinHours by gender

Effect of gender on Length of Stay(LOS)

- **Analysis of Variance** According to the ANOVA results, there is no evidence that gender has a substantial impact on LOSinhours. which have an F-value of 0.4490 and a p-value of 0.5028, there is no significant difference in the mean LOSinhours between males and females.
- **Median Scores**:The median scores for the variable "LOSinHours" were compared between genders in the analysis. According to the findings, males (N=3198) had median scores of 1584.66 and females (N=3574) had median scores of 1801.34. The chi-square test revealed that there was no gender difference in the median scores (p=0.4745).
- **Van der Waerden Scores (Normal)**: Similar to the median scores, Van der Waerden scores, and Savage scores, which had p-values of 0.4845, 0.4126, and 0.7131, respectively, do not demonstrate any significant differences between males and female.



Effects of Gender on LOSinhours

The NPAR1WAY Procedure

Median Scores (Number of Points Above Median) for Variable LOSinHours Classified by Variable Gender

Gender	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Female	3574	1801.34375	1787.0	20.517211	0.504013
Male	3198	1584.65625	1599.0	20.517211	0.495515

Average scores were used for ties.

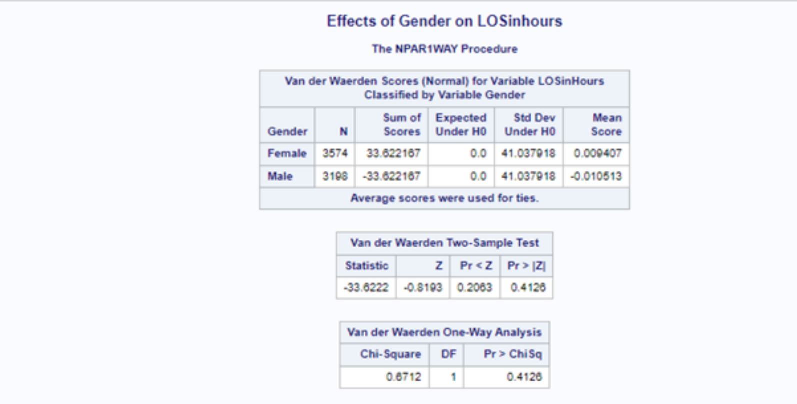
Median Two-Sample Test

Statistic	Z	Pr < Z	Pr > |Z|
1584.656	-0.6991	0.2422	0.4845

Median One-Way Analysis

Chi-Square	DF	Pr > ChiSq
0.4888	1	0.4845

Figure 4.25: Median scores for LOSinhours by gender



Effects of Gender on LOSinhours

The NPAR1WAY Procedure

Van der Waerden Scores (Normal) for Variable LOSinHours Classified by Variable Gender

Gender	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Female	3574	33.622167	0.0	41.037918	0.009407
Male	3198	-33.622167	0.0	41.037918	-0.010513

Average scores were used for ties.

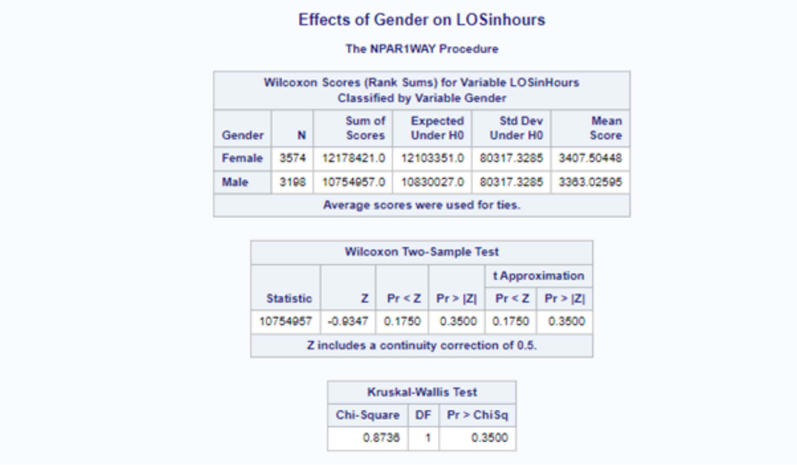
Van der Waerden Two-Sample Test

Statistic	Z	Pr < Z	Pr > |Z|
-33.6222	-0.8193	0.2063	0.4126

Van der Waerden One-Way Analysis

Chi-Square	DF	Pr > ChiSq
0.6712	1	0.4126

Figure 4.26: Van der Waerden scores for LOSinhours by gender



Effects of Gender on LOSinhours

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable LOSinHours Classified by Variable Gender

Gender	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Female	3574	12178421.0	12103351.0	80317.3285	3407.50448
Male	3198	10754957.0	10830027.0	80317.3285	3363.02595

Average scores were used for ties.

Wilcoxon Two-Sample Test

				t Approximation	
Statistic	Z	Pr < Z	Pr > |Z|	Pr < Z	Pr > |Z|
10754957	-0.9347	0.1750	0.3500	0.1750	0.3500

Z includes a continuity correction of 0.5.

Kruskal-Wallis Test

Chi-Square	DF	Pr > ChiSq
0.8736	1	0.3500

Figure 4.27: Wilcoxon scores(Rank Sums) for LOSinhours by gender

- **Wilcoxon Scores (Rank Sums)**: With a Z-statistic of -0.9347 and a p-value of 0.1750, the Wilcoxon scores (rank sums) likewise show that there is no discernible difference in the distribution of LOSinhours between females and males

Based on the abnormal data distribution,the analysis used nonparametric tests to look at the impact of diagnosis, age group, and gender on LOS. These tests included Analysis of Variance, Median Scores, Van der Waerden Scores, and

Wilcoxon Scores. Interesting findings about these factors were presented by the results. The Analysis of Variance, Median scores, Van der Waerden scores, and Wilcoxon scores all showed that the diagnosis had a substantial influence on LOS. These tests revealed significant variability in LOS among diagnoses, pointing to the possibility that particular medical problems might result in either longer or shorter hospitalisations. Secondly, it was observed that age had a big impact on LOS. This conclusion was validated by the Analysis of Variance, the Median scores, the Van der Waerden scores, and the Wilcoxon scores, suggesting that various age groups may have variable lengths of stays in hospitals. However, the Analysis of Variance, Median scores, Van der Waerden scores, and Wilcoxon scores did not show that gender had a significant effect on LOS.

4.3 Machine learning models for predictive diagnosis

To improve the efficiency of patients' hospital visits, predictive analytics can be used to diagnose patients illness. This can be achieved by using machine learning models which can predict patient illness based on their demographics and presenting complaints. In this section, Four Machine learning models, K-nearest Neighbours (KNN), random forest, logistic regression, and decision trees are created with python using jupyter notebook.

4.3.1 Data Preprocessing and exploration for model development

The data is first explored and preprocessed in such a way that it provides good accuracy results in the models.

Data exploration

- **Importing the libraries into Python environment.**

  Different Libraries are imported for easy data exploration, data preprocessing and predictive modelling.

- **Reading in the data**

  Using pd.read_excel, the excel file is imported into the python environment with the Jupyter notebook. The dataframe is named resp_patients_data for easy identification.

  Calling up the dataframe shows a total of 6774 rows and 14 columns.

- **Retrieving info about the data**

  To get information about the data, we use the resp_patients_data.info() to retrieve the details. Datatypes include datetime, float and Object.

  Calling up the dataframe shows a total of 6774 rows and 14 columns.

- **Retrieving Column details**

  This is achieved by using the resp_patient_data.columns function.

  Columns in the dataset as can be seen are EncounterSK, MonthCommence, DaysofWeek, TimeofArrival, TimeofDeparture, LOS, PresentingComplaint, DiagnosisDescription, ArrivalMode, TriagePriority, AgeOnArrival, Gender, Facility, Disposal.

- **Checking details of rows and columns**

  Using resp_patients_data.shape we confirm the details of number of rows and columns.

  The data has a total of 6774 rows and 14 columns.

- **Confirming missing values**

  The number of missing values are checked using the resp_patient_data.isnull().sum() function. The 'Timeofdeparture','TotalHoursspent',LOS and 'triagepriority' columns all have one missing values each.

Data Preprocessing

- **Dropping missing values**

Using dropna() function we eliminate the missing values in the dataset. Eliminating the missing data will help us build a better machine learning model. After the data is dropped, the number of rows left is 6772 rows and 14 columns.

- **Modify Column names**

  For proper labelling the column name 'Ageonarrival' is modified to AgeGroup while the column name'DiagnosisDescription' is renamed to Diagnosis.

- **Dropping columns not required in the model**

  To improve the data for the model, we drop columns which are not being used as features for the prediction. The Columns not needed are EncounterSK, ArrivalMode, TriagePriority, Facility, Disposal, Monthcommence, DaysofWeek, TimeofArrival and TimeofDeparture.

  After the columns are dropped, the data has 6772 rows and 5 columns left.

- **Filtering the target column**

  Because Asthma and Pneumonia have the highest cases we will be using the two illnesses to build the machine learning model, all other illness are filtered away.

  After filtering the data and eliminating all other illnesses, there are 6186 rows and 5 columns

- **Filtering the presenting complaints column**

  The top 5 presenting complaints are chosen to be used to train the model other complaints are filtered off.

  After filtering, the number of rows left are 4,584 and 5 columns.

- **Improving data quality of the target column**

  To ensure consistency with the Diagnosis, we remove the rows in which the pneumonia and asthma are spelled with the first letter as lower case since all the others are spelled with the first letter's as uppercase.

  The data becomes 4,573 rows and 5 columns showing that the deleted data is not a very significant number.

- **Converting the features to numerical data**

The feature columns, 'AgeGroup', 'Gender' and 'Presentingcomplaints' are all converted numerical data. This is achieved by first creating an instance of Label encoder and then doing a transformation using the fit_transform feature

4.3.2 Model development

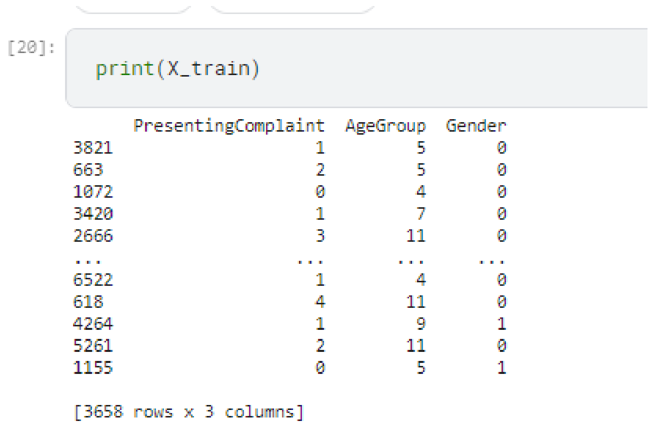
Splitting the data into train and test set

The data is split into different sets for training and testing and this is described as :

- **X_train**- Features for training
- **X_test**- Features for testing/evaluation
- **y_train**-Target variable for training
- **y_test**- Target variable for testing/evaluation

We now confirm the details of the data for both the training and testing using the print function.

- The training data has with the features 3,658 rows and 3 columns
- The test data with the features has 915 rows and 3 columns
- The train data with the target has 3,658 rows
- The test data with the target has 915 rows

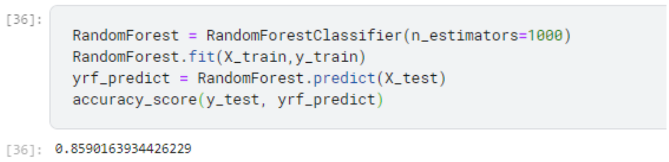


```
[20]:  print(X_train)

       PresentingComplaint  AgeGroup  Gender
3821                     1         5       0
663                      2         5       0
1072                     0         4       0
3420                     1         7       0
2666                     3        11       0
...                    ...       ...     ...
6522                     1         4       0
618                      4        11       0
4264                     1         9       1
5261                     2        11       0
1155                     0         5       1

[3658 rows x 3 columns]
```

Figure 4.28: Python code to retrieve details of the feature train data

Random forest model

When tree predictors are combined, the result is a forest known as a "random forest," where each tree is reliant on values from a random vector that was picked randomly and uniformly over the whole forest(Breiman, 2001).



```
[36]:  RandomForest = RandomForestClassifier(n_estimators=1000)
       RandomForest.fit(X_train,y_train)
       yrf_predict = RandomForest.predict(X_test)
       accuracy_score(y_test, yrf_predict)

[36]:  0.8590163934426229
```

Figure 4.29: Random forest accuracy

Accuracy is the proportion of properly categorised data instances over all data instances (Shung, 2018)

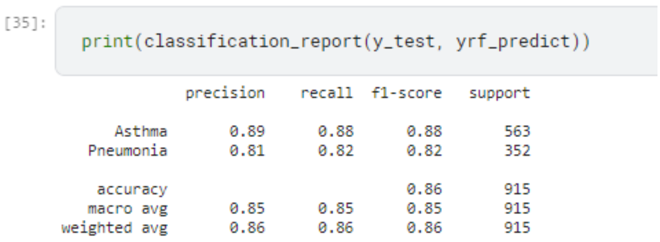
It is calculated as :

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

Figure 4.30: calculation of accuracy (Shung, 2018)

where TN= True Negatives TP= True Positives FP= False Positives FN= False Negatives

- **Random forest classification report**



```
[35]:  print(classification_report(y_test, yrf_predict))

              precision    recall  f1-score   support

      Asthma       0.89      0.88      0.88       563
   Pneumonia       0.81      0.82      0.82       352

    accuracy                           0.86       915
   macro avg       0.85      0.85      0.85       915
weighted avg       0.86      0.86      0.86       915
```

Figure 4.31: Python code and result for Random forest model classification report

Evaluating the performance of the model

Precision: The classifier's precision indicates how effectively it accurately recognises positive examples for each class. For instance, 89% of the instances of asthma predicted were truly cases of asthma, and 81% of the cases of pneumonia projected were indeed cases of pneumonia.

Recall: The classifier's capacity to identify all of the positive examples is measured by recall. In the cases of Asthma and Pneumonia, respectively, 88% and 82% of the actual cases of each disease were accurately classified.

F1-score: Precision and recall are combined into a single statistic known as the F1-score, which measures overall performance. The F1-score for asthma is 88%, whereas the F1-score for pneumonia is 82%.

Support: Support describes the quantity of samples required to assess each class's model. There are 563 samples for this situation, there are 563 samples for asthma and 352 samples for pneumonia

- **Random forest confusion matrix**

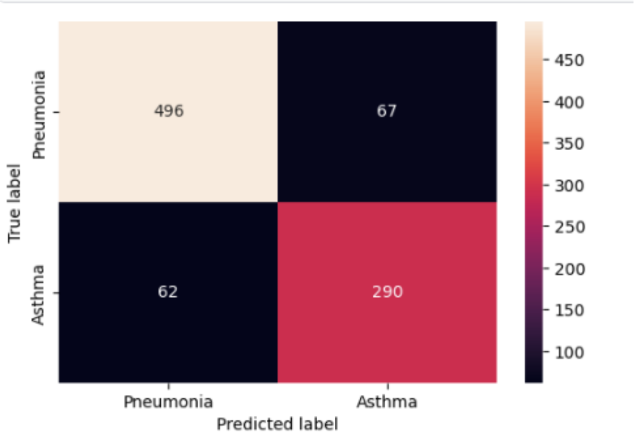
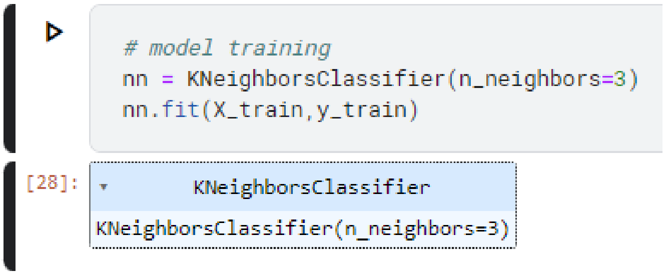


Figure 4.32: Random forest confusion matrix

The performance of the classification model for the classes Asthma and Pneumonia is usefully shown by the confusion matrix. With 496 cases of asthma and 290 cases of pneumonia properly diagnosed, the model obtained a good number of true positives. False positives also occurred; 67 occurrences that should have been classed as Pneumonia were instead labelled as Asthma. False negatives also occurred; 62 cases were wrongly labelled as non-pneumonia when they really involved cases of pneumonia. The model produced a high percentage of accurate positive predictions, as evidenced by the accuracy scores of 0.88 for asthma and 0.82 for pneumonia. Recall scores of 0.89 for Asthma and 0.81 for Pneumonia demonstrate the model's capability to properly identify each class. Overall, the research shows that the model has a fair mix between recall and precision, producing accurate predictions for the categories of asthma and pneumonia.

K-Nearest Neighbour(KNN)

The K-nearest neighbour (KNN) method is a supervised machine learning strategy that has been extensively employed in illness prediction. The supervised algorithm KNN makes predictions about how unlabelled data will be categorised by considering the traits and labels of the training data. (Uddin *et al*, 2022).



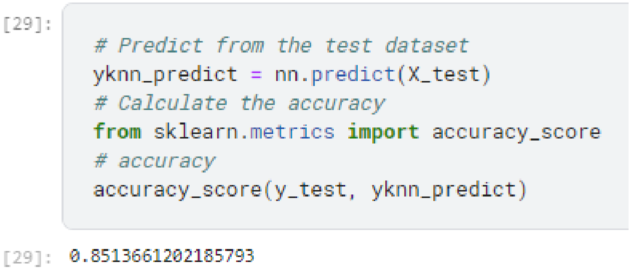
```
# model training
nn = KNeighborsClassifier(n_neighbors=3)
nn.fit(X_train,y_train)
```

```
[28]: KNeighborsClassifier
KNeighborsClassifier(n_neighbors=3)
```

Figure 4.33: Python code and result for selecting a k-value to train the data with

If n_neighbors=3 and nn = KNeighborsClassifier: By setting the argument n_neighbors to 3, this line creates a KNeighborsClassifier object. The number of neighbours utilised in the KNN algorithm is specified by the n_neighbors option. When making predictions in this scenario, the KNN algorithm will take into account the three closest neighbours. X_train and y_train in nn.fit The training data X_train (features) and y_train (target variable) are fitted by this line to the KNN model. By identifying patterns in the training data and creating a decision boundary to generate predictions, the fit technique trains the model.

This code estimates the accuracy of the predictions and uses a trained K-Nearest Neighbours (KNN) model to forecast the values of the target variable. It predicts the target values for the test dataset using the predict function, and then computes the accuracy score by contrasting the anticipated and actual values. Higher scores indicate greater performance. The accuracy score reflects how effectively the model predicts the right values.



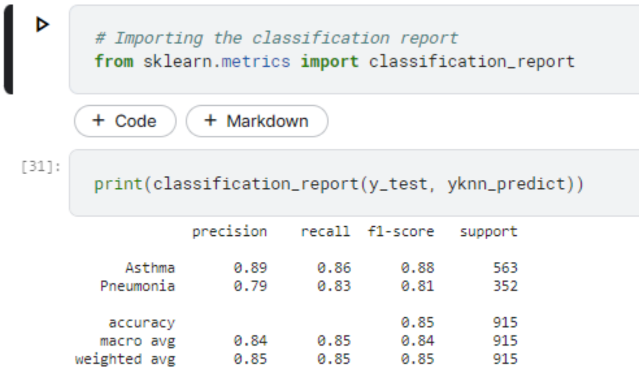
```
[29]:
# Predict from the test dataset
yknn_predict = nn.predict(X_test)
# Calculate the accuracy
from sklearn.metrics import accuracy_score
# accuracy
accuracy_score(y_test, yknn_predict)

[29]: 0.8513661202185793
```

Figure 4.34: Python code and result for KNN model accuracy

The KNN model has an accuracy of 0.851 approximately 85% accuracy level.

- **KNN Classification report**

  The classification report library is imported and the classification report is generated.



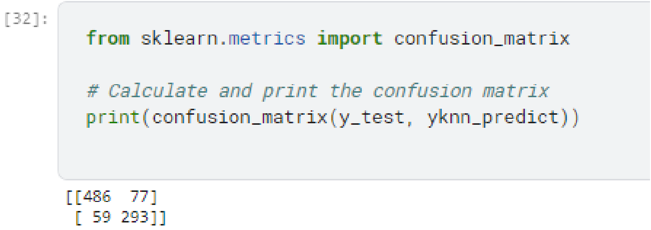
```
# Importing the classification report
from sklearn.metrics import classification_report
```

```
[31]:
print(classification_report(y_test, yknn_predict))

              precision    recall  f1-score   support

      Asthma       0.89      0.86      0.88       563
   Pneumonia       0.79      0.83      0.81       352

    accuracy                           0.85       915
   macro avg       0.84      0.85      0.84       915
weighted avg       0.85      0.85      0.85       915
```

Figure 4.35: Python code and result for KNN classification report

Precision indicates the proportion of cases when the projected favourable outcome really occurred. Regarding this: The accuracy is 0.89 for "Asthma". This indicates that 89% of the cases classified as "Asthma" were truly genuine positive cases of the disease. The accuracy is 0.79 for "Pneumonia". Accordingly, 79% of all cases that were correctly identified as "Pneumonia" predictions really were. Recall: Recall gauges how many occurrences of true positives were really accurately detected. Regarding this: The recall is 0.86 for "Asthma". In other words, 86% of all genuine positive "Asthma" occurrences were accurately detected by the model. The recall is 0.83 for "Pneumonia". In other words, the model successfully predicted all genuine positive "Pneumonia" cases.

- **KNN confusion matrix**



```
[32]:
from sklearn.metrics import confusion_matrix

# Calculate and print the confusion matrix
print(confusion_matrix(y_test, yknn_predict))

[[486  77]
 [ 59 293]]
```

Figure 4.36: Python code and results for KNN confusion matrix

The model accurately identified 486 of the 563 cases of asthma (referred to as "True Positives") out of the total. But in 77 instances, the model misdiagnosed them as cases of pneumonia when they actually belonged to the

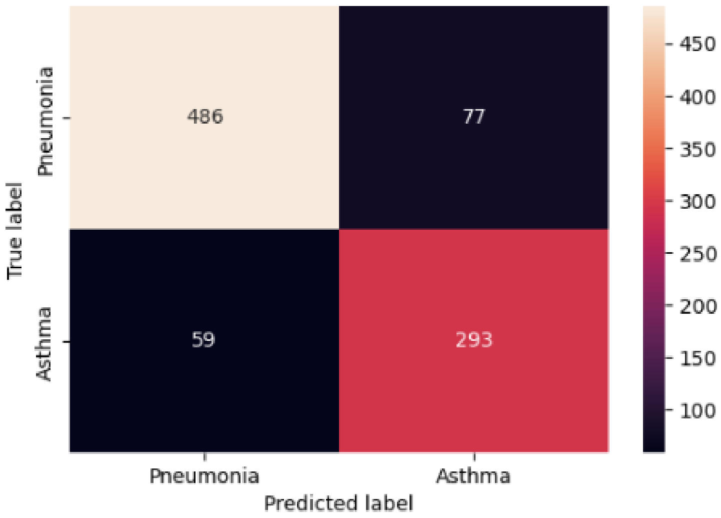
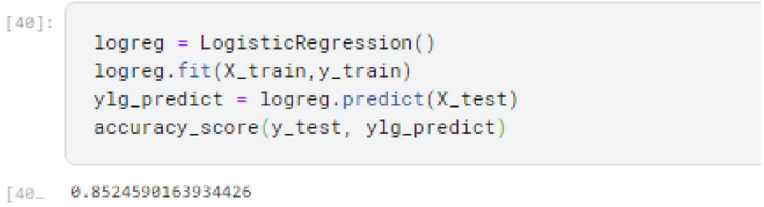


Figure 4.37: confusion matrix results

asthma class ("False Negatives"). The "Pneumonia" class's real occurrences are shown in the second row. The model mistakenly identified 59 cases of Pneumonia out of a total of 352 instances as Asthma when they actually belonged to the Pneumonia class ("False Positives"). 293 of the occurrences were accurately identified by the model as Pneumonia ("True Negatives").

Logistic regression model

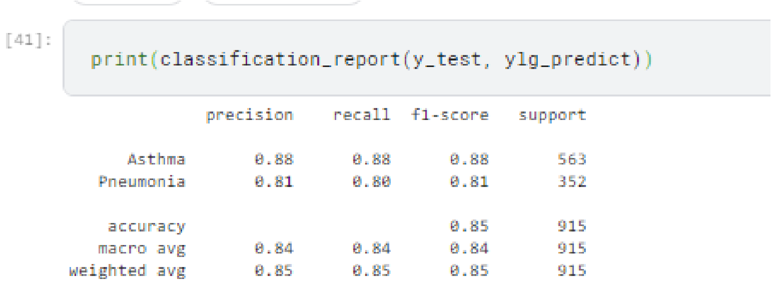
Predictions in logistic regression are made by comparing the likelihood that an observation belongs to a certain class. Through linear regression, the logarithm of these probability is computed (Blotwijk *et al*, 2023)



```
[40]:
logreg = LogisticRegression()
logreg.fit(X_train,y_train)
ylg_predict = logreg.predict(X_test)
accuracy_score(y_test, ylg_predict)

[40_  0.8524590163934426
```

Figure 4.38: Logistic regression code and accuracy result

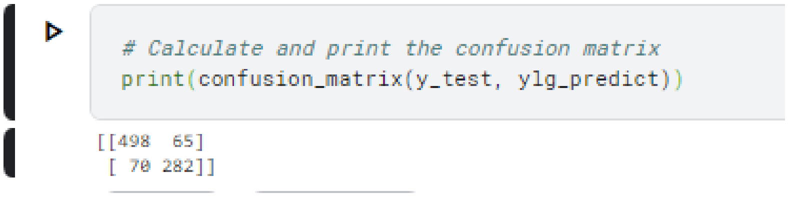
- **Classification Report for Logistic Regression** The precision for the "Asthma" class is 0.88, meaning that 88% of the cases that are predicted to be "Asthma" are indeed "Asthma." The accuracy is 0.81 for the "Pneumonia" class as well. With a recall of 0.88 for the "Asthma" class, 88% of the real "Asthma" occurrences were properly recognised. The recall is 0.80 for the "Pneumonia" class as well. The F1-score for the "Asthma" class is 0.88, showing a fair balance between recall and accuracy. The F1-score is 0.81 for the "Pneumonia" class as well. Support shows how many real instances of each class there are. There are 563 occurrences of the "Asthma" class and 352 instances of the "Pneumonia" class. Across all classes, accuracy measures how accurately predictions were made overall. Since the accuracy in this instance is 0.85, 85% of the occurrences have been accurately categorised. By treating each class equally, macro average determines the average metrics (precision, recall, and f1-score) over all classes. The accuracy, recall, and f1-score are all 0.84 on average on a macro level in this situation. With each class being given a different weight based on its support, weighted average determines the average metrics (precision, recall, and f1-score) over all classes. The weighted average accuracy, recall, and f1-score in this instance are all 0.85.



```
[41]:
print(classification_report(y_test, ylg_predict))

              precision    recall  f1-score   support

      Asthma       0.88      0.88      0.88       563
   Pneumonia       0.81      0.80      0.81       352

    accuracy                           0.85       915
   macro avg       0.84      0.84      0.84       915
weighted avg       0.85      0.85      0.85       915
```

Figure 4.39: Logistic regression classification report

- **Confusion matrix for Logistic Regression**



```
# Calculate and print the confusion matrix
print(confusion_matrix(y_test, ylg_predict))

[[498  65]
 [ 70 282]]
```

Figure 4.40: Logistic regression confusion matrix code and results

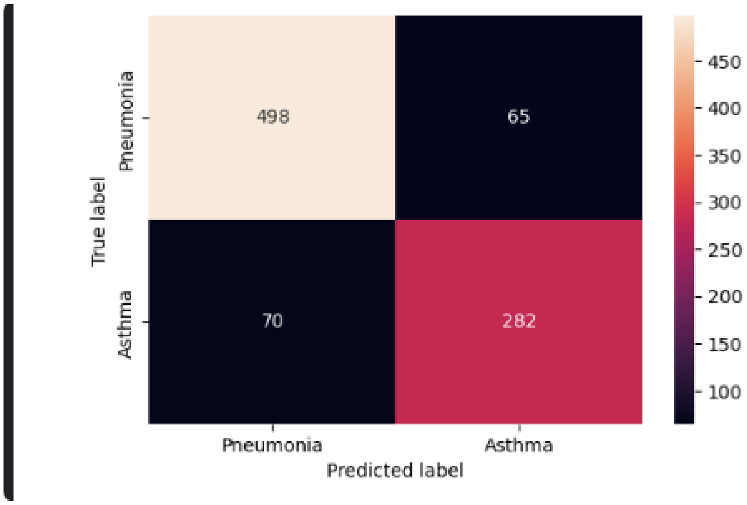
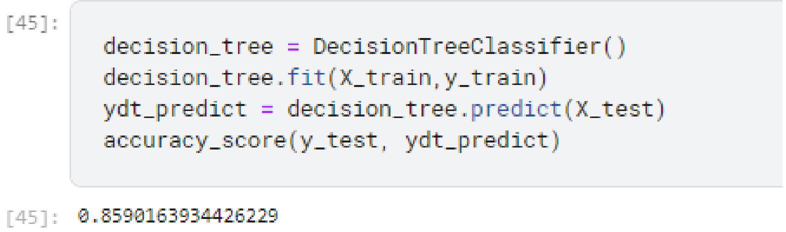


Figure 4.41: Logistic regression confusion matrix diagram

In the "Asthma" category True Positives: The model accurately identified 498 cases as "Asthma" (n = 498). False Positives (65): The model misdiagnosed 65 cases as "asthma" when they weren't. False Negatives (70): The model predicted 70 occurrences as not having "asthma" when they really did have "asthma." True Negatives (282): The model accurately identified 282 occurrences as not being "Asthma" in 282 instances. Regarding "Pneumonia":True Positives (282): In 282 occasions, the model accurately identified the condition as "Pneumonia".False Positives (70): The model misidentified 70 cases as "Pneumonia" when they weren't. False Negatives (65): The model mistakenly identified 65 occurrences as "Pneumonia" when they were "Pneumonia." True Negatives (498): The model accurately identified 498 occurrences as not being "Pneumonia" in 498 of the cases.

Decision tree model

This kind of machine learning technique divides the data into smaller groups recursively depending on the value of one or a few selected attributes. A specific measure, such maximising information gain or minimising entropy, guides the selection of splits. The result is a structure that resembles a tree, where each node is connected to the primary class of the data it contains and judgements are made at each branch (Blotwijk et al, 2023)

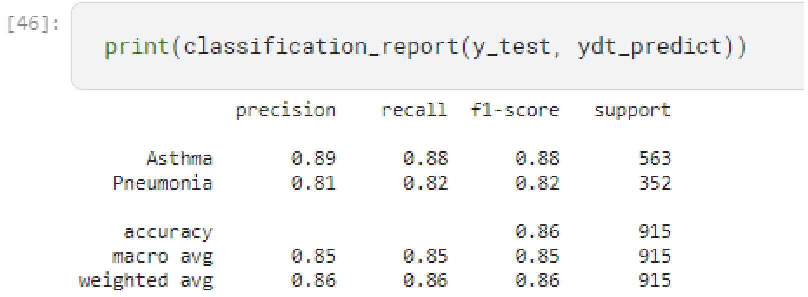


```
[45]:  decision_tree = DecisionTreeClassifier()
       decision_tree.fit(X_train,y_train)
       ydt_predict = decision_tree.predict(X_test)
       accuracy_score(y_test, ydt_predict)

[45]: 0.8590163934426229
```

Figure 4.42: Python code for decision tree model accuracy

- **Classification report for decision tree**

  The accuracy for the "Asthma" class is 0.89, which indicates that 89% of the cases that were predicted to be "Asthma" are indeed "Asthma." The accuracy is 0.81 for the "Pneumonia" class as well. With a recall of 0.88 for the "Asthma" class, 88% of the real "Asthma" occurrences were properly recognised. With a recall of 0.82 for the "Pneumonia" class, 82% of the real "Pneumonia" cases are properly recognised. The F1-score for the "Asthma" class is 0.88. The F1-score is 0.82 for the "Pneumonia" class as well. Support: There are 352 cases of "Pneumonia" and 563 cases of "Asthma" in this instance. Reliability: The model's overall reliability is 0.86, which means that 86% of the occurrences are properly categorised. The macro average for the F1-score, recall,

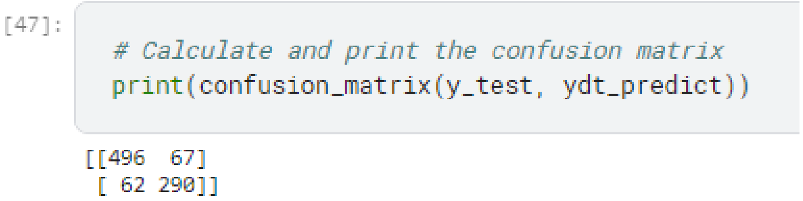


```
[46]: print(classification_report(y_test, ydt_predict))

              precision    recall  f1-score   support

      Asthma       0.89      0.88      0.88       563
   Pneumonia       0.81      0.82      0.82       352

    accuracy                           0.86       915
   macro avg       0.85      0.85      0.85       915
weighted avg       0.86      0.86      0.86       915
```

Figure 4.43: Python code and results for decision tree classification report

and accuracy in this instance is 0.85. The weighted average accuracy, recall, and F1-score in this instance are all 0.86.

- **Confusion matrix for decision tree**



```
[47]: # Calculate and print the confusion matrix
      print(confusion_matrix(y_test, ydt_predict))

[[496  67]
 [ 62 290]]
```

Figure 4.44: Python code and results for decision tree confusion matrix

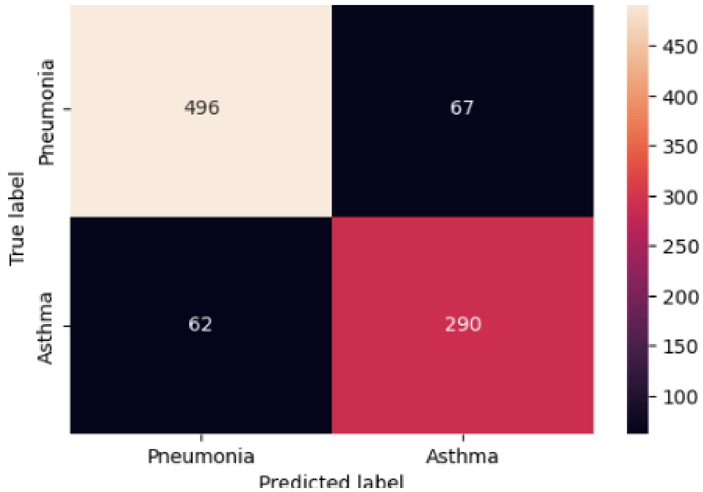
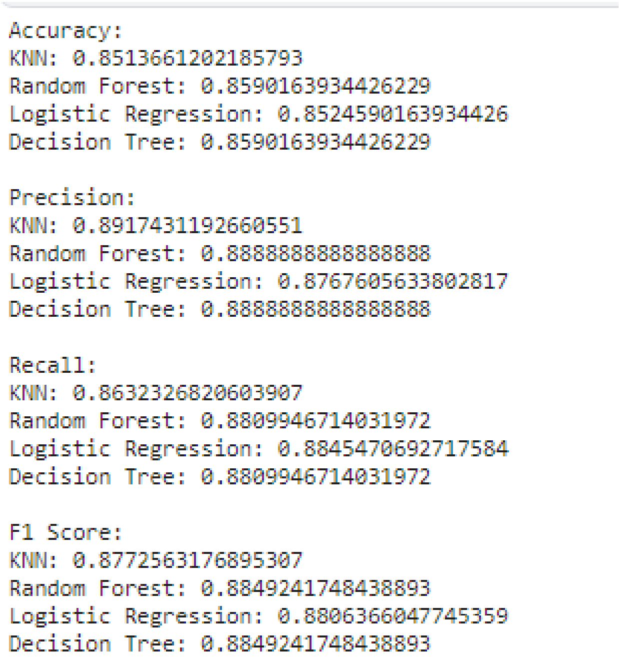


Figure 4.45: Decision tree confusion matrix diagram

For the class "Asthma": True Positives (496): The model accurately identified 496 occurrences as "Asthma" in these circumstances. False Positives (67): 67 events were falsely identified as "asthma" by the model despite not really being "asthma." False Negatives (62): The model misidentified 62 occurrences as having "Asthma" when in fact they did. True Negatives (290): 290 cases where the model properly identified them as not being "Asthma" occurred. In relation to "Pneumonia": True Positives (290): 290 cases when the model accurately identified "Pneumonia" as the diagnosis. False Positives (62): The model was implemented improperly

4.3.3 Model comparism and evaluation

- **Prediction for Asthma:**



```
Accuracy:
KNN: 0.8513661202185793
Random Forest: 0.8590163934426229
Logistic Regression: 0.8524590163934426
Decision Tree: 0.8590163934426229

Precision:
KNN: 0.8917431192660551
Random Forest: 0.8888888888888888
Logistic Regression: 0.8767605633802817
Decision Tree: 0.8888888888888888

Recall:
KNN: 0.8632326820603907
Random Forest: 0.8809946714031972
Logistic Regression: 0.8845470692717584
Decision Tree: 0.8809946714031972

F1 Score:
KNN: 0.8772563176895307
Random Forest: 0.8849241748438893
Logistic Regression: 0.8806366047745359
Decision Tree: 0.8849241748438893
```

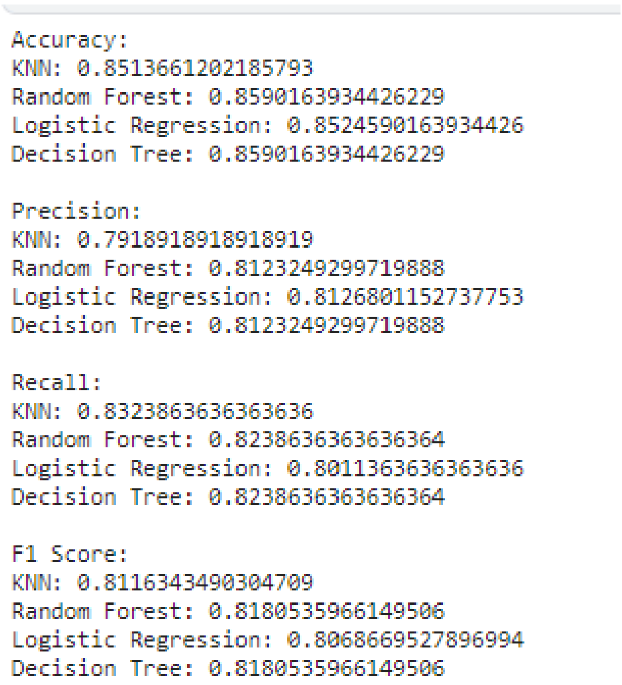
Figure 4.46: Model evaluation for the prediction of asthma

Accuracy: The best accuracy ratings were obtained by the Decision Tree and Random Forest models (85.9%), suggesting strong overall prediction ability.

Precision: The highest precision scores were obtained by the KNN and Random Forest models (88.9% to 89.2%), suggesting a low percentage of false positives.

Recall: The Logistic Regression model obtained the greatest recall score (88.5%), successfully catching a significant portion of positive instances of asthma. F1 Score: With an F1 score of 88.5%, the Random Forest model had the best precision and recall ratios.

- **Prediction for Pneumonia:**



```
Accuracy:
KNN: 0.8513661202185793
Random Forest: 0.8590163934426229
Logistic Regression: 0.8524590163934426
Decision Tree: 0.8590163934426229

Precision:
KNN: 0.7918918918918919
Random Forest: 0.8123249299719888
Logistic Regression: 0.8126801152737753
Decision Tree: 0.8123249299719888

Recall:
KNN: 0.8323863636363636
Random Forest: 0.8238636363636364
Logistic Regression: 0.8011363636363636
Decision Tree: 0.8238636363636364

F1 Score:
KNN: 0.8116343490304709
Random Forest: 0.8180535966149506
Logistic Regression: 0.8068669527896994
Decision Tree: 0.8180535966149506
```

Figure 4.47: Model Comparism results

Accuracy: The Decision Tree and Random Forest models scored the highest in terms of accuracy (85.9%), indicating great general predictive power.

Precision: The KNN model, with the highest precision score (79.2%), showed a minimal amount of false positives.

Recall: The KNN model was the most successful in identifying cases with confirmed pneumononia, with a recall score of 83.2%. The models with the highest F1 scores, Random Forest and Decision Tree (81.8%), displayed a good balance between accuracy and recall.

Model selection:

In conclusion, both the Random Forest and Decision Tree models consistently outperformed other assessment criteria for both the prediction of Pneumonia and Asthma. They received high ratings for precision, recall, accuracy, and F1, indicating that they were successful in anticipating both situations. The performance of the KNN model was somewhat worse but still acceptable, but the recall and F1 scores for the Asthma prediction in the Logistic Regression model were marginally worse.Overall, according to the assessment findings, it appears that the Random Forest model received great scores for accuracy, precision, recall, and F1 score For both scenarios . As a result, the Random Forest model may be regarded as the best model for predicting both Pneumonia and Asthma based on these performance parameters.

CHAPTER 5

CONCLUSIONS AND RECOMMENDATION

5.1 Conclusions

This chapter mainly discusses the results of the examination of the data on respiratory ailments as well as results from the model creation and evaluation. The findings of this study provide valuable insight on patients waiting times or lenth of stay and critically accesses important factors that contribute to how long hospital stays last for patients with respiratory ailments. In-depth investigation was carried out and machine learning models to enhance prediction abilities was created. Resources used include Microsoft power BI, SAS programming, and Python.

5.1.1 Demographic Analysis

Initial analysis of the data showed that hospital visits for respiratory illnesses were mostly common with patients between the ages of 71 to 80, followed closely followed by ages 25 to 34 and 81 to 90. What this means is that elderly people between the ages of 71 to 90 were mostly affected. This emphasises the requirement for specialised care and resources to meet the particular demands of this age group. Analysis by gender shows that female(52.78%) patients were slightly more than male patients(47.22%) which is not a significant difference.

5.1.2 Time series analysis

An interesting discovery is the difference in patient visits by month and year. Review by month shows that in December, there was a higher number of respiratory patients that in any other month while a review of the number of respiratory patient visits by year revealed that from 2019 to 2022, there was a steady rise in the total number of patients. This trend could be attributed to seasonal variations, flu epidemics, or other environmental factors. This information may be used by healthcare organisations to better plan for times of high demand and distribute resources appropriately.

Intriguing trends in patient visits by weekday were also found through the investigation. The days with the most visits were Mondays, while Saturdays had the fewest visits. The availability of sufficient personnel and facilities on days with a larger intake of respiratory patients may be ensured with the use of this information, which can be helpful for hospital scheduling and resource allocation

5.1.3 Most frequent respiratory ailments

According to the data, pneumonia and asthma are the two most prevalent respiratory illnesses, with pneumonia accounting for 41.33% of all cases and asthma for 49.79% of them. At 5.95%, influenza cases also showed up in substantial numbers. This demonstrates the substantial burden these illnesses place on healthcare resources and the demand for efficient management and treatment plans for asthma and pneumonia patients. There were fewer cases of other respiratory illnesses such as acute bronchitis, lung emphysema, haemoptysis, pleurisy, and sleep apnea.

5.1.4 Diagnostic testing

To get insight into the diagnosis process, the research concentrated on medical examinations and testing, eliminating prescription orders. The top 10 procedures were an X-ray of the chest, an ECG, a complete blood count, clotting screen, C-reactive protein, electrolytes and bicarbonate level, electrolyte level, blood culture MCS, and liver function test. These tests reveal information on organ, blood, and respiratory functions. The fact that these tests are included demonstrates the thorough comprehensive approach used to identify and track respiratory illnesses.

5.1.5 Length of Stay(LOS)

The length of hospital stay (LOS) for patients with respiratory conditions was also looked at in the research. The standard deviation of 3.13 and an average LOS of almost 4.6 hours were recorded. From 0.08 hours (the shortest stay) to 121.32 hours (the longest stay), the range of LOS was observed. The distribution of the data had a considerable rightward skew and kurtosis, both signs of a non-normal distribution. This implies that traditional parametric tests might not be adequate for further study and that non-parametric tests might be more useful. The rightward skew in the LOS (length of hospital stay) distribution is an indication that there were more cases with longer stays than shorter stays.

5.1.6 Presenting complaints

Additionally, the research on the most common patient complaints and requested diagnostic procedures offers important new information on the particular requirements of patients with respiratory conditions. The most frequent problems patients mentioned was asthma, adult shortness of breath, feeling unwell,sepsis, cough aand chest pain.

5.1.7 Relationship between variables

The Effect of Gender, Agegroup and Diagnosis on LOS were accessed, Analysis of Variance(ANOVA), Median scores, Van Da Waerden scores and Wilcoxon scores were all used to test. Results showed that Age group and diagnosis has an effect on LOS, what this means is that people who were within a certain age group spent longer time during their hospital visits, also the diagnosis type showed an effect in the waiting time. Gender had no significant impact in patients'length of stay.

5.1.8 Model creation and evaluation

Using machine learning techniques, a prediction model for patient sickness was created in this study based on demographics and presenting complaints. Python was used to implement the model.

To provide acceptable accuracy outcomes in the models, the data was first examined and preprocessed. The dataset included 6774 rows and 14 columns, and it included several datatypes such as datetime, float, and object. The redundant columns were removed, and the column names were changed.

The data included 5 columns and 6772 rows after preprocessing. The two ailments with the greatest number of instances in the analysis were pneumonia and asthma. Other ailments were filtered off

The top five presenting complaints were selected and added as features for the model.

The data was split into training and testing sets in order to get it ready for model building. The test set comprised 915 rows, whereas the training set had 3658 rows. Likewise, the target variable was divided.

Four machine learning models were created, K-nearest Neighbours (KNN), Random Forest, Logistic Regression, and Decision Tree. Three nearest neighbours were used by the KNN model to categorise the unlabeled data. For asthma and pneumonia, it had an accuracy of 0.89 and 0.79, respectively. For both classes, the accuracy, recall, and F1-scores were also computed.

By evaluating the likelihood of an observation belonging to several classes, logistic regression created predictions. For both pneumonia and asthma, it had an accuracy of 0.81 and 0.88, respectively. For both classes, the precision, recall, and F1-scores were computed.

In order to produce a structure that resembled a tree, the decision tree model separated the data depending on chosen qualities. For asthma and pneumonia, it had an accuracy of 0.89 and 0.81, respectively. For both classes, the precision, recall, and F1-scores were computed.

For the Random forest model, With an accuracy rate of 89% in detecting asthma patients, the classifier did well at predicting asthma cases. In addition, it properly identified 81% of the pneumonia cases that were anticipated. With 88% of genuine instances of asthma and 82% of actual cases of pneumonia properly diagnosed, recall rates were likewise extremely high. According to the F1-score, the total performance was 88% for asthma and 82% for pneumonia.

Summarily, all the models were compared against each other and the Random forest model is selected as the best model as it performed very well in all the classification parameters.

5.2 Recommendations

1. Reduced patient wait times for common respiratory diseases can be achieved with the use of predictive analytics. When used to predict asthma and pneumonia cases, the generated models, like the random forest model, showed great accuracy, precision, and recall. For those with certain conditions, this information can be used to prioritise patient treatment and shorten wait times.
2. With the use of machine learning models, patients who visit the hospital can be pre-diagnosed and basic tests such as full blood count, ECG, urea level, clotting screen, C-reactive protein, electrolytes and bicarbonate level, electrolyte level, blood culture MCS can be ordered for patients ahead of seeing the doctor. This will cut down on patients' time in the hospital because the analysis revealed that some patients spend as much as 23 hours before test order.
3. Healthcare institutions can spot patterns and trends that lead to greater wait times for patients with respiratory illnesses by using demographic data and presenting complaints. Also, based on the time series analysis,some days of the week and time of the month revealed that respiratory ailments are more prevalent on some days and months. This knowledge may help with forecasting future events and assist in decision making as regards with staffing selection, process optimisation, and resource allocation to remove bottlenecks and boost patient triage and care delivery efficiency during periods of seasonal fluctuation.
4. Healthcare professionals may use predictive models, such the random forest model, to help them identify high-risk patients for example the elderly patients who are more likely to have longer wait times. Healthcare institutions may use this data to prioritise patients based on urgency and severity, ensuring that those who need treatment the most get it as soon as possible.
5. The results also emphasise how crucial data preparation and feature selection are for creating precise prediction models. The models may be optimised to focus on the most important factors for predicting patients' ailments by filtering and choosing pertinent characteristics, such as the top presenting complaints.
6. To maintain the prediction models' efficacy over time, they must be evaluated and improved continuously. For the models to retain their predictive power and accuracy as healthcare systems change and new elements are taken into consideration, continuous updating and recalibration is necessary.
7. The study shows how predictive analytics has the ability to improve patient care and deal with inefficiencies in healthcare systems. Healthcare facilities may streamline operations, manage resources wisely, and enhance patient experience by lowering wait times by utilising data-driven insights.

In conclusion, the application of predictive analytics to help shorten patient wait times for common respiratory disorders shows significant potential. Healthcare institutions may make wise judgements to prioritise resources, improve patient care, and eventually decrease patient wait times by utilising demographic data, presenting complaints, and machine learning models.

CHAPTER 6

CRITICAL REFLECTIONS

- **Methodology and Research Analysis**: The technique selected for the study seemed appropriate for analysing respiratory conditions and predicting patient outcomes. Comprehensive analysis and prediction were made possible through the application of machine learning models including K-nearest Neighbours, Random Forest, Logistic Regression, and Decision Tree. It is crucial to remember that the dataset may have had intrinsic flaws like incomplete data. A more thorough knowledge of respiratory illnesses would have been possible with a larger sample size and more varied data sources.

- **Findings and Interpretation**: The study's conclusions provide important light on a number of respiratory illnesses. The examination of demographic data revealed trends in patient visits, with the older group being more significantly affected. The time series study of respiratory patient visits showed seasonal fluctuations and an upward tendency. The most common respiratory ailments, pneumonia and asthma, were found to be asthma and pneumonia which reflects the impact of these illnesses in the healthcare sector. It is important to note, nevertheless, that the study might have taken a broader view by thoroughly examining other respiratory conditions.

- **Model Development and Evaluation**: In order to predict patient outcomes, machine learning models were developed and evaluated. The model with the highest accuracy, precision, recall, and F1 score for instances of both pneumonia and asthma emerged as the top performer. This demonstrates the promise of machine learning methods for precisely predicting respiratory illnesses. It is crucial to recognise, nonetheless, that the models' effectiveness strongly depends on the quality and variety of the data. By including additional useful features and investigating more advanced algorithms, further improvements may be realised with the predictive models.

- **Limitations and Biases**: This research has certain inherent limitations and possible biases, just like any other study. Selection bias could have been introduced since the dataset used may not have been completely representative of the total population. Furthermore, because patient medical histories and accurate symptom reporting may differ, the results may have been vulnerable to reporting biases. In order to achieve a more precise knowledge of respiratory illnesses, it is imperative to recognise and solve these limitations.

- **Areas for Improvement**; Several areas for improvement should be taken into account in order to strengthen the study. First, a more thorough study would result from enlarging the dataset to include information from other healthcare institutions or areas. Incorporating additional varied elements, such environmental aspects or patients' socioeconomic backgrounds, might also yield insightful results.

Bibliography

[1] Alanazi, R. (2022) Identification and prediction of chronic diseases using machine learning approach, Journal of Healthcare Engineering. Available at: https://www.hindawi.com/journals/jhe/2022/2826127/ (Accessed: 31 May 2023).

[2] Beedle, J. (2023) Data show 1.65 million patients in England faced 12-hour waits from time of arrival in a&es in 2022, RCEM. Available at: https://rcem.ac.uk/data-show-1-65-million-patients-in-england-faced-12-hour-waits-from-time-of-arrival-in-aes-in-2022 (Accessed: 09 June 2023).

[3] Benevento, E., Aloini, D. and Squicciarini, N. (2023) 'Towards a real-time prediction of waiting times in emergency departments: A Comparative Analysis of Machine Learning Techniques', International Journal of Forecasting, 39(1), pp. 192–208. doi:10.1016/j.ijforecast.2021.10.006.

[4] Breiman, L. (2001) Machine Learning, 45(1), pp. 5–32. doi:10.1023/a:1010933404324.

[5] Chu, H. et al. (2019) 'The psychology of The wait time experience – what clinics can do to manage the waiting experience for patients: A longitudinal, qualitative study', BMC Health Services Research, 19(1). doi:10.1186/s12913-019-4301-0.

[6] Clayton, T. et al. (2022) Respiratory disease: Understanding the future service and workforce needs, www.hee.nhs.uk. Available at: https://www.hee.nhs.uk/sites/default/files/documents/Respiratory

[7] El Naqa, I. and Murphy, M.J. (2015) 'What is machine learning?', Machine Learning in Radiation Oncology, pp. 3–11. doi:10.1007/978-3-319-18305-3_1.

[8] Horwitz, L.I., Green, J. and Bradley, E.H. (2010) 'US emergency department performance on Wait Time and length of visit', Annals of Emergency Medicine, 55(2), pp. 133–141. doi:10.1016/j.annemergmed.2009.07.023.

[9] Li, L. et al. (2022) Prediction and diagnosis of respiratory disease by combining convolutional neural network and bi-directional long short-term memory methods [Preprint]. doi:10.21203/rs.3.rs-1332028/v1.

[10] Li, X. et al. (2021) 'Artificial intelligence-assisted reduction in patients' waiting time for outpatient process: A retrospective cohort study', BMC Health Services Research, 21(1). doi:10.1186/s12913-021-06248-z.

[11] Li, X. et al. (2022) 'Using artificial intelligence to reduce queuing time and improve satisfaction in pediatric outpatient service: A randomized clinical trial', Frontiers in Pediatrics, 10. doi:10.3389/fped.2022.929834.

[12] Mall, S. et al. (2022) 'Implementation of machine learning techniques for disease diagnosis', Materials Today: Proceedings, 51, pp. 2198–2201. doi:10.1016/j.matpr.2021.11.274.

[13] Mangat, P.K. and Saini, K.S. (2021) 'Health Care Prediction Using Predictive Analytics', 2021 10th International Conference on System Modeling amp; Advancement in Research Trends (SMART), pp. 64–70. doi:10.1109/smart52563.2021.9676220.

[14] Naser, A.Y. et al. (2021) 'Hospital admission trends due to respiratory diseases in England and Wales between 1999 and 2019: An ecologic study', BMC Pulmonary Medicine, 21(1). doi:10.1186/s12890-021-01736-8.

[15] NHS (2023) A and E waiting times, england.nhs.uk. Available at: https://www.nuffieldtrust.org.uk/resource/a-e-waiting-times: :text=Background,orNHS, digital (2021) NHS choices. Available at: https://digital.nhs.uk/data-and-information/publications/statistical/provisional-monthly-hospital-episode-statistics-for-admitted-patient-care-outpatient-and-accident-and-emergency-data/april-2020—march-2021-m13 (Accessed: 27 June 2023).

[16] Nithya, B. and Ilango, V. (2017) 'Predictive analytics in health care using machine learning tools and Techniques B. Nithya and V. Ilango, ', 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 492–499. doi:10.1109/iccons.2017.8250771.

[17] Paling, S. et al. (2020) 'Waiting times in emergency departments: Exploring the factors associated with longer patient waits for emergency care in England using routinely collected daily data', Emergency Medicine Journal [Preprint]. doi:10.1136/emermed-2019-208849.

[18] Pappas, P.A. and DePuy, V., 2004. An overview of non-parametric tests in SAS: when, why, and how. Paper TU04. Duke Clinical Research Institute, Durham, pp.1-5.

[19] Poreva, A., Karplyuk, Y. and Vaityshyn, V. (2017) 'Machine learning techniques application for lung diseases diagnosis', 2017 5th IEEE Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE) [Preprint]. doi:10.1109/aieee.2017.8270528.

[20] Prasad, B.D., Prasad, P.E. and Sagar, Y. (2011) 'A comparative study of machine learning algorithms as expert systems in medical diagnosis (asthma)', Advances in Computer Science and Information Technology, pp. 570–576. doi:10.1007/978-3-642-17857-3_56.

[21] Rasjid, Z.E. (2021) 'Predictive analytics in Healthcare: The use of machine learning for diagnoses', 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET), pp. 1–6. doi:10.1109/icecet52533.2021.9698508.

[22] Reichert, A. and Jacobs, R. (2018) 'The impact of waiting time on patient outcomes: Evidence from early intervention in psychosis services in England', Health Economics, 27(11), pp. 1772–1787. doi:10.1002/hec.3800.

[23] SAS (2013) The NPAR1WAY procedure - SAS support. Available at: https://support.sas.com/documentation/onlinedoc/stat/131/npar1way.pdf (Accessed: 19 August 2023).

[24] Stewart, C. (2022) Topic: Respiratory disease in the UK, Statista. Available at: https://www.statista.com/topics/5908/respiratory-disease-in-the-uk/#topicOverview (Accessed: 13 June 2023).

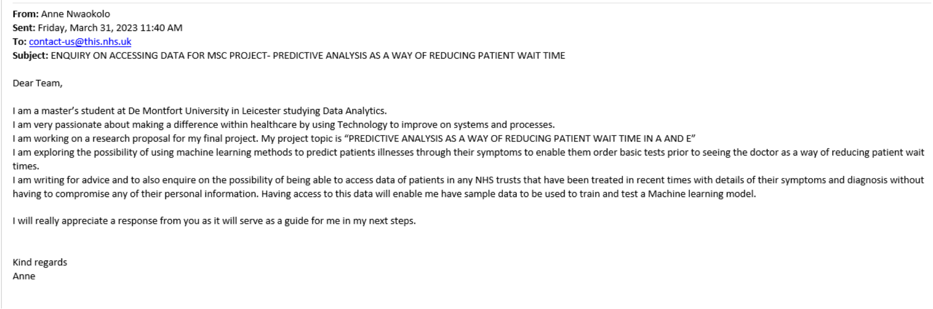
[25] THIS (2022) Predictive analytics - this. Available at: https://www.this.nhs.uk/fileadmin/content_uploads/insights/predictive-analytics/THIS-Predictive-Analytics-White-Paper.pdf (Accessed: 12 July 2023).

APPENDIX A

Appendix

A.1 Project Resources

- **Data source**



From: Anne Nwaokolo
Sent: Friday, March 31, 2023 11:40 AM
To: contact-us@this.nhs.uk
Subject: ENQUIRY ON ACCESSING DATA FOR MSC PROJECT- PREDICTIVE ANALYSIS AS A WAY OF REDUCING PATIENT WAIT TIME

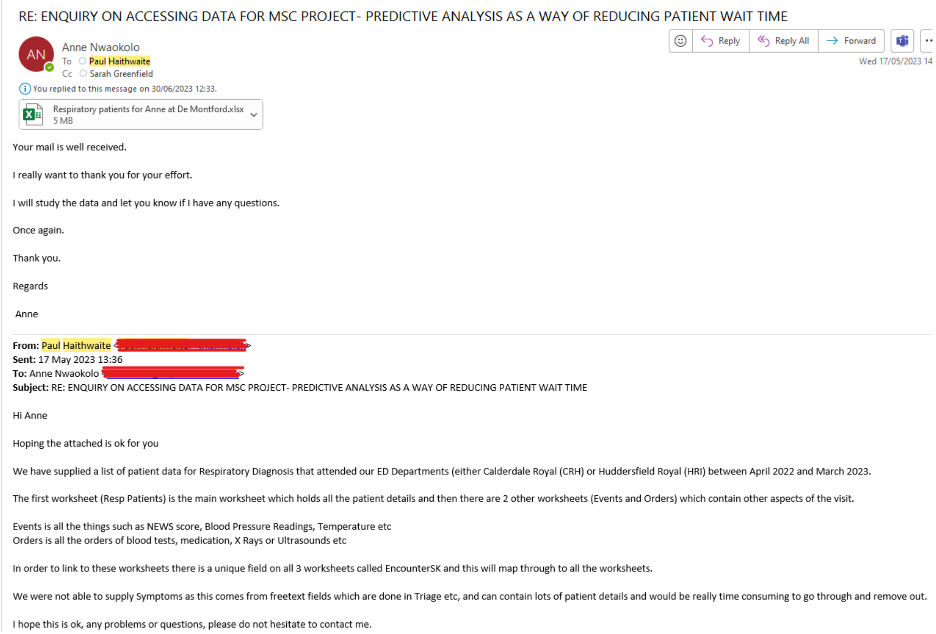
Dear Team,

I am a master's student at De Montfort University in Leicester studying Data Analytics.
I am very passionate about making a difference within healthcare by using Technology to improve on systems and processes.
I am working on a research proposal for my final project. My project topic is "PREDICTIVE ANALYSIS AS A WAY OF REDUCING PATIENT WAIT TIME IN A AND E"
I am exploring the possibility of using machine learning methods to predict patients illnesses through their symptoms to enable them order basic tests prior to seeing the doctor as a way of reducing patient wait times.
I am writing for advice and to also enquire on the possibility of being able to access data of patients in any NHS trusts that have been treated in recent times with details of their symptoms and diagnosis without having to compromise any of their personal information. Having access to this data will enable me have sample data to be used to train and test a Machine learning model.

I will really appreciate a response from you as it will serve as a guide for me in my next steps.

Kind regards
Anne

Figure A.1: Email requesting data from The Health Informatics Service(THIS)



RE: ENQUIRY ON ACCESSING DATA FOR MSC PROJECT- PREDICTIVE ANALYSIS AS A WAY OF REDUCING PATIENT WAIT TIME

AN Anne Nwaokolo
To ○ Paul Haithwaite
Cc ○ Sarah Greenfield

Wed 17/05/2023 14

You replied to this message on 30/06/2023 12:33.

Respiratory patients for Anne at De Montford.xlsx
5 MB

Your mail is well received.

I really want to thank you for your effort.

I will study the data and let you know if I have any questions.

Once again.

Thank you.

Regards

Anne

From: Paul Haithwaite <[redacted]>
Sent: 17 May 2023 13:36
To: Anne Nwaokolo <[redacted]>
Subject: RE: ENQUIRY ON ACCESSING DATA FOR MSC PROJECT- PREDICTIVE ANALYSIS AS A WAY OF REDUCING PATIENT WAIT TIME

Hi Anne

Hoping the attached is ok for you

We have supplied a list of patient data for Respiratory Diagnosis that attended our ED Departments (either Calderdale Royal (CRH) or Huddersfield Royal (HRI) between April 2022 and March 2023.

The first worksheet (Resp Patients) is the main worksheet which holds all the patient details and then there are 2 other worksheets (Events and Orders) which contain other aspects of the visit.

Events is all the things such as NEWS score, Blood Pressure Readings, Temperature etc
Orders is all the orders of blood tests, medication, X Rays or Ultrasounds etc

In order to link to these worksheets there is a unique field on all 3 worksheets called EncounterSK and this will map through to all the worksheets.

We were not able to supply Symptoms as this comes from freetext fields which are done in Triage etc, and can contain lots of patient details and would be really time consuming to go through and remove out.

I hope this is ok, any problems or questions, please do not hesitate to contact me.

Figure A.2: Email response with data from The Health Informatics Service(THIS)

- **Project proposal**

The Use of Predictive Analytics in Assisting the Reduction of Patient Waiting Times: A Study of Common Respiratory Ailments.

Anne Awele Nwaokolo (P2743914@my365.dmu.ac.uk)
Supervisor: Sarah Greenfield
De Montfort University, Leicester, UK

BACKGROUND:

Several studies indicate a strong inverse relationship between patient satisfaction and waiting time. Patients should not have to wait too long for appointments and consultations in a well-designed healthcare system (Li *et al*., 2021) Longer wait times in emergency rooms and longer visits have a negative impact on care quality and raise the risk of bad outcomes (Horwitz *et al*., 2010).

The Handbook to the NHS Constitution has a commitment to the four-hour A&E waiting time objective. At least 95% of patients who visit A&E should be hospitalised, transferred, or released within four hours, according to the operational criteria established in 2010. A 76% intermediate threshold objective was adopted in December 2022 with a projected improvement of 2% in 2024/25(NHS, 2023). According to NHS England, 347,703 patients had to wait 12 hours between the decision to admit them and their admission in 2022.

The Royal College of Emergency Medicine (Beedle, 2023) suggests that there is one patient mortality for every 72 patients who wait 8 to 12 hours after arriving at an emergency room. By using the Standard Mortality Ratio (SMR) to determine the total number of 12-hour time of arrival delays for 2022, we can estimate that there were 23,003 patient fatalities in England because of long wait times.

In England, respiratory illness is the third leading cause of mortality and affects one in five people (Clayton, *et al).* The mortality rate from respiratory disease in 2020 was 89 deaths per 100,000 women and 130 deaths per 100,000 men in England. In the winter, older adults are more vulnerable to respiratory illnesses. In England and Wales during the winter of 2019–20, underlying respiratory disorders were the likely cause of almost 8.3 thousand winter deaths among people over 75. About a sixth of people between the ages of 75 and 79 (Stewart, 2022).

Due to the introduction of Information Technology, the healthcare industry has advanced more recently. IT is used in healthcare with the goal of improving patient care while lowering costs and enhancing comfort (Alanazi 2022).

Predictive analytics is a subset of advanced analytics that makes predictions about the unknown. Predictive analytics examines current discoveries to create predictions about the future by utilizing a variety of approaches from data mining, statistics, modelling, machine learning, and artificial intelligence (Nithya,2017)

Various healthcare provider areas are supported by predictive analytics. It attempts to improve clinical outcomes, improve patient care, optimise resource use, and properly diagnose illnesses. By maximising the cost, predictive analytics aids organisations in planning for health care. The application of predictive analytics in this sector is anticipated to produce effective results by raising the level of Patient

Satisfaction. The future of the health care sector will be transformed by predictive analytics (Nithya,2017)

This study aims to critically analyse the current wait time in NHS hospitals, examine the demographics and other activities of patients with common respiratory illnesses with a view of introducing machine learning predictions that can be used to perform quick diagnosis of patients conditions prior to them seeing a doctor in order for quick checks and tests to be carried out as a way of reducing the amount of time spent in the hospital and improve the patients outcome.

INTRODUCTION:

All healthcare systems have long struggled with Emergency Department (ED) overcrowding. The ageing population and the widespread adoption of rigorous cost reduction strategies are the major causes of this imbalance between supply and demand for emergency care (Benevento, *et al* 2023)

Predictive analytics helps several healthcare sectors. To enhance patient care, resource efficiency, and accurate disease diagnosis. Through improved patient experience, it is projected that the use of predictive analytics in this industry would lead to successful outcomes. Predictive analytics will change the way that the health care industry operates in the future (Nithya, 2017).

In recent years, the fast-developing topic of predictive analytics in healthcare has drawn increased interest due to its potential to enhance patient outcomes and resource efficiency.
Healthcare providers understand how crucial it is to cut down patient wait times, especially for those who have urgent medical needs like common respiratory diseases.

Machine Learning is a field of computing algorithms and is constantly developing and aims to replicate human intelligence by learning from the environment. In the brand-new era of "big data," they are regarded as the workhorse. Machine learning methods have been effectively used in a variety of industries, including banking, entertainment, biomedicine, pattern recognition, computer vision, spacecraft engineering, and computational biology. Ionising radiation (radiotherapy), which is the primary therapeutic option for advanced stages of local illness, is given to more than half of cancer patients (El Naga & Murphy 2015).

It has been demonstrated that waiting times not only have a negative impact on patient satisfaction but can also put patient outcomes and safety at risk while increasing the chances of mortality rate (Beedle, 2023)

Data obtained from NHS Digital, in a summary of primary diagnosis for finished consultancy episodes for 2021 to 2022, It shows that for Pneumonia and Influenza, which are common respiratory illnesses, a total of 414,802 cases were recorded which is a very high number as compared to other illnesses and it falls within the top 20 ailments for consultancy episodes (NHS, 2021).

In summary, using predictive analytics in the healthcare industry has the potential of enhancing patient outcomes, resource effectiveness, and precise illness identification. Healthcare professionals can ensure effective healthcare delivery by having knowledge of high-risk patients, especially for those with urgent medical needs such common respiratory infections. Big data, machine learning, and predictive analytics have the potential to completely change the healthcare industry and the way doctors provide care. This study aims to add to the growing body of knowledge on the effectiveness of predictive analytics in the healthcare industry, particularly in reducing patient waiting times for common respiratory ailments.

PROPOSED WORK:

AIMS

- To assess the existing waiting times at NHS hospitals for individuals with common respiratory conditions.
- To create a Machine Learning predictive model that healthcare professionals may employ to help cut down on patient wait times.

OBJECTIVES

- Carry out literature review.
- Obtain data on patient demographics, symptoms, and waiting times for respiratory illnesses from selected healthcare facilities.
- Data preparation, cleaning, and transformation to usable format.
- Analyse the data to find trends, patterns, and correlations.
- Assess the wait times for patients with respiratory conditions.
- Develop predictive analytics models.
- Perform model validation and comparison.
- Select the most accurate model for implementation.
- Analyse the study's findings.
- Finalise the report with the results, conclusions, and recommendations.

RESEARCH QUESTIONS:

1. How long do patients with respiratory conditions spend in hospitals?
2. How do factors such as demographics and hospital activities affect the length of time for patients with respiratory conditions?
3. Can predictive analytics be used to reduce the waiting times and optimize patients' experience?

PROJECT REQUIREMENTS:

- Software: Python, Jupyter Notebooks, SAS enterprise Miner, Power BI, Tableau etc.
- Hardware: Computers from DMU labs and personal computers.
- Data: Access to patient data from selected healthcare facilities.
- Academic literature: Books, articles, and journals.

- Access to Kimberlin Library.
- Access to De Montfort University computer labs.

PROJECT RISKS:

- Non access to computer labs
- Non access to Kimberlin library

PROJECT SCHEDULE:

Phase 1 – Literature review, Data Collection, Preparation and Analysis (March to May 2023)

- Carry out literature review.
- Gather and compile useful resources such as articles and journals.
- Obtain data on patient demographics, symptoms, and waiting times for respiratory illnesses from selected healthcare facilities.
- Data preparation, cleaning, and transformation to usable format.
- Analyse the data to find trends, patterns, and correlations.
- Assess the wait times for patients with respiratory conditions.

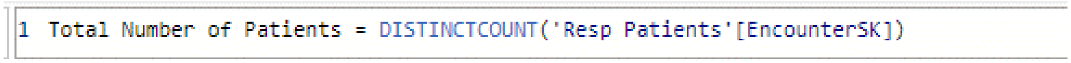
Phase 2 – Model Development, Results, Recommendations and Report writing (June to August 2023)

- Develop predictive analytics models.
- Perform model validation and comparison.
- Select the most accurate model for implementation.
- Analyse the study's findings.
- Finalise the report with the results, conclusions, and recommendations.

Project Work Plan

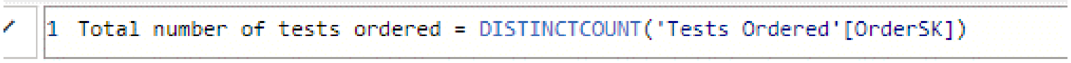
The use of predictive Analytics in Assisting the Reduction of Patient Waiting times: A study of common respiratory Ailments						
		Phase 1			**Phase 2**	
	Week 22 -26 (March 1st to March 31st 2023)	Week 27 - 29(April 4th to April 30th	Week 31-34(May 1st to May 28th 2023)	Week 35-39 (June 1st to June 30th 2023)	Week 40-43 (July 3rd to July 28th	Week 44-47 (August 1st to August 27th
Literature review, Data Collection, Preparation and Analysis						
• Carry out literature review	🟦					
• Obtain data on patient demographics, symptoms, and waiting times for respiratory illnesses from selected healthcare facilities.		🟦				
• Data preparation, cleaning and transformation to usable format			🟦			
• Analyse the data to find trends, patterns, and correlations.				🟦		
• Assess the wait times for patients with respiratory conditions.				🟦		
Model Development, Results, Recommendations and Report writing						
• Develop predictive analytics models.					🟦	
• Perform model validation and comparison.					🟦	
• Select the most accurate model for implementation.					🟦	
• Analyse the study's findings.						🟦
• Prepare report with the results, conclusions, and recommendations.	🟦	🟦	🟦	🟦	🟦	🟦

A.2 Power BI measures



```
Total Number of Patients = DISTINCTCOUNT('Resp Patients'[EncounterSK])
```

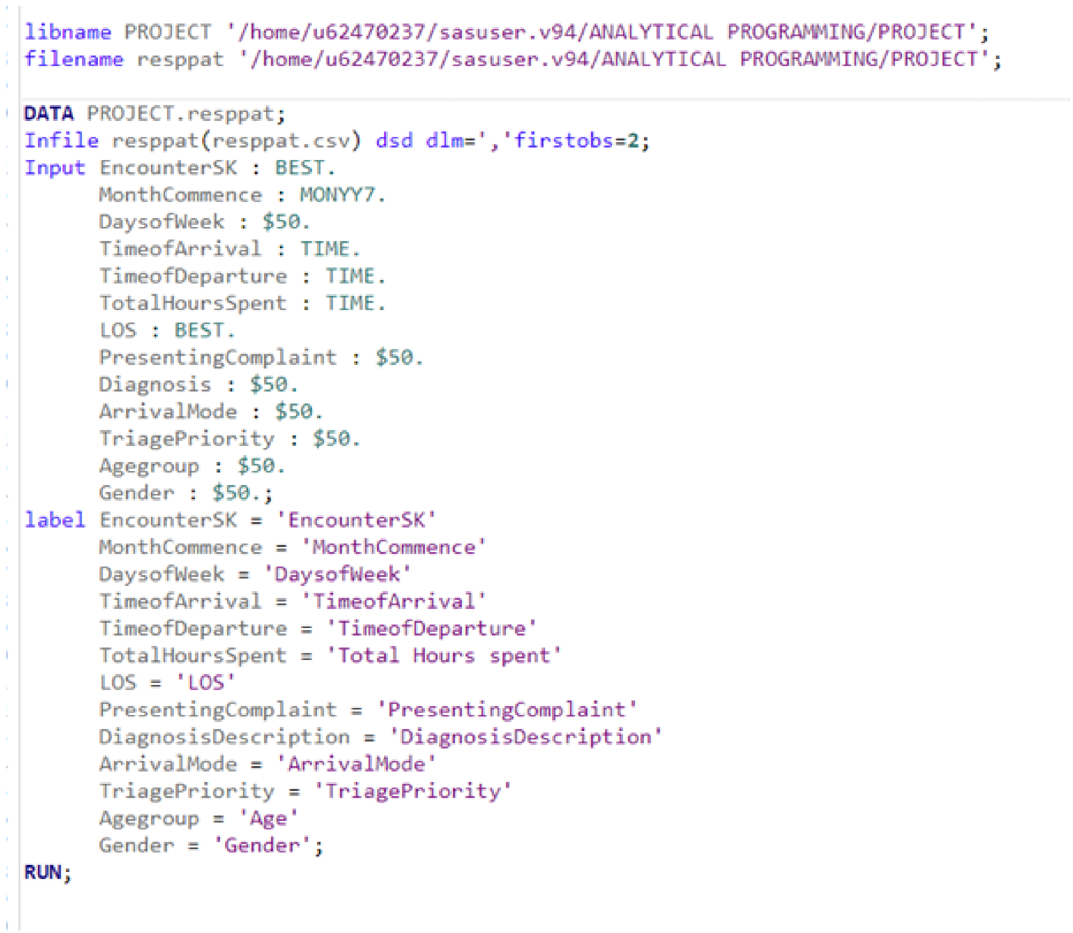
Figure A.3: Power BI measure for ailments by total finished consultant episodes 2021 to 2022



```
Total number of tests ordered = DISTINCTCOUNT('Tests Ordered'[OrderSK])
```

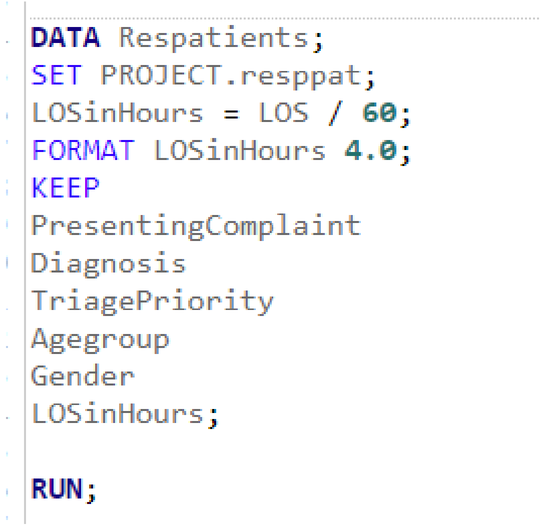
Figure A.4: Power BI measure for total number of tests ordered

A.3 SAS Codes



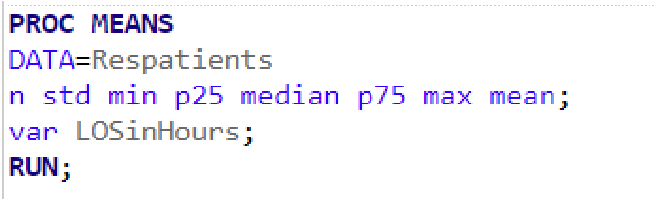
```
libname PROJECT '/home/u62470237/sasuser.v94/ANALYTICAL PROGRAMMING/PROJECT';
filename resppat '/home/u62470237/sasuser.v94/ANALYTICAL PROGRAMMING/PROJECT';

DATA PROJECT.resppat;
Infile resppat(resppat.csv) dsd dlm=','firstobs=2;
Input EncounterSK : BEST.
      MonthCommence : MONYY7.
      DaysofWeek : $50.
      TimeofArrival : TIME.
      TimeofDeparture : TIME.
      TotalHoursSpent : TIME.
      LOS : BEST.
      PresentingComplaint : $50.
      Diagnosis : $50.
      ArrivalMode : $50.
      TriagePriority : $50.
      Agegroup : $50.
      Gender : $50.;
label EncounterSK = 'EncounterSK'
      MonthCommence = 'MonthCommence'
      DaysofWeek = 'DaysofWeek'
      TimeofArrival = 'TimeofArrival'
      TimeofDeparture = 'TimeofDeparture'
      TotalHoursSpent = 'Total Hours spent'
      LOS = 'LOS'
      PresentingComplaint = 'PresentingComplaint'
      DiagnosisDescription = 'DiagnosisDescription'
      ArrivalMode = 'ArrivalMode'
      TriagePriority = 'TriagePriority'
      Agegroup = 'Age'
      Gender = 'Gender';
RUN;
```

Figure A.5: SAS code for library and file path creation and data importation



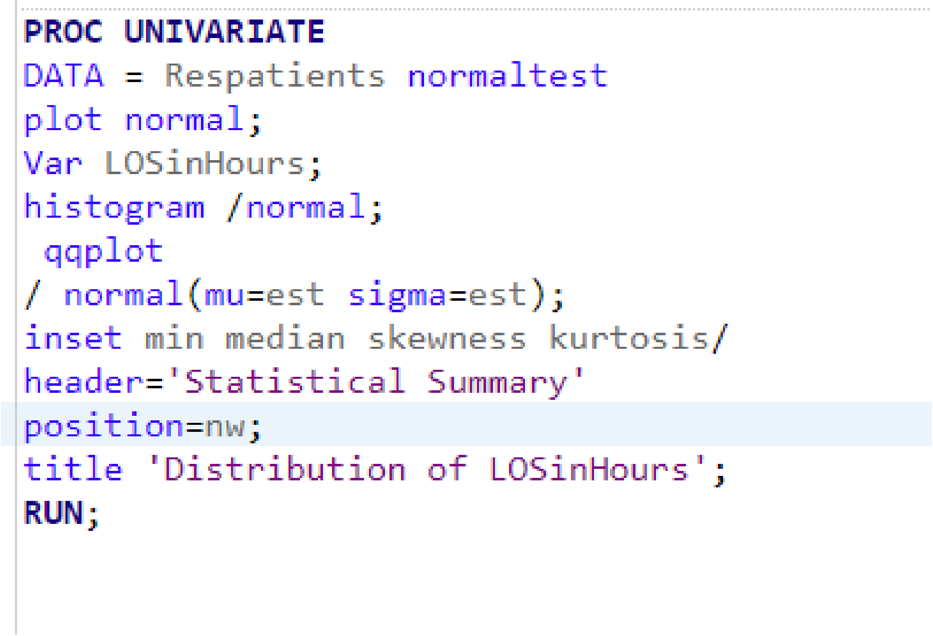
```
DATA Respatients;
SET PROJECT.resppat;
LOSinHours = LOS / 60;
FORMAT LOSinHours 4.0;
KEEP
PresentingComplaint
Diagnosis
TriagePriority
Agegroup
Gender
LOSinHours;

RUN;
```

Figure A.6: SAS code for LOSinHours variable creation.png



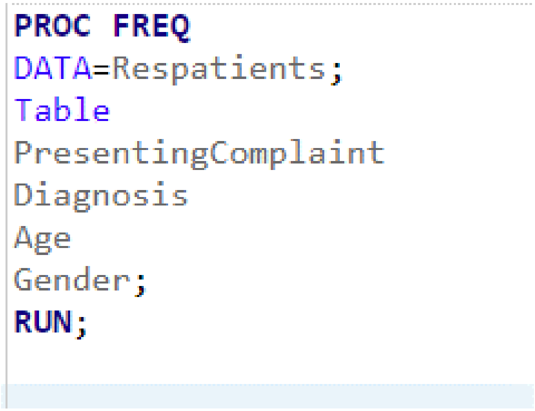
```
PROC MEANS
DATA=Respatients
n std min p25 median p75 max mean;
var LOSinHours;
RUN;
```

Figure A.7: Proc means statement for LOS



```
PROC UNIVARIATE
DATA = Respatients normaltest
plot normal;
Var LOSinHours;
histogram /normal;
 qqplot
/ normal(mu=est sigma=est);
inset min median skewness kurtosis/
header='Statistical Summary'
position=nw;
title 'Distribution of LOSinHours';
RUN;
```

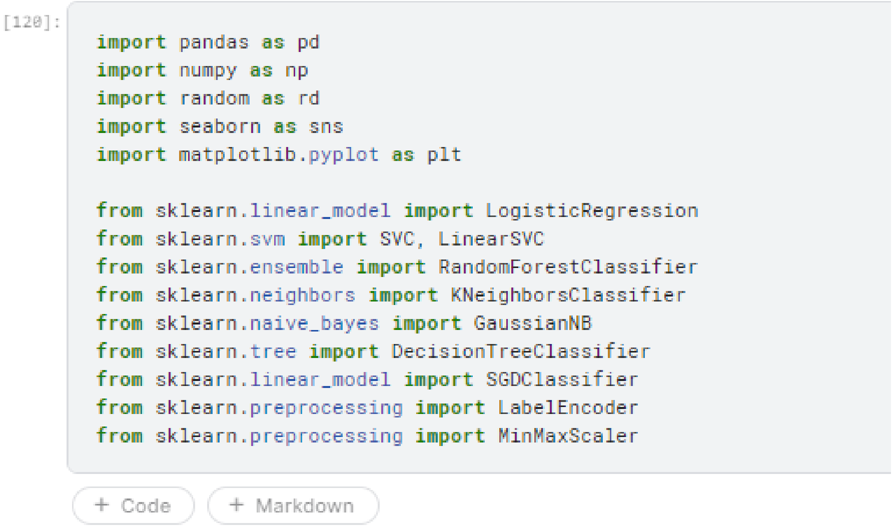
Figure A.8: proc univariate for normality testing



```
PROC FREQ
DATA=Respatients;
Table
PresentingComplaint
Diagnosis
Age
Gender;
RUN;
```

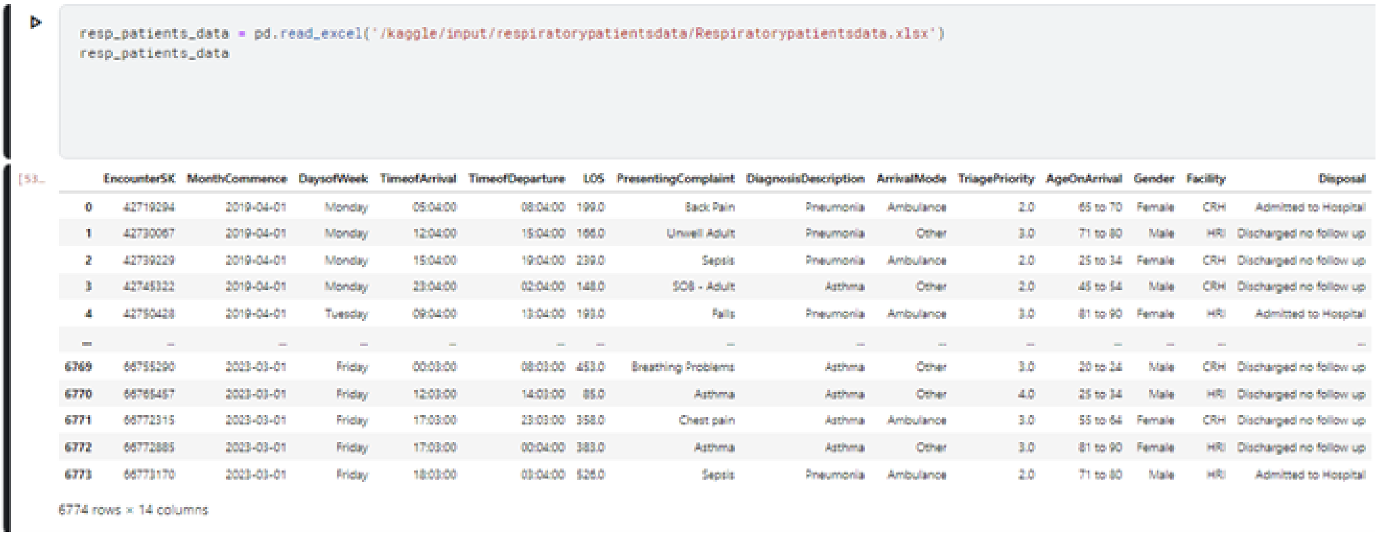
Figure A.9: Proc freq statement for frequency of the variables

A.4 Python codes.



```
[120]:
import pandas as pd
import numpy as np
import random as rd
import seaborn as sns
import matplotlib.pyplot as plt

from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC, LinearSVC
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.linear_model import SGDClassifier
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import MinMaxScaler
```

Figure A.10: Python code for importing libraries



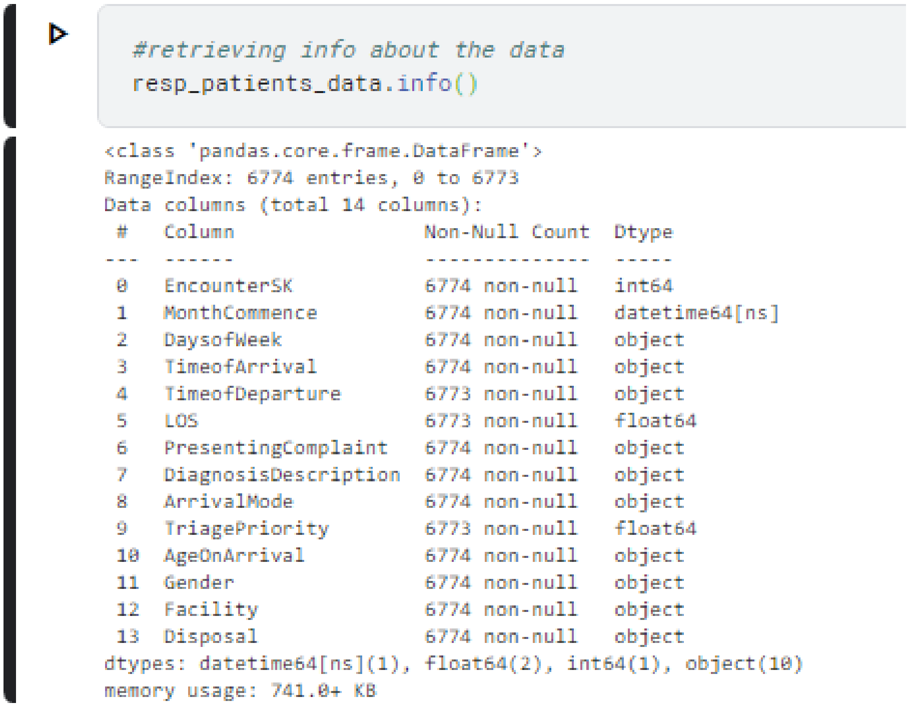
```
resp_patients_data = pd.read_excel('/kaggle/input/respiratorypatientsdata/Respiratorypatientsdata.xlsx')
resp_patients_data
```

[53…

	EncounterSK	MonthCommence	DaysofWeek	TimeofArrival	TimeofDeparture	LOS	PresentingComplaint	DiagnosisDescription	ArrivalMode	TriagePriority	AgeOnArrival	Gender	Facility	Disposal
0	42719294	2019-04-01	Monday	05:04:00	08:04:00	199.0	Back Pain	Pneumonia	Ambulance	2.0	65 to 70	Female	CRH	Admitted to Hospital
1	42730067	2019-04-01	Monday	12:04:00	15:04:00	166.0	Unwell Adult	Pneumonia	Other	3.0	71 to 80	Male	HRI	Discharged no follow up
2	42739229	2019-04-01	Monday	15:04:00	19:04:00	239.0	Sepsis	Pneumonia	Ambulance	2.0	25 to 34	Female	CRH	Discharged no follow up
3	42745322	2019-04-01	Monday	23:04:00	02:04:00	148.0	SOB - Adult	Asthma	Other	2.0	45 to 54	Male	CRH	Discharged no follow up
4	42750428	2019-04-01	Tuesday	09:04:00	13:04:00	193.0	Falls	Pneumonia	Ambulance	3.0	81 to 90	Female	HRI	Admitted to Hospital
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
6769	66755290	2023-03-01	Friday	00:03:00	08:03:00	453.0	Breathing Problems	Asthma	Other	3.0	20 to 24	Male	CRH	Discharged no follow up
6770	66765457	2023-03-01	Friday	12:03:00	14:03:00	85.0	Asthma	Asthma	Other	4.0	25 to 34	Male	HRI	Discharged no follow up
6771	66772315	2023-03-01	Friday	17:03:00	23:03:00	358.0	Chest pain	Asthma	Ambulance	3.0	55 to 64	Female	CRH	Discharged no follow up
6772	66772885	2023-03-01	Friday	17:03:00	00:04:00	383.0	Asthma	Asthma	Other	3.0	81 to 90	Female	HRI	Discharged no follow up
6773	66773170	2023-03-01	Friday	18:03:00	03:04:00	526.0	Sepsis	Pneumonia	Ambulance	2.0	71 to 80	Male	HRI	Admitted to Hospital

6774 rows × 14 columns

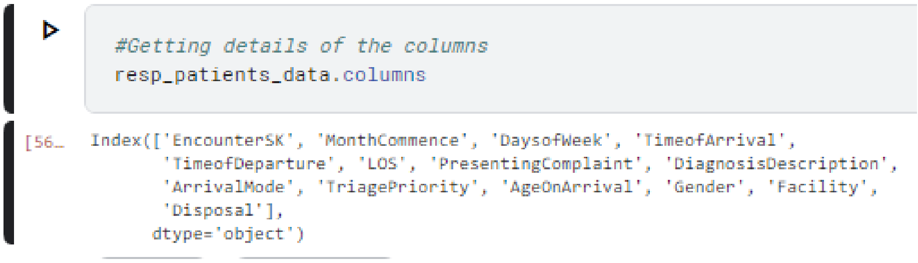
Figure A.11: Python code for reading in the data



```
#retrieving info about the data
resp_patients_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6774 entries, 0 to 6773
Data columns (total 14 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   EncounterSK          6774 non-null   int64
 1   MonthCommence        6774 non-null   datetime64[ns]
 2   DaysofWeek           6774 non-null   object
 3   TimeofArrival        6774 non-null   object
 4   TimeofDeparture      6773 non-null   object
 5   LOS                  6773 non-null   float64
 6   PresentingComplaint  6774 non-null   object
 7   DiagnosisDescription 6774 non-null   object
 8   ArrivalMode          6774 non-null   object
 9   TriagePriority       6773 non-null   float64
 10  AgeOnArrival         6774 non-null   object
 11  Gender               6774 non-null   object
 12  Facility             6774 non-null   object
 13  Disposal             6774 non-null   object
dtypes: datetime64[ns](1), float64(2), int64(1), object(10)
memory usage: 741.0+ KB
```

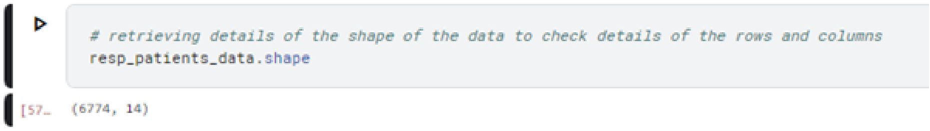
Figure A.12: Python code for retrieving info about the data



```
#Getting details of the columns
resp_patients_data.columns
```

```
[56…  Index(['EncounterSK', 'MonthCommence', 'DaysofWeek', 'TimeofArrival',
             'TimeofDeparture', 'LOS', 'PresentingComplaint', 'DiagnosisDescription',
             'ArrivalMode', 'TriagePriority', 'AgeOnArrival', 'Gender', 'Facility',
             'Disposal'],
            dtype='object')
```

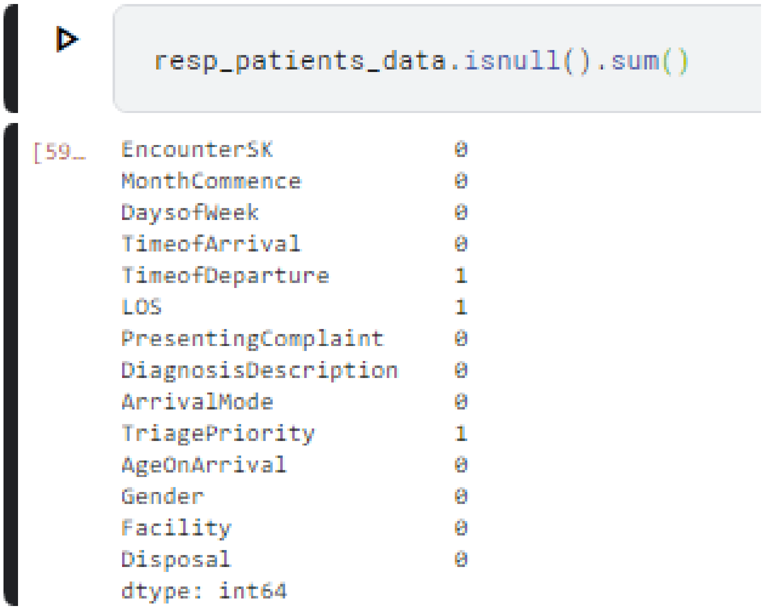
Figure A.13: Python code and result for column details



```
# retrieving details of the shape of the data to check details of the rows and columns
resp_patients_data.shape
```

```
[57…  (6774, 14)
```

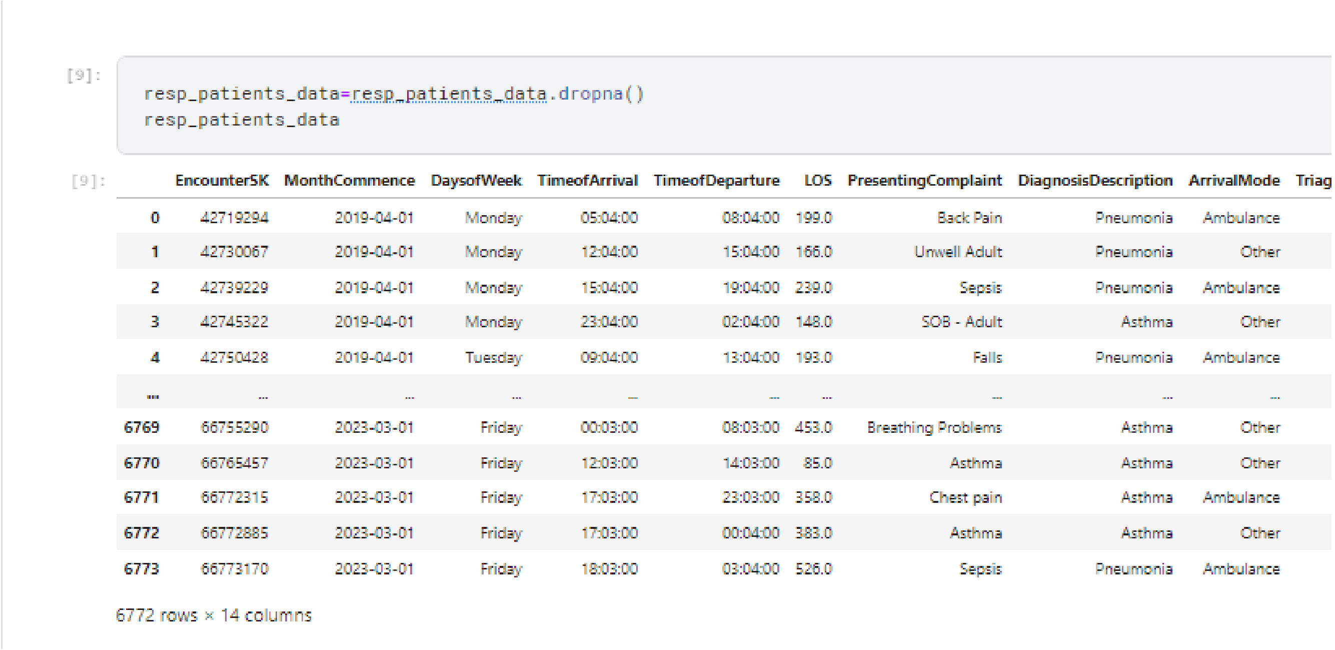
Figure A.14: Python code for column details



```
resp_patients_data.isnull().sum()
```

```
[59_  EncounterSK             0
      MonthCommence           0
      DaysofWeek              0
      TimeofArrival           0
      TimeofDeparture         1
      LOS                     1
      PresentingComplaint     0
      DiagnosisDescription    0
      ArrivalMode             0
      TriagePriority          1
      AgeOnArrival            0
      Gender                  0
      Facility                0
      Disposal                0
      dtype: int64
```

Figure A.15: Python code and result for counting missing values



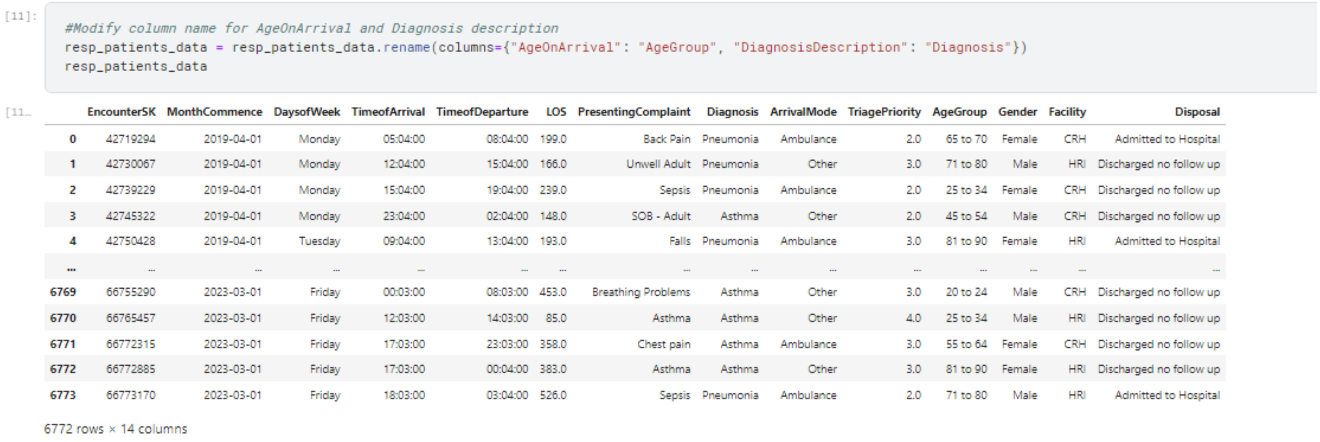
```
[9]:  resp_patients_data=resp_patients_data.dropna()
      resp_patients_data
```

[9]:

	EncounterSK	MonthCommence	DaysofWeek	TimeofArrival	TimeofDeparture	LOS	PresentingComplaint	DiagnosisDescription	ArrivalMode	Triag
0	42719294	2019-04-01	Monday	05:04:00	08:04:00	199.0	Back Pain	Pneumonia	Ambulance	
1	42730067	2019-04-01	Monday	12:04:00	15:04:00	166.0	Unwell Adult	Pneumonia	Other	
2	42739229	2019-04-01	Monday	15:04:00	19:04:00	239.0	Sepsis	Pneumonia	Ambulance	
3	42745322	2019-04-01	Monday	23:04:00	02:04:00	148.0	SOB - Adult	Asthma	Other	
4	42750428	2019-04-01	Tuesday	09:04:00	13:04:00	193.0	Falls	Pneumonia	Ambulance	
...	...	...	...	...	...	...	...	...	...	
6769	66755290	2023-03-01	Friday	00:03:00	08:03:00	453.0	Breathing Problems	Asthma	Other	
6770	66765457	2023-03-01	Friday	12:03:00	14:03:00	85.0	Asthma	Asthma	Other	
6771	66772315	2023-03-01	Friday	17:03:00	23:03:00	358.0	Chest pain	Asthma	Ambulance	
6772	66772885	2023-03-01	Friday	17:03:00	00:04:00	383.0	Asthma	Asthma	Other	
6773	66773170	2023-03-01	Friday	18:03:00	03:04:00	526.0	Sepsis	Pneumonia	Ambulance	

6772 rows × 14 columns

Figure A.16: Python code and result for dropping missing values

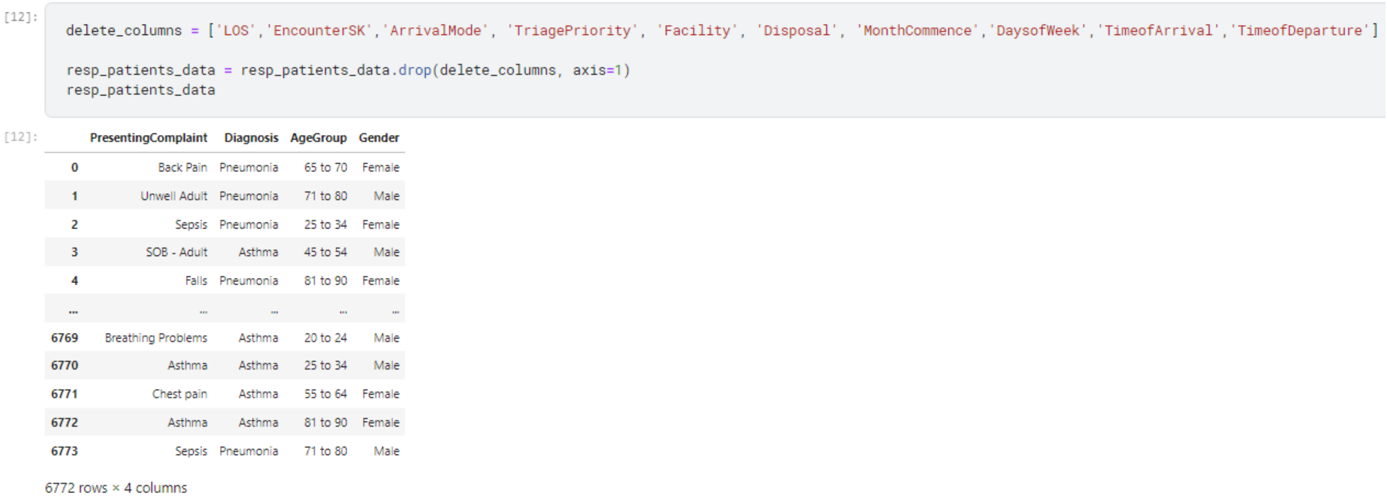


```python
#Modify column name for AgeOnArrival and Diagnosis description
resp_patients_data = resp_patients_data.rename(columns={"AgeOnArrival": "AgeGroup", "DiagnosisDescription": "Diagnosis"})
resp_patients_data
```

	EncounterSK	MonthCommence	DaysofWeek	TimeofArrival	TimeofDeparture	LOS	PresentingComplaint	Diagnosis	ArrivalMode	TriagePriority	AgeGroup	Gender	Facility	Disposal
0	42719294	2019-04-01	Monday	05:04:00	08:04:00	199.0	Back Pain	Pneumonia	Ambulance	2.0	65 to 70	Female	CRH	Admitted to Hospital
1	42730067	2019-04-01	Monday	12:04:00	15:04:00	166.0	Unwell Adult	Pneumonia	Other	3.0	71 to 80	Male	HRI	Discharged no follow up
2	42739229	2019-04-01	Monday	15:04:00	19:04:00	239.0	Sepsis	Pneumonia	Ambulance	2.0	25 to 34	Female	CRH	Discharged no follow up
3	42745322	2019-04-01	Monday	23:04:00	02:04:00	148.0	SOB - Adult	Asthma	Other	2.0	45 to 54	Male	CRH	Discharged no follow up
4	42750428	2019-04-01	Tuesday	09:04:00	13:04:00	193.0	Falls	Pneumonia	Ambulance	3.0	81 to 90	Female	HRI	Admitted to Hospital
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
6769	66755290	2023-03-01	Friday	00:03:00	08:03:00	453.0	Breathing Problems	Asthma	Other	3.0	20 to 24	Male	CRH	Discharged no follow up
6770	66765457	2023-03-01	Friday	12:03:00	14:03:00	85.0	Asthma	Asthma	Other	4.0	25 to 34	Male	HRI	Discharged no follow up
6771	66772315	2023-03-01	Friday	17:03:00	23:03:00	358.0	Chest pain	Asthma	Ambulance	3.0	55 to 64	Female	CRH	Discharged no follow up
6772	66772885	2023-03-01	Friday	17:03:00	00:04:00	383.0	Asthma	Asthma	Other	3.0	81 to 90	Female	HRI	Discharged no follow up
6773	66773170	2023-03-01	Friday	18:03:00	03:04:00	526.0	Sepsis	Pneumonia	Ambulance	2.0	71 to 80	Male	HRI	Admitted to Hospital

6772 rows × 14 columns

Figure A.17: Python code and result for modifying column names Age group and Diagnosis

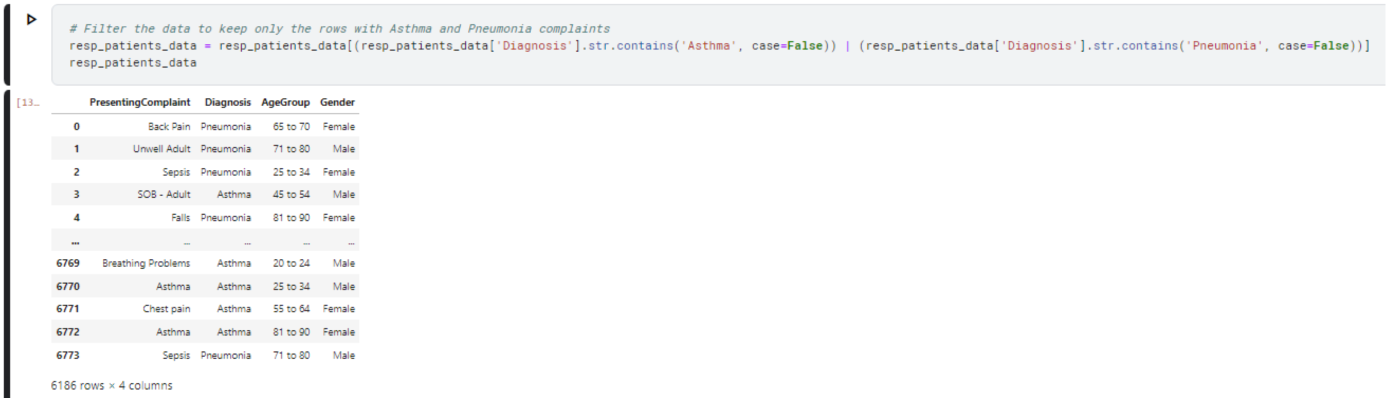


```python
delete_columns = ['LOS','EncounterSK','ArrivalMode', 'TriagePriority', 'Facility', 'Disposal', 'MonthCommence','DaysofWeek','TimeofArrival','TimeofDeparture']

resp_patients_data = resp_patients_data.drop(delete_columns, axis=1)
resp_patients_data
```

	PresentingComplaint	Diagnosis	AgeGroup	Gender
0	Back Pain	Pneumonia	65 to 70	Female
1	Unwell Adult	Pneumonia	71 to 80	Male
2	Sepsis	Pneumonia	25 to 34	Female
3	SOB - Adult	Asthma	45 to 54	Male
4	Falls	Pneumonia	81 to 90	Female
...	...	...	...	...
6769	Breathing Problems	Asthma	20 to 24	Male
6770	Asthma	Asthma	25 to 34	Male
6771	Chest pain	Asthma	55 to 64	Female
6772	Asthma	Asthma	81 to 90	Female
6773	Sepsis	Pneumonia	71 to 80	Male

6772 rows × 4 columns

Figure A.18: Python code and results for dropping columns not required

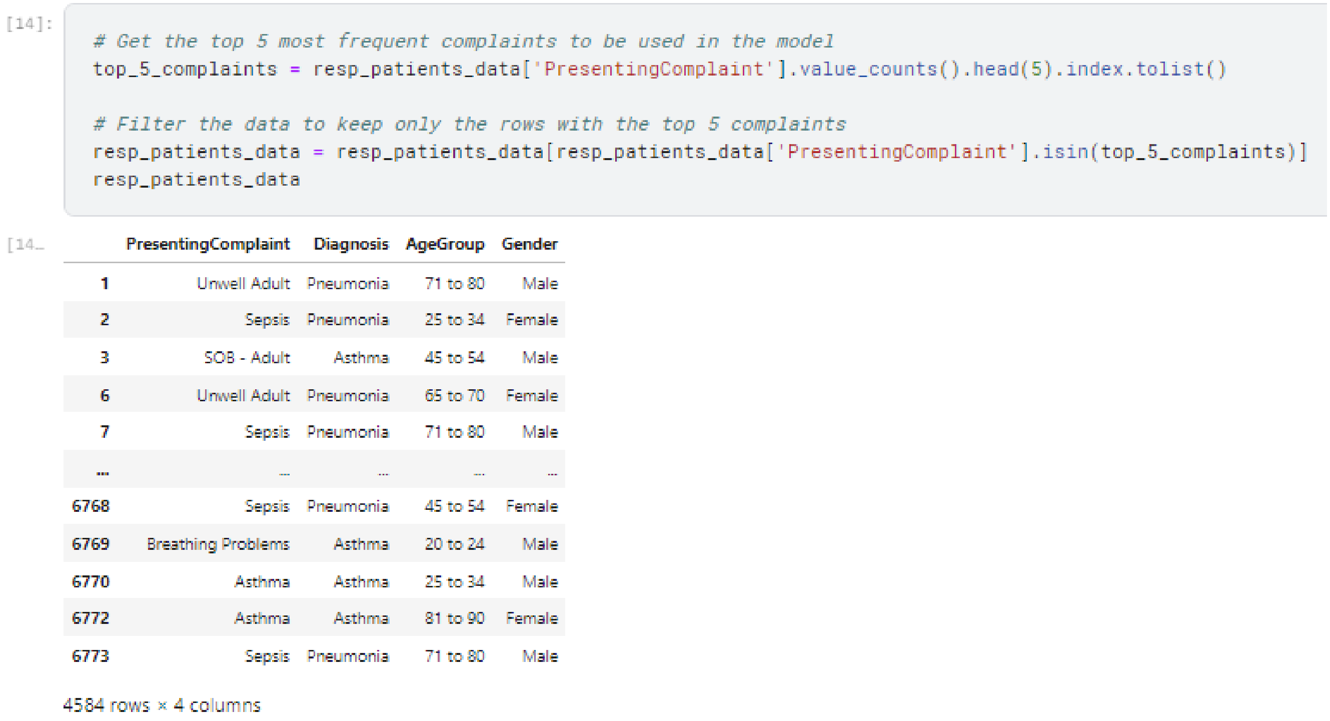


```python
# Filter the data to keep only the rows with Asthma and Pneumonia complaints
resp_patients_data = resp_patients_data[(resp_patients_data['Diagnosis'].str.contains('Asthma', case=False)) | (resp_patients_data['Diagnosis'].str.contains('Pneumonia', case=False))]
resp_patients_data
```

	PresentingComplaint	Diagnosis	AgeGroup	Gender
0	Back Pain	Pneumonia	65 to 70	Female
1	Unwell Adult	Pneumonia	71 to 80	Male
2	Sepsis	Pneumonia	25 to 34	Female
3	SOB - Adult	Asthma	45 to 54	Male
4	Falls	Pneumonia	81 to 90	Female
...	...	...	...	...
6769	Breathing Problems	Asthma	20 to 24	Male
6770	Asthma	Asthma	25 to 34	Male
6771	Chest pain	Asthma	55 to 64	Female
6772	Asthma	Asthma	81 to 90	Female
6773	Sepsis	Pneumonia	71 to 80	Male

6186 rows × 4 columns

Figure A.19: Python code and result for filtering the target column



[14]:

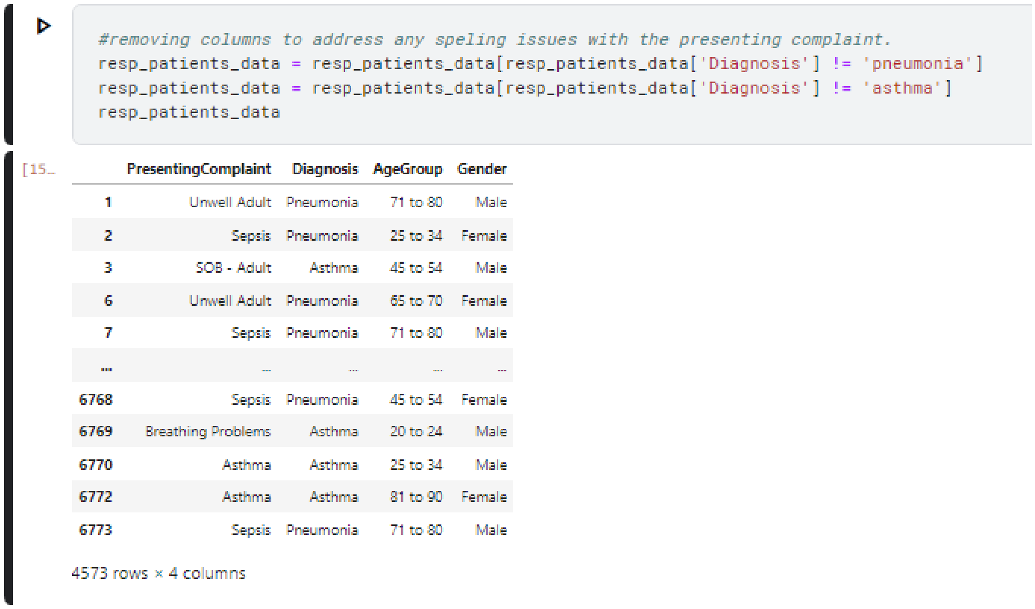
```python
# Get the top 5 most frequent complaints to be used in the model
top_5_complaints = resp_patients_data['PresentingComplaint'].value_counts().head(5).index.tolist()

# Filter the data to keep only the rows with the top 5 complaints
resp_patients_data = resp_patients_data[resp_patients_data['PresentingComplaint'].isin(top_5_complaints)]
resp_patients_data
```

[14…

	PresentingComplaint	Diagnosis	AgeGroup	Gender
1	Unwell Adult	Pneumonia	71 to 80	Male
2	Sepsis	Pneumonia	25 to 34	Female
3	SOB - Adult	Asthma	45 to 54	Male
6	Unwell Adult	Pneumonia	65 to 70	Female
7	Sepsis	Pneumonia	71 to 80	Male
...	...	...	...	...
6768	Sepsis	Pneumonia	45 to 54	Female
6769	Breathing Problems	Asthma	20 to 24	Male
6770	Asthma	Asthma	25 to 34	Male
6772	Asthma	Asthma	81 to 90	Female
6773	Sepsis	Pneumonia	71 to 80	Male

4584 rows × 4 columns

Figure A.20: Python code and result for filtering presenting complaints column



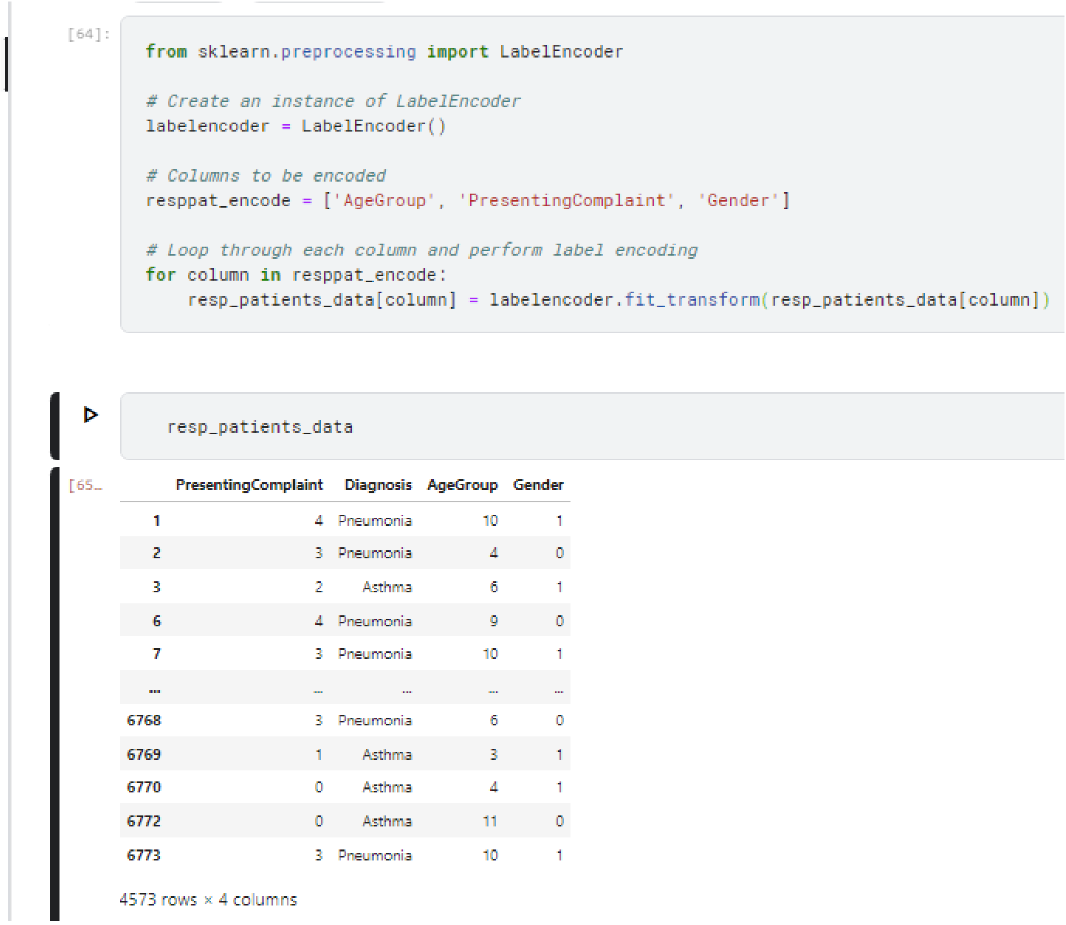
```python
#removing columns to address any speling issues with the presenting complaint.
resp_patients_data = resp_patients_data[resp_patients_data['Diagnosis'] != 'pneumonia']
resp_patients_data = resp_patients_data[resp_patients_data['Diagnosis'] != 'asthma']
resp_patients_data
```

[15…

	PresentingComplaint	Diagnosis	AgeGroup	Gender
1	Unwell Adult	Pneumonia	71 to 80	Male
2	Sepsis	Pneumonia	25 to 34	Female
3	SOB - Adult	Asthma	45 to 54	Male
6	Unwell Adult	Pneumonia	65 to 70	Female
7	Sepsis	Pneumonia	71 to 80	Male
...	...	...	...	...
6768	Sepsis	Pneumonia	45 to 54	Female
6769	Breathing Problems	Asthma	20 to 24	Male
6770	Asthma	Asthma	25 to 34	Male
6772	Asthma	Asthma	81 to 90	Female
6773	Sepsis	Pneumonia	71 to 80	Male

4573 rows × 4 columns

Figure A.21: Python code for deleting values with spelling issues

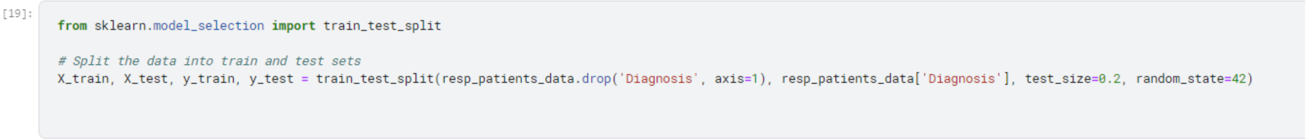


```
[64]:
from sklearn.preprocessing import LabelEncoder

# Create an instance of LabelEncoder
labelencoder = LabelEncoder()

# Columns to be encoded
resppat_encode = ['AgeGroup', 'PresentingComplaint', 'Gender']

# Loop through each column and perform label encoding
for column in resppat_encode:
    resp_patients_data[column] = labelencoder.fit_transform(resp_patients_data[column])
```

```
resp_patients_data
```

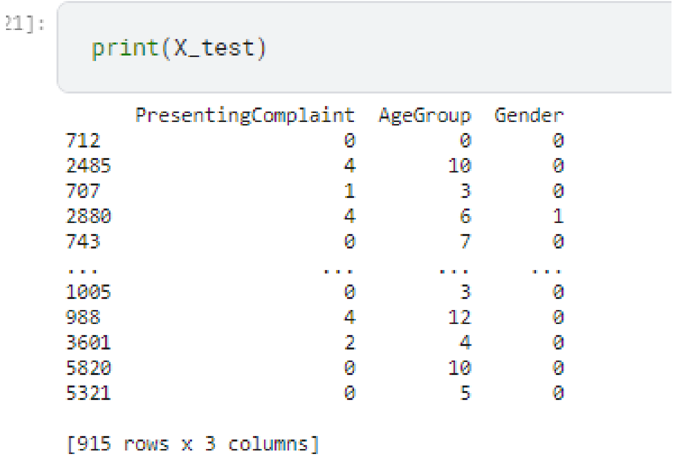
```
[65…
      PresentingComplaint  Diagnosis  AgeGroup  Gender
   1                    4  Pneumonia        10       1
   2                    3  Pneumonia         4       0
   3                    2     Asthma         6       1
   6                    4  Pneumonia         9       0
   7                    3  Pneumonia        10       1
 ...                  ...        ...       ...     ...
6768                    3  Pneumonia         6       0
6769                    1     Asthma         3       1
6770                    0     Asthma         4       1
6772                    0     Asthma        11       0
6773                    3  Pneumonia        10       1

4573 rows × 4 columns
```

Figure A.22: Python code and results for converting the features to numerical data



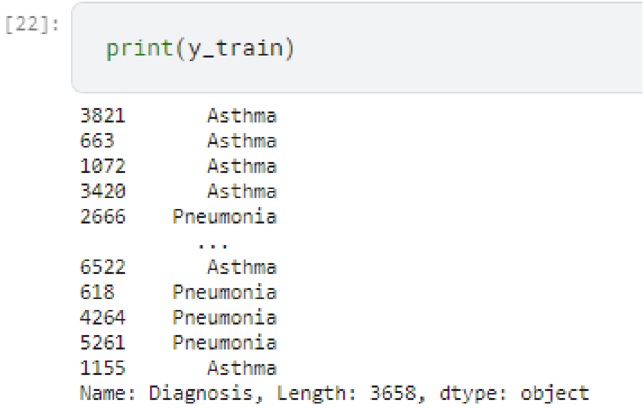
```
[19]:
from sklearn.model_selection import train_test_split

# Split the data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(resp_patients_data.drop('Diagnosis', axis=1), resp_patients_data['Diagnosis'], test_size=0.2, random_state=42)
```

Figure A.23: Python code to split the data



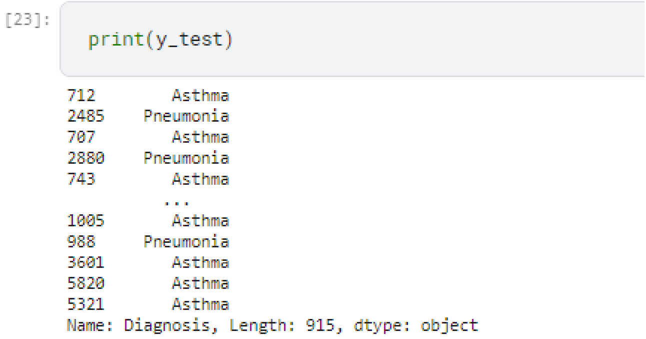
```
21]:  print(X_test)

      PresentingComplaint  AgeGroup  Gender
712                     0         0       0
2485                    4        10       0
707                     1         3       0
2880                    4         6       1
743                     0         7       0
...                   ...       ...     ...
1005                    0         3       0
988                     4        12       0
3601                    2         4       0
5820                    0        10       0
5321                    0         5       0

[915 rows x 3 columns]
```

Figure A.24: Python code to retrieve details of the feature test data



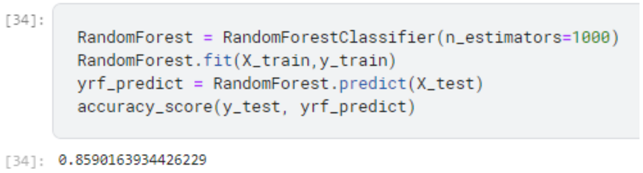
```
[22]:  print(y_train)

3821       Asthma
663        Asthma
1072       Asthma
3420       Asthma
2666    Pneumonia
          ...
6522       Asthma
618     Pneumonia
4264    Pneumonia
5261    Pneumonia
1155       Asthma
Name: Diagnosis, Length: 3658, dtype: object
```

Figure A.25: Python code to retrieve details of the target train data



```
[23]:  print(y_test)

712        Asthma
2485    Pneumonia
707        Asthma
2880    Pneumonia
743        Asthma
          ...
1005       Asthma
988     Pneumonia
3601       Asthma
5820       Asthma
5321       Asthma
Name: Diagnosis, Length: 915, dtype: object
```

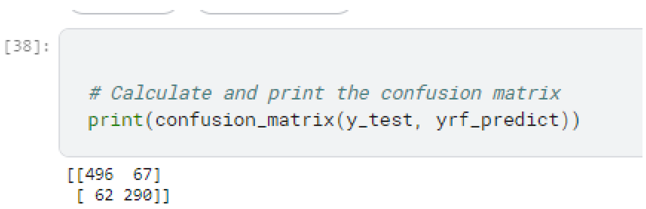
Figure A.26: Python code to retrieve details of the target test data



```
[34]:
RandomForest = RandomForestClassifier(n_estimators=1000)
RandomForest.fit(X_train,y_train)
yrf_predict = RandomForest.predict(X_test)
accuracy_score(y_test, yrf_predict)
```

```
[34]: 0.8590163934426229
```

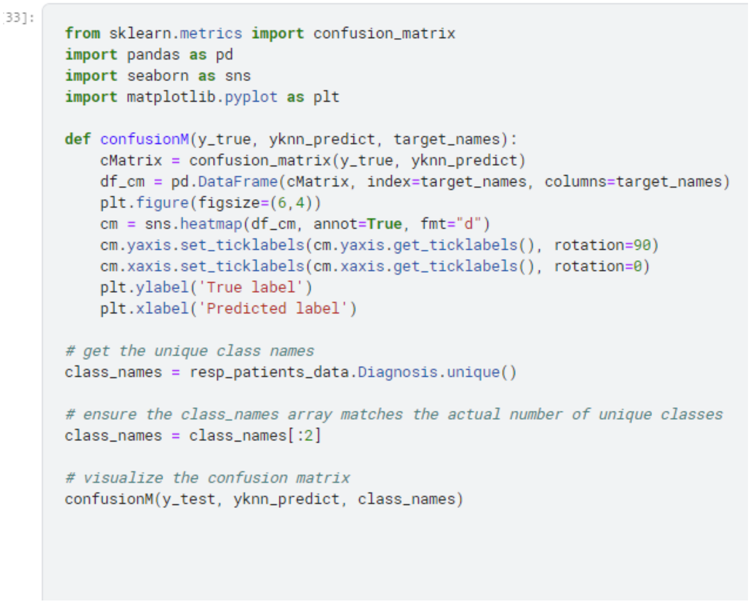
Figure A.27: Python code and result for Random forest model accuracy



```
[38]:
# Calculate and print the confusion matrix
print(confusion_matrix(y_test, yrf_predict))
```

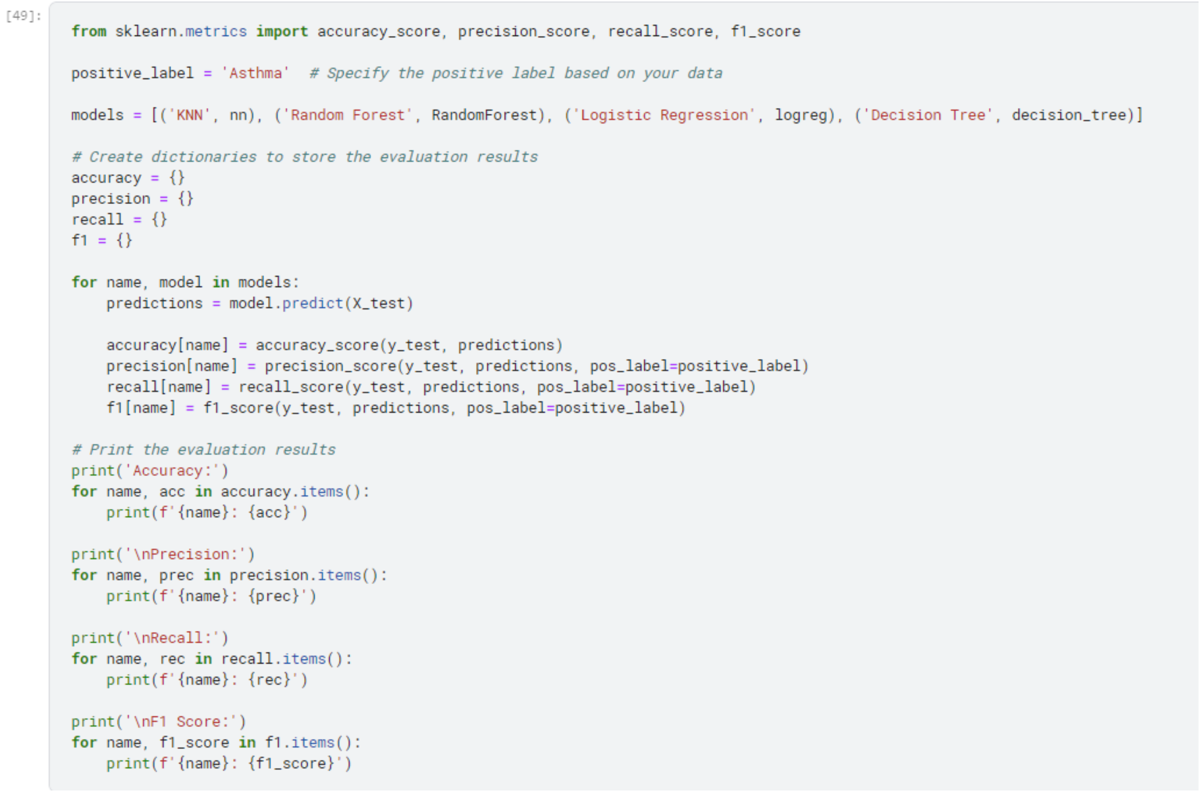
```
[[496  67]
 [ 62 290]]
```

Figure A.28: Python code and result for Random forest confusion matrix



```
33]:
from sklearn.metrics import confusion_matrix
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

def confusionM(y_true, yknn_predict, target_names):
    cMatrix = confusion_matrix(y_true, yknn_predict)
    df_cm = pd.DataFrame(cMatrix, index=target_names, columns=target_names)
    plt.figure(figsize=(6,4))
    cm = sns.heatmap(df_cm, annot=True, fmt="d")
    cm.yaxis.set_ticklabels(cm.yaxis.get_ticklabels(), rotation=90)
    cm.xaxis.set_ticklabels(cm.xaxis.get_ticklabels(), rotation=0)
    plt.ylabel('True label')
    plt.xlabel('Predicted label')

# get the unique class names
class_names = resp_patients_data.Diagnosis.unique()

# ensure the class_names array matches the actual number of unique classes
class_names = class_names[:2]

# visualize the confusion matrix
confusionM(y_test, yknn_predict, class_names)
```

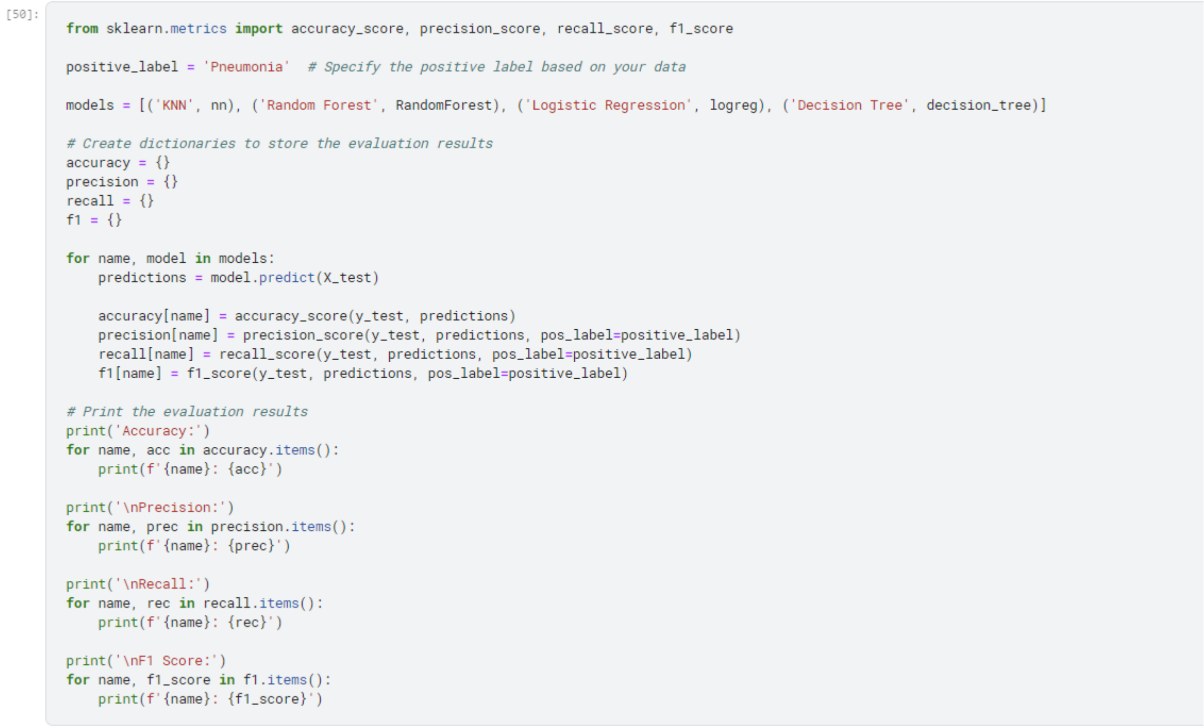
Figure A.29: Python code for generating confusion matrix diagram



[49]:

```
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

positive_label = 'Asthma'  # Specify the positive label based on your data

models = [('KNN', nn), ('Random Forest', RandomForest), ('Logistic Regression', logreg), ('Decision Tree', decision_tree)]

# Create dictionaries to store the evaluation results
accuracy = {}
precision = {}
recall = {}
f1 = {}

for name, model in models:
    predictions = model.predict(X_test)

    accuracy[name] = accuracy_score(y_test, predictions)
    precision[name] = precision_score(y_test, predictions, pos_label=positive_label)
    recall[name] = recall_score(y_test, predictions, pos_label=positive_label)
    f1[name] = f1_score(y_test, predictions, pos_label=positive_label)

# Print the evaluation results
print('Accuracy:')
for name, acc in accuracy.items():
    print(f'{name}: {acc}')

print('\nPrecision:')
for name, prec in precision.items():
    print(f'{name}: {prec}')

print('\nRecall:')
for name, rec in recall.items():
    print(f'{name}: {rec}')

print('\nF1 Score:')
for name, f1_score in f1.items():
    print(f'{name}: {f1_score}')
```

Figure A.30: Python code and results for model comparism for prediction of asthma



[50]:

```
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

positive_label = 'Pneumonia'  # Specify the positive label based on your data

models = [('KNN', nn), ('Random Forest', RandomForest), ('Logistic Regression', logreg), ('Decision Tree', decision_tree)]

# Create dictionaries to store the evaluation results
accuracy = {}
precision = {}
recall = {}
f1 = {}

for name, model in models:
    predictions = model.predict(X_test)

    accuracy[name] = accuracy_score(y_test, predictions)
    precision[name] = precision_score(y_test, predictions, pos_label=positive_label)
    recall[name] = recall_score(y_test, predictions, pos_label=positive_label)
    f1[name] = f1_score(y_test, predictions, pos_label=positive_label)

# Print the evaluation results
print('Accuracy:')
for name, acc in accuracy.items():
    print(f'{name}: {acc}')

print('\nPrecision:')
for name, prec in precision.items():
    print(f'{name}: {prec}')

print('\nRecall:')
for name, rec in recall.items():
    print(f'{name}: {rec}')

print('\nF1 Score:')
for name, f1_score in f1.items():
    print(f'{name}: {f1_score}')
```

Figure A.31: Python code and for model comparism