|  |  |  |
| --- | --- | --- |
| **crrao.png** | **C R RAO Advanced Institute of Mathematics, Statistics and Computer Science (AIMSCS) in Collaboration with Jawaharlal Nehru Technological University Hyderabad** | **71B14E20-F721-4482-B367-C06C9CF82633-300x281.png** |

Real Time Projects/Societal Needed Projects

### A dissertation submitted to the CRRAO AIMSCS in partial fulfillment of the requirements for the Completion of the Course

### Real Time Projects/Societal Needed Projects

by

#### AZHAR-22XV1M6706

#### HARSHIT-22XV1M6709

#### HARDHIK-22XV1M6717

#### P.VAISHNAVI-22XV1M6723

#### YASHASWINI-22XV1M6726

# Projects outside CRRAO

Supervisor: Dr.G.padmavathi

Mentor: for internship mentor name

# Projects with in CRRAO

Date of Submission: 19th July 2024

CERTIFICATE

This is to certify that the dissertation entitled ----------------------submitted by **Team,** bearing Roll nos **,** in partial fulfillment of the requirements for the award of Master of Technology in Information Security, is a bonafide work carried out by him under my supervision and guidance.

The dissertation has not been submitted previously in part or in full to this or any other University or Institution for the award of any degree or diploma.

Dr.G. Padmavathi

B.Tech Coordinator

CRRAO,AIMSCS

## Declaration

I,\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_here by declare that this dissertation entitled \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ submitted by me under the supervision of Dr.G.Padmavathi and Mentored by \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_. I also declare that it has not been submitted previously in part or in full to this University or any other University or Institution for the award of any degree or diploma or Projects

A report of plagiarism statistics from the college is enclosed.

Date:19/07/2024

Place:

#### AZHAR-22XV1M6706

#### HARSHIT-22XV1M6709

#### HARDHIK-22XV1M6717

#### P.VAISHNAVI-22XV1M6723

#### YASHASWINI-22XV1M6726

(Dr.G.Padmavathi)

Supervisor& HOD

### Acknowledgement

It gives me utmost pleasure to express my profound sense of gratitude and sincere thanks to Dr.G. Padmavathi and Mentor name\_\_\_\_\_\_\_\_\_\_\_ for her continuous guidance, support, and encouragement in carrying out this work.

I would like to express my sincere regards to my family and friends for their continuous support.

With Sincere Regards, Team Members:

AZHAR

HARSHIT

HARDHIK

P.VAISHNAVI

YASHASWINI

# Table of Contents

|  |
| --- |
| CHAPTER-1 : INTRODUCTION Devops |
| CHAPTER-1.1 : INTRODUCTION Diabetes Dataset |
| CHAPTER-1.2 : MOTIVATION |
| CHAPTER-2 : LITERATURE SURVEY |
| CHAPTER-2.1 : INTRODUCTION TO DATASETS |
| CHAPTER-3 : GOAL AND OBJECTIVES |
| CHAPTER-3.1 : OBJECTIVE-1 |
| CHAPTER-4 : ALGORITHMS |
| CHAPTER-4.1 : OBJECTIVE-2 |
| CHAPTER-5 : CONCLUSION AND FUTURE WORK |
| CHAPTER-6 : INTERNSHIP |
| REFERENCE & LINKS |

**List of Figures**

Figure 1 : Analysis types on a dataset

Figure 2 : Aspects of Covid-19 dataset

Figure 3 : Aspects of the ADNI dataset

Figure 4 : Various aspects of The Cardiovascular Diseases dataset

Figure 5 : Pictorial overview of a melanoma dataset

Figure 6 : Overview of Nationwide Inpatient Sample dataset

Figure 7 : Insights on choosing to analyze the 10 Years Diabetes Dataset over other datasets

Figure 8 : Data Preprocessing

Figure 9 : Key concepts of Random Forest Algorithm

Figure 10 : Key concepts of SVM Algorithm

Figure 11: Key concepts of Logistic Regression Algorithm

Figure 12: Table for accuracy Values of Different Datasets

**Chapter-1**

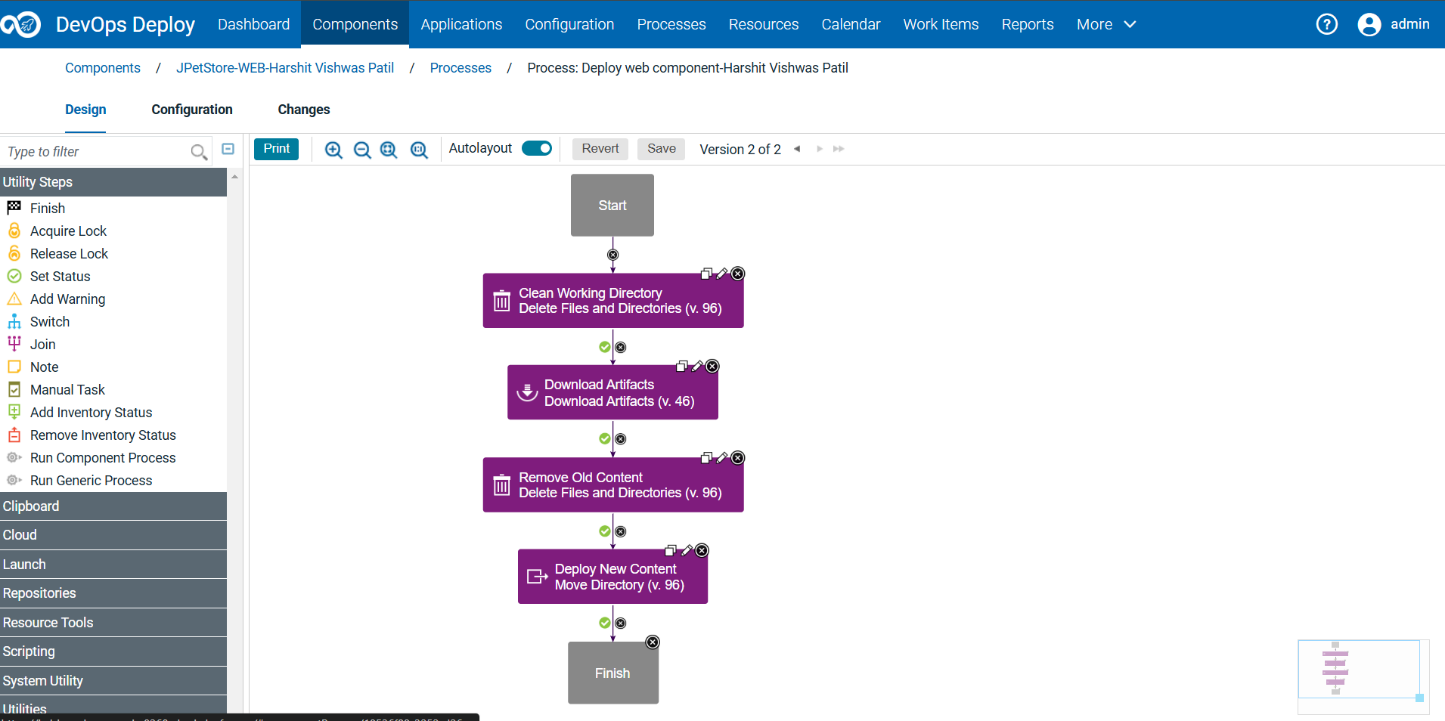
## **INTRODUCTION:**

**Devops-**

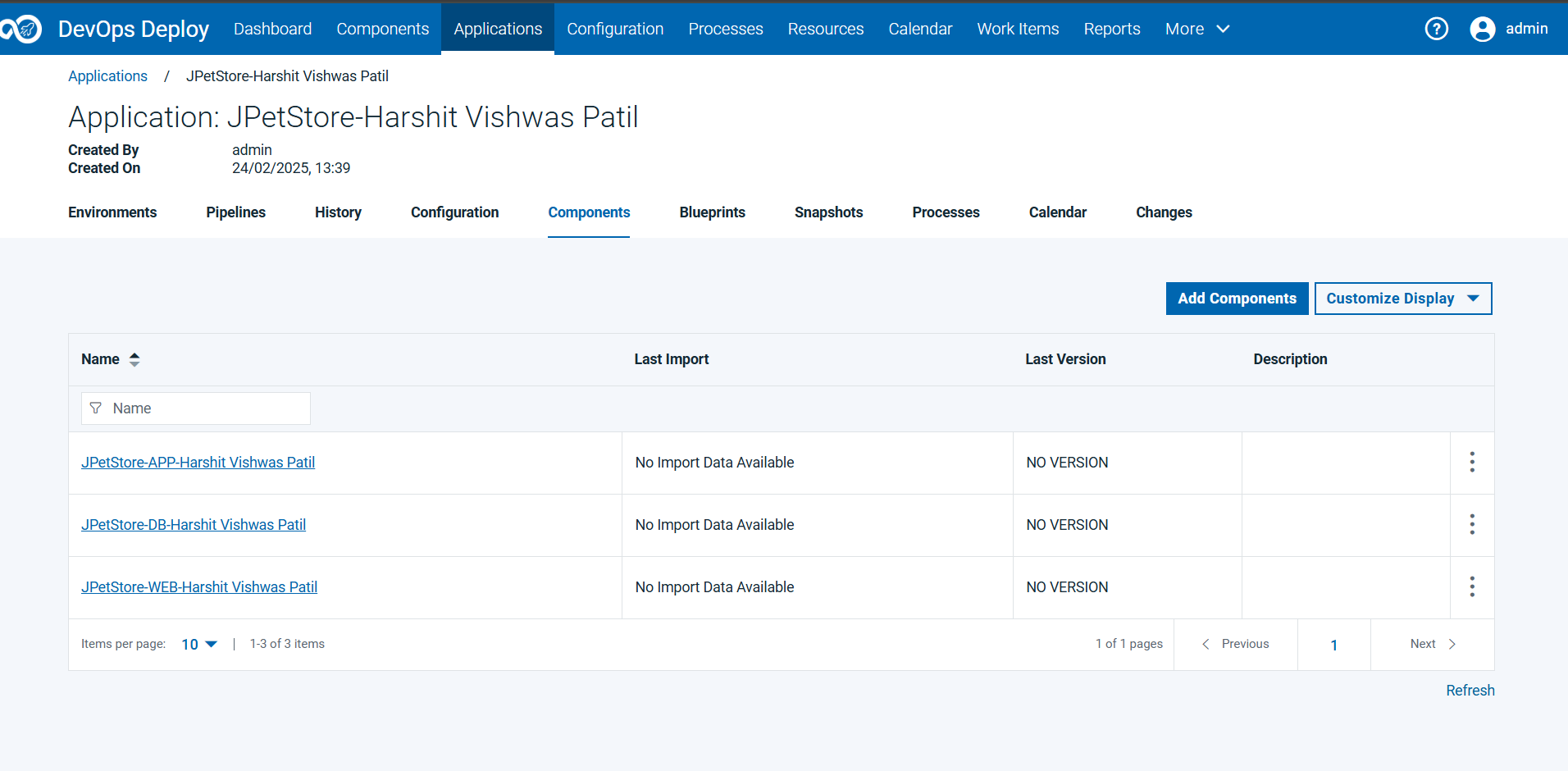
We recently Completed an 8-week DevOps program – "Project Sunshine\_AI in DevOps Using HCL DevOps Deploy."

Submitted an assessment on:

* Configuring components and component processes.
* Creating components.
* Creating component processes.
* Creating an application.



Component process



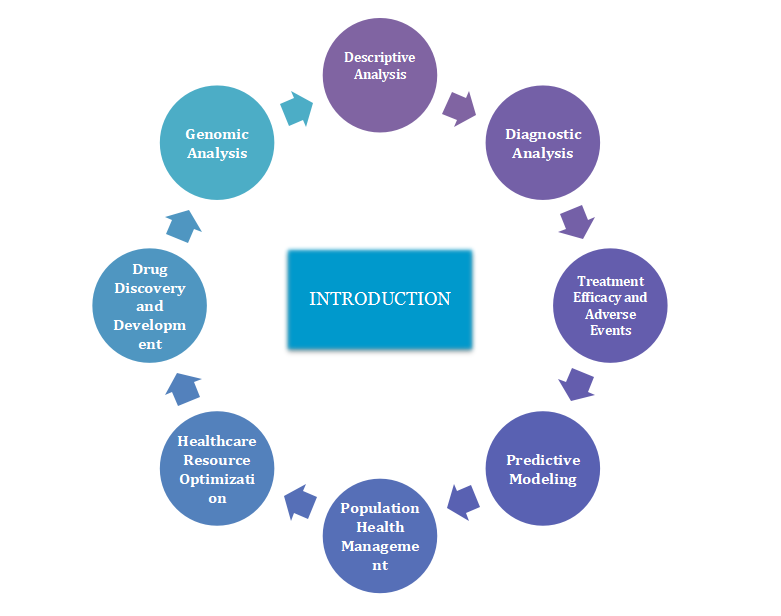
Creating Application

**Chapter-1.1**

## **INTRODUCTION:**

It has huge series data of patient demographics, clinical observations, diagnostic tests, treatment plans, and outcomes, among many others. Such indispensable resources make studying trends, patterns, and insights for the improvement of patient care, advancement of medical research, and healthcare systems optimization easier to achieve for researchers, data scientists, and healthcare practitioners.

Figure 1: Analysis types on a dataset



**❖Analysing Traits Descriptively:** This basic research describes to identifiable features within a patient group that have to be known to understand healthcare needs and create actions. For example, geospatial data might give out the variation of geography-based access to health facilities; at the same time, demographic features like age distribution may help in recognizing health problems unique to a particular age group.

* **﻿Diagnostic exams:** Using diagnostic information, researchers can develop algorithms designed to facilitate analysis and type. One instance is while system getting to know fashions trained with scientific photo statistics assist radiologists identify anomalies indicative of cancer or neurological sickness In the same vein, early detection and threat evaluation may be improved through diagnostic biomarkers because of lab take a look at results**.**
* **Effectiveness of Treatment:** One can consider treatment outcomes along with adverse events if one is a fitness care provider interested in assessing the effectiveness and safety of treatment strategies. Comparative-effectiveness research includes the evaluation of the results of opportunity treatments as a way to guide proof-based totally remedy alternatives. Additionally, postmarket drug protection is controlled through drug monitoring efforts that use actual-global data to discover and examine drug abuse
* **Predictive Modeling**: It makes use of historical data in predicting future events or results. Such insights are very helpful in terms of proactive intervention and chance management. One example of predictive analytics could be the identification of sufferers at high danger of hospital readmission or complication, thus allowing healthcare suppliers to take preventive measures or alter remedy plans. Furthermore, predictive methods can help personalized medicine initiatives by personalizing remedy methods to a particular person traits and threat profiles.
* **Population Health Management**: Cluster of population-level analyses provides a microscopic view of health trends and diversity within a network or population institution. Epidemiological researches make use of the medical data sets in order to track disease incidence, detect risk factors, and monitor the relative impact of public health interventions. This information is used by population health management strategies in the design of focused interventions aimed at improving health.
* **Optimization of Healthcare Resources:** Information from the investigation of styles in healthcare use will help organizations draw out inefficiencies and cleanse workflows to optimize aid allocation. Using data-driven insights that permit strategic allocation of assets in a way that meets affected person wishes, health systems may be capable of assignment demand for healthcare services and optimize staffing degrees.Beyond that, the analytics in healthcare discover opportunities to comprise value and decrease waste to make sure standard efficiency and sustainability in healthcare shipping.
* **Drug Discovery and Development:** Medical datasets come in handy in accelerating drug discovery and development by performing target identification, compound screening, and clinical trial designs. Bioinformatics analyses bring together genomic, proteomic, and metabolomic data to understand the mechanism of the disease, identify probable targets that the drug can act on, and prioritize some candidate compounds for further investigation. The analysis of clinical trial data assures the safety and efficacy of investigative drugs and thus helps in decision-making at each step of drug development.
* **Genomic Analysis:** Genomic data analysis has huge potential for personalized medicine, in which it can help treatment decisions based on the genetic profile of the patient. Genomic data can be integrated into clinical information about a patient for the detection of genetic variation, predisposition to diseases, response to treatment, and metabolism of drugs. Pharmacogenomic analysis drives medication selection and dosing in the light of patient genotype to maximize therapeutic efficacy while avoiding adverse drug reactions.  
    
  Such diverse analyses improve, in an aggregate way, medical knowledge, patient outcomes, and healthcare delivery across the continuum of care. Stakeholders can activate the potential of medical data to extract actionable insights that help address pressing healthcare challenges and drive innovation in clinical practice and research.There are many Healthcare datasets available for analysis , some are:

Here are some disease-specific datasets:

1. ﻿**COVID-19 datasets**:

Stacks of datasets immediately and not directly connected to the COVID-19 pandemic, like case counts, vaccination prices, and genomic sequences, are to be had through sources just like the WHO, CDC, and Johns Hopkins University.Johns Hopkins COVID-19 Data Repository: [[Github Repository link]](https://github.com/CSSEGISandData/COVID-19) .This dataset includes a big extent of information on the pandemic, especially on day by day cases, deaths, recovered instances, and vaccination costs from all over the international. Data is organized at a country level and from time to time at a local level within countries. It has had huge usage amongst researchers, decision-makers, and the general public in tracking the development of the pandemic.

Figure 2 : Aspects of Covid-19 dataset

1.**Case Counts**: This dataset contains daily case counts of confirmed cases, death, and recovery cases of COVID-19. On the whole, these case counts are normally reported at the country or regional level and are aggregated from many sources, including official health departments, hospitals, and research institutions.

﻿2. **Vaccination rates**: It also presents info on vaccination costs in phrases of the range of administered doses, completely vaccinated people, and the share proportion of the population this is vaccinated towards COVID-19. This statistics will significantly assist in tracing how vaccination campaigns are going for walks and reviewing their effect on the spread of this virus.

3. **Geographic Coverage**: The dataset has data for countries and regions from all parts of the world. This gives a worldwide dimension to the pandemic. In that way, one would look into trends and responses across different countries and regions.

4. **Time Series Data**: Probably one of the most interesting features of this dataset is that the data reported is time series in nature, with some extended time instances being accounted for on a daily basis. This will enable trend analysis over time and the identification of peaks, troughs, and seasonal patterns in case counts and vaccination rates.

5. **Data Quality and Sources**: This dataset assures that a very high level of requirement regarding the quality of the statistics to be delivered accurately and fresh, within the proposed counts. It identifies sources that shall be used, such as WHO—World Health Organization, countrywide fitness departments, and other credible sources.

6. **Data Format**: The dataset is generally to be had in a based format, such as a CSV file, with every row representing a particular date and region (e.G., country or region) and columns representing distinct metrics (e.G., confirmed cases, deaths, vaccinations).

7. **Data Visualization**: This dataset will very often be used to construct visualizations—including, but not limited to, line graphs, bar charts, maps, and so forth—to understand the spread of the virus and the effect of vaccination campaigns, among other trends that emerge.

8. **Data Analysis**: This dataset enables a number of varying analyses, such as epidemiological modeling, trend analysis, and forecasting, all of which help inform broader public health policies and certain other interventions put in place to control the virus.

9. **Global Collaboration**: The dataset becomes part of the global effort to trace and act on COVID-19. Much of the time, that helps with other datasets and research efforts to provide an overall understanding of the impact of the pandemic.

The COVID-19 dataset serves as a very valuable tool for researchers, policy decision-makers, and the general public in the tracing of progress in the pandemic, comprehension of its impact, and informing effective responses to mitigate its spread.

**2. ADNI[Alzheimer's Disease Neuroimaging Initiative]:**

A dataset about Alzheimer's disease, including neuroimaging data, genetic data, and other features such as cognitive assessments.  
ADNI: This Alzheimer's Disease Neuroimaging Initiative dataset is a long-term, multi-modal database with deep coverage of subjects having AD and related neurodegenerative diseases. It is a multicenter project involving academic research centers, pharmaceutical companies, and government agencies in an effort to expedite the discovery of biomarkers for the early identification and monitoring of AD progression. The ADNI database houses all sorts of data modalities: neuroimaging, genetic information, clinical assessments, data on cognitive functions, and, last but not least, biospecimens—a rich resource for those studying different aspects of the pathophysiology, diagnosis, and treatment of AD.

[ADNI](https://adni.loni.usc.edu/)

Here's a detailed exploration of each aspect of the ADNI dataset:

Figure 3 : Aspects of the ADNI dataset

1. **Neuroimaging Data**: Volumetric structural magnetic resonance imaging, purposeful magnetic resonance imaging, and positron emission tomography image data are available for AD patients, human beings with mild cognitive impairment, and healthy controls in ADNI. Functional magnetic resonance imaging measures interest and connectivity patterns within the thoughts that supply perception into changes in neural networks underpinning cognitive decline. Imaging with positron emission tomography and radiotracers targeting amyloid-β and tau protein accumulations lets in the in vivo detection of AD pathological hallmarks and is beneficial in disease evaluation and monitoring of its development. Imaging with positron emission tomography and radiotracers targeting amyloid-β and tau protein accumulations allows the in vivo detection of AD pathological hallmarks and is beneficial in disorder evaluation and tracking of its development.

2. **Genetic Data**: Genetic data is obtained for each subject; the data gathered include genotyping and whole-genome sequencing. By leveraging genetic studies, the ADNI looks into genetic risk factors that further susceptibility to AD, such as apolipoprotein E genotype and other genetic variants from GWAS that affect susceptibility and disease progression of AD and, recently, treatment response. Taken together, several genes in novel AD risk loci are identified using genome-wide association studies, after which polygenic risk scores are derived by aggregating several genetic variants in order to predict a person's risk of developing AD or any other kind of related dementia.

3. **Clinical Assessments**: ADNI includes a set of medical tests and neuropsychological checks performed at normal intervals in members. The checks have included measures for cognitive feature, memory, executive feature, language skills, and activities of every day residing. Standardized cognitive assessment devices, like the Mini-Mental State Examination and Alzheimer's Disease Assessment Scale–Cognitive Subscale, permit for assessment of trade in topics' cognitive overall performance over time and calculation of ailment severity. It is further characterized by demographic data, medical comorbidities, and drug use, which are obtained from clinical interviews and questionnaires on the medical history of the participants.

4. **Cognitive Function**: In ADNI, detailed data are available for the participants' cognitive function, assessed by a large variety of psychometric tests and informant and clinician reports. Longitudinal assessments quantify decline in the domains of cognition typically affected by AD. Summary composite scores of performance on several standardized cognitive tests emphasize global cognitive performance and serve as outcome measures in clinical trials testing investigational treatments of possible AD.

5. **Biospecimens**: ADNI collects blood, CSF, and a number of organic samples from topics to be used for evaluation of biomarkers. It'd be the molecular markers of AD pathology, collectively with the degrees of amyloid-beta and tau protein in the CSF and blood-based mostly markers pointing towards neuroinflammation, synaptic dysfunction, and neuronal hurt.Biomarker profiles from these analyses of biospecimens contribute to the early detection of disease, differential diagnosis, and monitoring response to treatment in the clinical trials of AD.

6. **Longitudinal Follow-up**: This is one of the greatest strengths of ADNI: the longitudinal design that allows for the continued monitoring of subjects over several years. In this respect, longitudinal data are collected to capture changes in the disease course, cognitive decline, and biomarker profiles over time, thus allowing the description of the natural history of AD and identification of predictors of disease course. Longitudinal analyses within ADNI identify trajectories of cognitive decline, pinpoint biomarkers of disease progression, and assess the efficacy of interventions that are putative to slow or halt the progress of AD.  
  
A dataset such as ADNI is therefore an invaluable resource for multidisciplinary research into AD and related neurogenerative disorders in general. It will enrich current understanding of the mechanisms of disease and the development of novel diagnostic tools and therapeutic targets. Integrate neuroimaging, genetic, clinical, cognitive, and biomarker data through ADNI to foster collaborative research designed to accelerate the development of effective treatments of AD and improvement in patient outcomes.

**3. Cardiovascular disease diagnosis:**

Cardiovascular disease datasets are very instrumental in understanding and overcoming a challenge to public health like cardiovascular diseases, which remain a leading cause of mortality globally.   
[<https://github.com/caravanuden/cardio>]  
  
The Cardiovascular Diseases dataset: This is an open-source dataset shared on Kaggle, which contains 70,000 records—34,979 show patients who suffer from cardiovascular disease, and 35,021 do not show cardiovascular diseases.﻿It has eleven capabilities structured into demographic, examination, and social records functions with 4 variables every except the social records, which has most effective 3 variables.  
  
The dimensions variously describe the subsequent in precise:

Figure 4 : Various aspects of The Cardiovascular Diseases dataset.

1. **Demographic Information**: Information about the people that make up your statistics set, as an instance, age, sex, ethnicity/race. Demographic statistics is useful to understand trends and hazard elements for precise populations.  
  
﻿2. **Clinical Measurements**: Some of the essential clinical measurements have to do with key health metrics, blood stress, levels of cholesterol, and blood glucose tiers which move further to determine body mass index. These metrics offer insights into cardiovascular risk elements and sickness improvement.  
  
﻿3. **Medical History**: It includes facts about the affected person's past cardiovascular occasions, associated ailments—which include high blood pressure and diabetes—and present day medicines. It publications evaluation of the recognition and dreams of the character for treatment.  
  
4. **Diagnostic Tests and Procedures**: This accommodates the results from diagnostic tests and strategies, consisting of electrocardiograms, echocardiography, and cardiac catheterization, and it assists in arriving at the diagnoses of the cardiovascular conditions and estimating their severity.  
  
5. **Outcomes and Follow-up**: Outcomes data include incidence of cardiovascular events—e.g., myocardial infarction or strokes—and mortality rates, and longitudinal data for following up on changes over time. The results of some interventions or treatments need to be known due to the evaluation effectiveness.  
  
6. **Lifestyle Factors**: Dietary habits, alcohol intake, and stressful life are significant contributory factors to cardiovascular health related to lifestyle factors. Monitoring these factors allows for the identification of modifiable risk factors and formulates appropriate preventive strategies.  
  
7. **Genetic and Biomarker Data**: Genetic variants and biomarkers offer an understanding of the genetic predispositions and biochemical processes associated with cardiovascular diseases. Genetic and biomarker data can be integrated to enhance the prediction of risk and develop personalized modes of treatment.  
Some of them are numerical features; some are categorical, assigned by codes, while others are binary. The classes are balanced, but there were more female than male patients. Besides, continuous-valued features are almost normally distributed, and most categorical-valued features are skewed towards "normal" as opposed to "high" levels of potentially pathological features.

**4. The Cancer Genome Atlas (TCGA):** This is a large dataset of genomic and clinical data on many types of cancer. Concisely, multi-dimensional genomic data in the TCGA dataset consists of DNA sequencing, gene expression profiling, DNA copy number variations, and epigenetic modifications across many types of cancers.  
  
TCGA: [The cancer genome atlas](https://www.cancer.gov/ccg/research/genome-sequencing/tcga)  
The Cancer Genome Atlas Program. NCI

Figure 5 : Pictorial overview of a melanoma dataset  
  
﻿Thus,that is a precise of statistics contained in the melanoma dataset:

1) **Information of sufferers**:  
•Demographic information: statistics approximately age, sex, ethnicity, or every other characteristic of relevance.  
• Family History: Details of family history of cancer or various skin cancers;  
• Risk Factors: Information should be shared about risk factors, including sun exposure, a history of severe sunburns, tanning beds, presence of atypical moles.

2) **Clinical Information**  
• Tumour Characteristics: Melanoma tumor characteristics : anatomic site of melanoma, size, thickness in millimeter, Breslow thickness, and ulceration status  
• Tumor Stage: Melanoma staging AJCC staging system.  
• Histopathological Features: Information on histological subtypes of melanoma and other histopathological features observed in biopsy specimens.  
• Presence of Metastasis: Information on whether or not melanoma has invaded regional lymph nodes or other distant organs.

3) **Diagnostic Procedures**:  
• Biopsy Results: Findings from skin biopsy procedures that were done to confirm the diagnosis of melanoma and characterize it.  
﻿Imaging studies: Results of ultrasound, MRI, CT, or PET experiment imaging to investigate the amount of disorder and metastases.

4)**Treatments**:  
•Surgical interventions: Information about surgical interventions for cancer, including wide local excision, lymph node dissection, and sentinel lymph node biopsy.  
•Adjuvant Therapy: Data on adjuvant chemotherapy, immunotherapy, targeted therapy, or radiation therapy given post-operatively.  
Systemic Therapy: Treatment given in the setting of metastatic melanoma, involving systemic therapies like immunotherapy or targeted therapy, including inhibitors of the checkpoints and BRAF.

5) **Treatment Response**:  
• Treatment Response: Whether or not patients showed regression, stabilization, or progression to the different treatments administered.  
• Adverse Events: The adverse events during the treatment process that the patients have gone through.

6) **Outcome Variables**  
• Overall Survival: time from diagnosis, or the start of treatment to death from any cause.  
• Progression-Free Survival: The duration from diagnosis or the beginning of treatment up to when the first signs of relapse or death are manifested..  
• ﻿Recurrence: The information regarding recurrence of cancer submit-preliminary treatment.  
This furnished dataset might be beneficial for a big range of packages, including medical research, growing fashions of affected person final results predictions, determining prognostic elements, measuring the effectiveness of remedies, and accomplishing higher medical control strategies in patients affected by cancer.

**5. Nationwide Inpatient Sample:**

This is a data set of statistics regarding in-patient stay at a U.S. Health center for the diagnosis and procedures of various diseases.  
NIS [<https://www.hcup-us.ahrq.gov/nisoverview.jsp>]  
  
The National Inpatient Sample provides a great platform for looking into a number of problems associated with in-patient care throughout the U.S.Its breadth and versatility make it an important resource from which one can draw for knowledge, healthcare policies, and evidence-based intervention strategies.

Figure 6 :Overview of Nationwide Inpatient Sample dataset

**1**. **Scope and Coverage**: The National (Nationwide) Inpatient Sample is one of the core healthcare databases, offering an all-inclusive representation of hospital inpatient stays in the United States. Broad coverage includes most varieties of hospitals, thereby giving a representative sample of inpatient care across the country. Because of this broad scope, it allows for overarching trend analyses but still retains the ability to examine localized patterns and disparities in health care. Such care nuances can be captured across different regions, types of hospitals, and a variety of patient populations by using data from the NIS. By including large academic medical centers and small community hospitals, the NIS samples a wide patient population. As such, it represents all the complexity and diversity of health care delivery in the United States.  
  
**2. Data content**: Detailed diagnoses, procedures, patient demographics, hospital characteristics—Researchers will find broad coverage of information within the NIS.  
The richness of these data contents enables sophisticated analyses of health care utilization, outcomes, and disparities. One is allowed to undertake in-depth research on the kind of primary and secondary diagnoses patients are given and procedures conducted on them during their stay in hospitals. In view of such details, research can be focused on nuanced analysis whereby one is not simply restricted to the knowledge of prevalence of any condition, but also the kind of intervention used in dealing with them. It also contains patient demographics and hospital characteristics that provide contextual information to be used in the analyses being performed on the dataset. Therefore, because of the comprehensiveness of the data content within the NIS, researchers are better positioned to investigate the very multifaceted dimensions of inpatient care and healthcare delivery.  
  
**3. Diagnoses and Procedures**: Within the dataset are both primary and secondary diagnoses coded in ICD-CM, along with procedures that are coded in HCPCS or ICD-CM procedure codes. This kind of granularity offers the ability for in-depth research into the prevalence of diseases and treatment trends. These data will allow studies of trends in diagnosis and procedures over time, variation in clinical practice between different regions or types of hospitals, and the effectiveness of certain interventions on patient outcomes. In addition, detailed coding information is provided so that cross-referencing with other datasets and clinical registries can be done by researchers to seek validation and for conducting robust sensitivity analyses. Rich diagnostic and procedural data available in the NIS will be very important because they will help researchers come up with important insights into the complications of healthcare delivery, thus bringing forth evidence-based practices.  
  
**4. Patient Demographics**: This includes demographic information about patients, such as their age, gender, race, insurance status, among others, which gives insight into the characteristics of the patients that have undergone inpatient care. This will afford researchers an opportunity to look into the access disparities to healthcare and healthcare outcomes across different population groups. The demographic trends within the NIS can also be assessed with a view to determine which populations are vulnerable, either to facing barriers in access to healthcare services or bearing disparities in quality of care.It also helps in estimating the impact of social determinants of health on patient outcomes and patterns of healthcare use with the demographic data. This demographic understanding of the patient profile from the NIS sets interventions and policies in such a way that it really works to improve the needs of various populations of patients, thereby ensuring the reducement of inequity in improving health outcomes for all.  
  
**5. Hospital Characteristics**: Information on length of hospital, region, teaching status, and type of possession contributes a few institutional context to the versions attended upon in fitness care transport and outcomes. In this regard, one is in a position to analyze how the numerous traits of hospitals effect a huge range of phenomena associated with the outcomes of patients, styles of healthcare use, and aid allocation choices. Other studies have used the NIS to assess whether the teaching status of hospitals influences patient outcomes, whether access to specialized care was disproportionately low in rural hospitals, and whether there were differences in treatment practices related to hospital nonprofit and for-profit status. Then, accounting for hospital-level characteristics puts researchers in an improved position to interpret how patient-level variables influence outcomes and control for factors that otherwise might either confound or modify these relationships. Only then will this holistic approach add robustness and generalizability to findings derived from NIS and empower researchers to make meaningful conclusions about this complex interaction between healthcare delivery and institutional context.  
  
﻿**6. Temporal Trends**: The annual replace of the NIS will permit one to have a look at the temporal traits in in-patient care. Longitudinal analyses deliver insights into changes in exercise patterns and as a result ought to aid destiny policymaking. Trends in diagnoses, strategies, affected person demography, and results are tracked over the years to perceive rising changes, estimate the effectiveness of coverage interventions or first-rate development tasks, and project future needs or worrying situations in healthcare. For instance, this NIS has been applied in measuring the upward thrust in persistent situations, shifting traits in remedy styles with the introduction of new treatment options or guideline adjustments, and fluctuating charges for fitness care use related to public health emergencies or economic downturns. The NIS is longitudinal in nature and provides the possibility to perform strong trend analyses that get at the dynamic nature of healthcare transport and inform proof-based totally strategies for improving patient care and populace fitness.  
  
**7. Research Applications**: Outcome research, fitness disparities, cost-effectiveness studies, and epidemiologic investigations are a number of the topics investigated the usage of the NIS. Its versatility offers great potential for improvements in healthcare delivery. Researchers have drawn on the NIS for studies on the racial and ethnic disparities in access to care and outcomes, effectiveness of interventions targeted at preventing health care-associated infections or readmissions, and estimation of the economic burden of certain diseases or health care policies. Furthermore, it has been on the forefront in driving the progress made toward better understanding of rare diseases, capturing real-world treatment patterns and outcomes, and thus informing clinical practice guidelines and policy decisions. Based on the breadth and depth of data in the NIS, it is possible for researchers to come up with actionable insights that mean a great deal of improvement to patients, quality of health care, and health outcomes at a population level.

**Chapter-1.2**

**Motivation**

* Analyzing the 10 Years Diabetes Dataset over other datasets:

1. COVID-19 Datasets
2. Alzheimer's Disease Neuroimaging Initiative (ADNI)
3. Cardiovascular Disease dataset
4. The Cancer Genome Atlas (TCGA)
5. National (Nationwide) Inpatient Sample (NIS)

While each dataset offers unique insights and research opportunities,

choosing to analyze the 10 Years Diabetes Dataset over other

datasets has been aligned with our research objectives, expertise, and

the specific insights we aim to gain from the analysis.

Figure 7: Insights on choosing to analyze the 10 Years Diabetes Dataset over other datasets

**Chapter-2**

**Literature Survey**

* **Paper-1**:[Neha Prerna Tigga , Shruti Garg]

Prediction of Type 2 Diabetes using Machine Learning Classification Methods

* **Paper-2**:[Wee, B.F., Sivakumar, S., Lim, K.H. et al.]

Diabetes detection based on machine learning and deep learning approaches

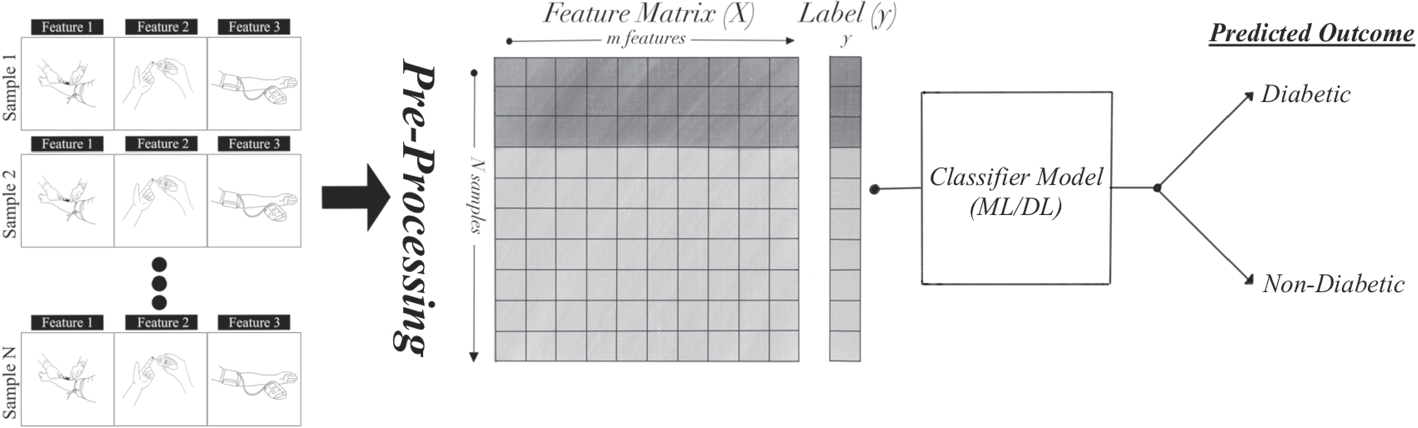


Figure 8 : Data Preprocessing

**Chapter-2.1**

**INTRODUCTION TO DATASETS**

Diabetes mellitus remains a significant global health challenge, with its prevalence steadily increasing across diverse populations. Early detection and effective management are crucial in mitigating its debilitating complications, including cardiovascular diseases, kidney failure, and neuropathy. In response to these challenges, this project explores the application of machine learning algorithms on diverse datasets to enhance diabetes prediction and diagnosis.The project leverages three distinct datasets: the Pima Indians Diabetes Dataset (PIDD), the National Health and Nutrition Examination Survey (NHANES), and the Diabetes 130-US Hospitals for Years 1999-2008 dataset. Each dataset offers unique insights into diabetes-related factors, ranging from genetic predispositions and lifestyle habits to clinical outcomes across a broad demographic spectrum.

**Dataset-1: Pima Indians Diabetes Dataset (PIDD)**

The PIDD dataset provides valuable insights into diabetes among Pima Indians, a population with a high prevalence of the disease. This dataset includes features such as glucose levels, insulin levels, BMI, and age, crucial for understanding the onset and progression of diabetes in this community. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

**Content**

The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

**Dataset-2: National Health and Nutrition Examination Survey (NHANES)**

NHANES dataset offers a comprehensive view of diabetes and related health metrics across the United States population. It encompasses a wide range of variables including demographic data, dietary habits, physical activity levels, and clinical measurements, making it ideal for exploring broader epidemiological trends in diabetes.

The [National Health and Nutrition Examination Survey (NHANES)](https://www.cdc.gov/Nchs/Nhanes/about_nhanes.htm) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations. NHANES is a major program of the National Center for Health Statistics (NCHS). NCHS is part of the Centers for Disease Control and Prevention (CDC) and has the responsibility for producing vital and health statistics for the Nation.The NHANES program began in the early 1960s and has been conducted as a series of surveys focusing on different population groups or health topics. In 1999, the survey became a continuous program that has a changing focus on a variety of health and nutrition measurements to meet emerging needs. The survey examines a nationally representative sample of about 5,000 persons each year. These persons are located in counties across the country, 15 of which are visited each year.

The NHANES interview includes demographic, socioeconomic, dietary, and health-related questions. The examination component consists of medical, dental, and physiological measurements, as well as laboratory tests administered by highly trained medical personnel.

To date, [thousands of research findings have been published using the NHANES data](https://www.ncbi.nlm.nih.gov/pubmed?orig_db=PubMed&term=NHANES&cmd=search).

**Dataset-3: Diabetes 130-US Hospitals for Years 1999-2008**

The Diabetes 130-US Hospitals dataset spans a decade and provides extensive hospital-based records related to diabetes care. It includes patient demographics, treatment regimens, and outcomes, offering insights into healthcare delivery and patient management strategies over time.

**What do the instances in this dataset represent?**

The instances represent hospitalized patient records diagnosed with diabetes.

**Are there recommended data splits?**

No recommendation. The standard train-test split could be used. Can use three-way holdout split (i.e., train-validation-test) when doing model selection.

**Does the dataset contain data that might be considered sensitive in any way?**

Yes. The dataset contains information about the age, gender, and race of the patients.

**Additional Information**

The dataset represents ten years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes over 50 features representing patient and hospital outcomes. Information was extracted from the database for encounters that satisfied the following criteria. (1) It is an inpatient encounter (a hospital admission). (2) It is a diabetic encounter, that is, one during which any kind of diabetes was entered into the system as a diagnosis. (3) The length of stay was at least 1 day and at most 14 days. (4) Laboratory tests were performed during the encounter. (5) Medications were administered during the encounter. The data contains such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab tests performed, HbA1c test result, diagnosis, number of medications, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc.

This project employs a structured approach encompassing data cleaning, exploratory data analysis, and the application of machine learning algorithms. The datasets undergo rigorous preprocessing to handle missing values, normalize features, and ensure compatibility across different sources. Exploratory data analysis uncovers patterns and correlations essential for feature selection and model validation.

To predict and classify diabetes outcomes, three primary algorithms are implemented: Random Forest, Support Vector Machines (SVM), and Logistic Regression. These algorithms are chosen for their ability to handle complex data relationships, classify outcomes with accuracy, and provide insights into feature importance critical for clinical decision-making.

By integrating machine learning techniques with diverse datasets, this project aims to enhance the accuracy and efficiency of diabetes diagnosis and prediction. The findings not only contribute to advancing predictive healthcare analytics but also pave the way for personalized medicine strategies tailored to diverse patient populations.

In conclusion, this project underscores the transformative potential of data-driven approaches in addressing the multifaceted challenges posed by diabetes mellitus. By harnessing the power of machine learning algorithms and leveraging diverse datasets, it seeks to empower healthcare practitioners with robust tools for early intervention and improved patient outcomes.

**Chapter – 3**

**Objectives**

**Goal –** To explore different machine learning algorithms

### 1. Pima Indians Diabetes Database

* **Predict Diabetes Onset**: Develop models to predict diabetes based on physiological data.
* **Identify Risk Factors**: Determine key risk factors for diabetes in the Pima Indian population.
* **Improve Diagnostic Criteria**: Enhance early diagnosis criteria through data analysis.

### 2. NHANES 2013-2014

* **Estimate Prevalence**: Determine the prevalence of diabetes and prediabetes in the U.S.
* **Analyze Nutrition and Lifestyle**: Examine dietary and lifestyle factors influencing diabetes.
* **Monitor Health Outcomes**: Track complications and health outcomes related to diabetes.
* **Inform Public Policy**: Provide data to support public health policies and programs.

### 3. Diabetes 130-US Hospitals for Years 1999-2008

* **Analyze Hospitalization Trends**: Study trends in diabetes-related hospitalizations.
* **Evaluate Treatment Efficacy**: Assess the effectiveness of various diabetes treatments.
* **Monitor Adverse Events**: Identify and evaluate adverse events and complications in hospitalized diabetic patients.

**Chapter-3.1**

**Objective**-1

**Data Preprocessing**

The data preprocessing stage is crucial in any data analysis pipeline. It involves cleaning and transforming raw data to ensure it is suitable for analysis. Raw data is often messy, containing missing values, outliers, and inconsistencies that can distort analysis and modelling results. Preprocessing transforms this raw data into a clean and structured format, enhancing its quality and usability.

In this project, data preprocessing was carried out on the files contained in the archive. The processed data was then saved in a cleaned format, ready for subsequent analysis and modelling tasks. This preprocessing step is vital as it lays the foundation for accurate and reliable data analysis, which is essential for deriving meaningful insights and making informed decisions.

**Dataset-1: Pima Indians Diabetes Dataset (PIDD)**

Steps Taken

1. Loading the Data:
   * The R script began by loading the necessary libraries for data manipulation and cleaning. Libraries such as those for data manipulation and reading were used.
   * The raw data files from the archive were then loaded into R data frames using appropriate functions, depending on the file formats.
2. Data Cleaning:
   * Handling Missing Values:
     + Missing values were identified and appropriately handled. Techniques such as imputation with mean/median values, or removal of rows/columns with excessive missing data, were applied.
   * Outlier Detection and Treatment:
     + Outliers in the data were detected using statistical methods and either corrected or removed to prevent skewing the analysis results.
   * Data Transformation:
     + Data was transformed to meet the requirements of the analysis. This included normalizing or standardizing numerical values, encoding categorical variables, and creating new derived variables if necessary.
   * Removing Duplicates:
     + Duplicate records were identified and removed to ensure the uniqueness of the data entries.
3. Data Integration:
   * If multiple files were involved, they were merged into a single data frame. This could involve joining datasets based on common keys or concatenating datasets with similar structures.
4. Feature Selection:
   * Unnecessary columns that did not contribute to the analysis were removed to reduce dimensionality and improve computational efficiency.
   * Relevant features were selected based on domain knowledge and preliminary analysis.
5. Saving the Cleaned Data:
   * After thorough cleaning and transformation, the final dataset was saved in a cleaned format using appropriate functions. This file is now ready for further analysis and modelling.

Code: -

library(ggplot2)

library(dplyr)

library(readxl)

install.packages("tidyverse")

library(tidyverse)

install.packages("corrplot")

library(corrplot)

install.packages("leaflet")

library(leaflet)

install.packages("xlsx")

library(xlsx)

data1 <- pidd

columns\_to\_check <- c("Blood pressure", "Skin thickness", "Insulin","Body mass index")

df\_clean <- data1[!apply(data1[columns\_to\_check], 1, function(row) any(row == 0)), ]

write.csv(df\_clean,file="cleaned2.csv")

**Dataset-2: National Health and Nutrition Examination Survey (NHANES)**

Steps Taken:

1.Dataset Merging & Attribute selection

* Merge Datasets:Datasets from the same cycle were merged using the unique participant identifier (SEQN). This integration included combining demographic, examination, laboratory, dietary, and questionnaire data into a single dataset.For analyses involving multiple cycles, datasets were harmonized by aligning variable names, formats, and coding schemes before merging.
* Attribute Selection:Relevant variables were selected based on the research objectives. This step ensured that only necessary attributes were included for further analysis.

2. Handling Missing Values

* Missing values were addressed by removing rows with excessive missing data and imputing missing values in numerical columns with the median or mean, and in categorical columns with the mode.

3. Feature Selection:

* Variable Transformation:

New features were created from existing ones to enhance analysis. For example, BMI was calculated from height and weight, and age was categorized into groups (e.g., Child, Teen, Adult, Senior).

4.Dimensionality Reduction:

* Techniques like Principal Component Analysis (PCA) or feature selection algorithms were used to reduce the number of variables while retaining essential information.

5.Correlation Analysis:

* Correlation analysis was performed to identify and remove highly correlated (redundant) variables.

6. Saving the Cleaned Data:

* After thorough cleaning and transformation, the final dataset was saved in a cleaned format using appropriate functions. This file is now ready for further analysis and modelling.

Code:-

import numpy as np

import pandas as pd

import pandas as pd

import seaborn as sns

import numpy as np

import matplotlib.pyplot as plt

import matplotlib as matplot

import re

import sklearn

import warnings

warnings.filterwarnings("ignore")

%matplotlib inline

df1 = pd.read\_csv('labs.csv')

df2 = pd.read\_csv('examination.csv')

df3 = pd.read\_csv('demographic.csv')

df4 = pd.read\_csv('diet.csv')

df5 = pd.read\_csv('questionnaire.csv')

df2.drop(['SEQN'], axis = 1, inplace=True)

df3.drop(['SEQN'], axis = 1, inplace=True)

df4.drop(['SEQN'], axis = 1, inplace=True)

df5.drop(['SEQN'], axis = 1, inplace=True)

df = pd.concat([df1, df2], axis=1, join='inner')

df = pd.concat([df, df3], axis=1, join='inner')

df = pd.concat([df, df4], axis=1, join='inner')

df = pd.concat([df, df5], axis=1, join='inner')

df.describe()

df.to\_csv('concatinated.csv', index=False)

from sklearn.feature\_selection import VarianceThreshold

df.dropna(axis=1, how='all')

df.dropna(axis=0, how='all')

df = df.rename(columns = {'SEQN' : 'ID',

'RIAGENDR' : 'Gender', 'DMDYRSUS' : 'Years\_in\_US', # Nan -> american iguess

'INDFMPIR' : 'Family\_income','LBXGH' : 'GlycoHemoglobin',

'BMXARMC' : 'ArmCircum','BMDAVSAD' : 'SaggitalAbdominal',

'MGDCGSZ' : 'GripStrength','DRABF' : 'Breast\_fed'})

df = df.loc[:, ['ID', 'Gender', 'Years\_in\_US', 'Family\_income','GlycoHemoglobin', 'ArmCircum',

'SaggitalAbdominal', 'GripStrength', 'Breast\_fed']]

df.describe()

df.info()

**Dataset-3: Diabetes 130-US Hospitals for Years 1999-2008**

Steps Taken:

1. Handling Missing Values

Missing values were addressed by removing rows with excessive missing data and imputing missing values in numerical columns with the median or mean, and in categorical columns with the mode.

2. Data Normalization

Normalization was applied to numerical features to standardize their scales. This process involved transforming the data to fall within a specific range, typically between 0 and 1.

3. Encoding Categorical Variables

Categorical variables were converted into numerical formats. Techniques such as one-hot encoding or label encoding were used to facilitate their inclusion in machine learning models.

4. Outlier Detection and Removal

Outliers were identified and removed using the Interquartile Range (IQR) method to prevent skewed results. Observations that fell significantly outside the IQR were considered outliers.

5. Feature Selection

To improve model performance and reduce complexity, irrelevant or redundant features were removed. Correlation analysis and variance thresholding helped identify features to exclude.

6. Data Splitting

The dataset was split into training and testing sets, typically using an 80-20 split ratio. This division allows for effective model evaluation and validation.

Code: -

library(ggplot2)

library(dplyr)

library(readxl)

install.packages("tidyverse")

library(tidyverse)

install.packages("corrplot")

library(corrplot)

install.packages("leaflet")

library(leaflet)

data<- read.csv("C:/Users/Harshit/Documents/R files/diabetic\_data.csv")

data1<- data %>%select(-encounter\_id,-patient\_nbr,-payer\_code,-admission\_type\_id,-discharge\_disposition\_id,-admission\_source\_id) %>% filter(complete.cases(.))

mask <- apply(data1, 1, function(row) any(row == '?'))

data\_ckeaned<- data1[!mask,]

**Chapter-4**

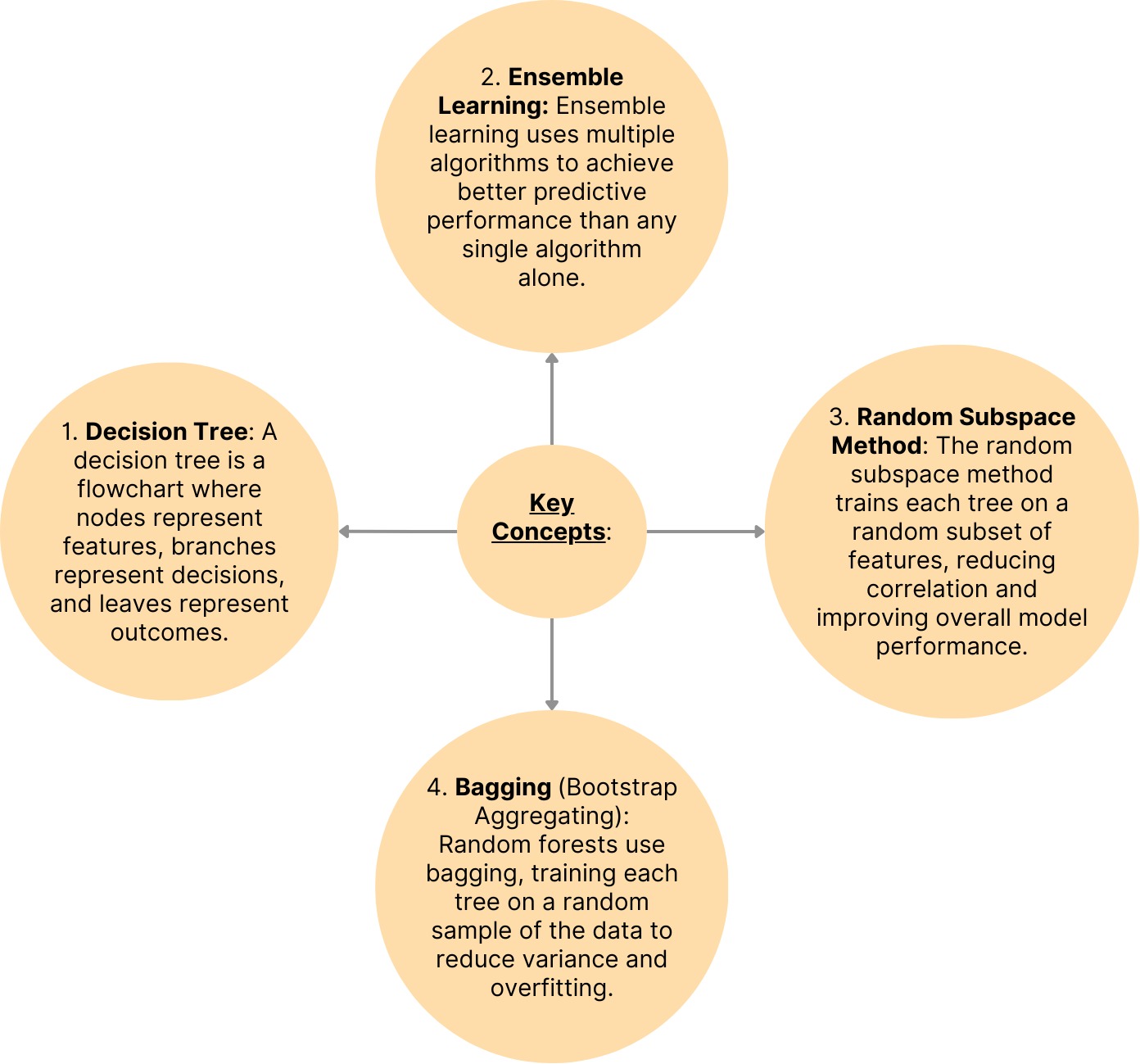
**Algorithms**

* **Random Forest Algorithm**

The Random Forest algorithm is a popular ensemble learning technique used for both classification and regression tasks. It operates by constructing a multitude of decision trees during training time and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Here are the key concepts and steps involved in the Random Forest algorithm:

Figure 9 : Key concepts of Random Forest Algorithm



**Advantages of Random Forest**

1. **Robustness**: It is less prone to overfitting compared to individual decision trees.
2. **Versatility**: It can handle both classification and regression tasks.
3. **Feature Importance**: It provides an estimate of feature importance, which can be useful for feature selection.
4. **Scalability**: It can handle large datasets efficiently.

**Disadvantages of Random Forest**

1. **Complexity**: It can be computationally intensive and require more memory.
2. **Interpretability**: The results are less interpretable compared to a single decision tree, though some interpretability can be gained through feature importance scores.
3. **Slower Predictions**: Predictions might be slower than some simpler models due to the ensemble of trees.

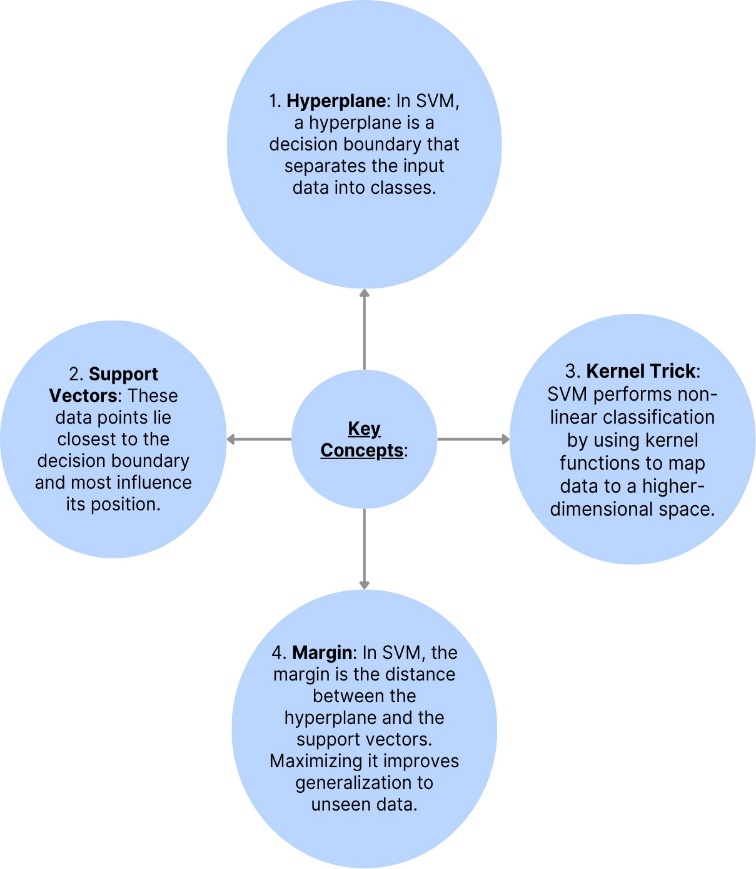
* **SVM (Support Vector Machine) algorithm**

SVM is a powerful supervised machine learning algorithm used for classification and regression tasks. It works by finding the optimal hyperplane that best separates data points belonging to different classes. Here’s an overview of the SVM algorithm:

**Advantages of SVM**

1. **Effective in High-Dimensional Spaces**: SVM performs well even when the number of dimensions is greater than the number of samples.
2. **Memory Efficient**: It uses a subset of training points (support vectors) in the decision function, making it memory efficient.
3. **Versatile**: SVM can be used for both classification and regression tasks, and with appropriate kernel functions, it can handle complex decision boundaries.
4. **Regularization**: SVM has a regularization parameter (C) that helps prevent overfitting by balancing between maximizing the margin and minimizing classification error.

Figure 10 : Key concepts of SVM algorithm



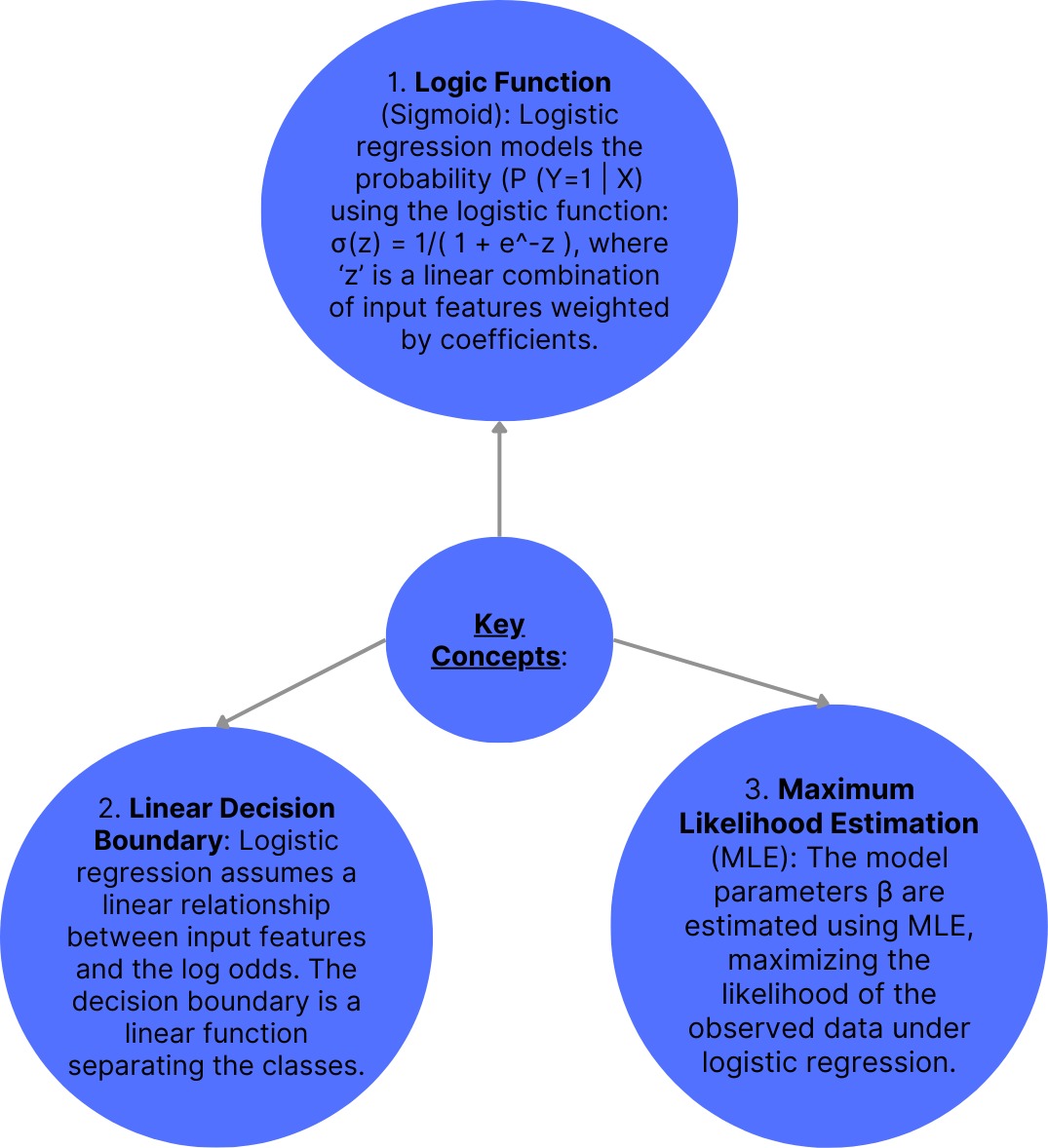
**Disadvantages of SVM**

1. **Computationally Expensive**: Training time can be high for large datasets.
2. **Choice of Kernel**: The performance of SVM depends heavily on the choice of kernel. Selecting the right kernel and tuning its parameters can be challenging.
3. **Difficulty in Interpretation**: Unlike decision trees or linear regression, SVM models are not easily interpretable.

**Logistic Regression Algorithm**

Logistic regression is a widely used statistical method for binary classification tasks. It predicts the probability of occurrence of an event by fitting data to a logistic function. Here are the key concepts and characteristics of logistic regression:

Figure 11 : Key concepts of logistic regression algorithm



**Advantages of Logistic Regression**

1. **Interpretability**: Logistic regression provides straightforward interpretation of coefficients. Each coefficient represents the change in log-odds of the outcome associated with a one-unit change in the predictor, holding other predictors constant.
2. **Efficiency**: It is computationally efficient and can be trained quickly, even with large datasets.
3. **Probabilistic Predictions**: Outputs probabilities that a given observation belongs to a particular class, enabling flexible thresholding for decision-making.
4. **Feature Importance**: Although not as direct as in tree-based models, feature importance can be inferred from the magnitude and sign of coefficients.

**Disadvantages of Logistic Regression**

1. **Linear Assumption**: Logistic regression assumes a linear relationship between the features and the log-odds of the outcome, which may not hold true for complex, non-linear relationships in the data.
2. **Limited Expressiveness**: It may not capture complex interactions between features without feature engineering or transformation.
3. **Binary Classification Limitation**: Primarily designed for binary classification tasks and may not perform well for multi-class problems without extensions like multinomial logistic regression.

**Chapter-4.1**

**Objective-2**

**Performance and Classification**

**Random Forest Algorithm on Dataset-1: Pima Indians Diabetes Dataset (PIDD)**

**Code: -**

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import classification\_report, accuracy\_score

# Load the dataset

df = pd.read\_csv('cleaned2.csv')

# Drop the unnamed column

df = df.drop(columns=['Unnamed: 0'])

# Define features (X) and target (y)

X = df.drop(columns=['Outcome'])

y = df['Outcome']

# Split the data into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Create a Random Forest Classifier

rf = RandomForestClassifier(n\_estimators=100, random\_state=42)

# Train the classifier

rf.fit(X\_train, y\_train)

# Make predictions

y\_pred = rf.predict(X\_test)

# Evaluate the model

accuracy = accuracy\_score(y\_test, y\_pred)

report = classification\_report(y\_test, y\_pred)

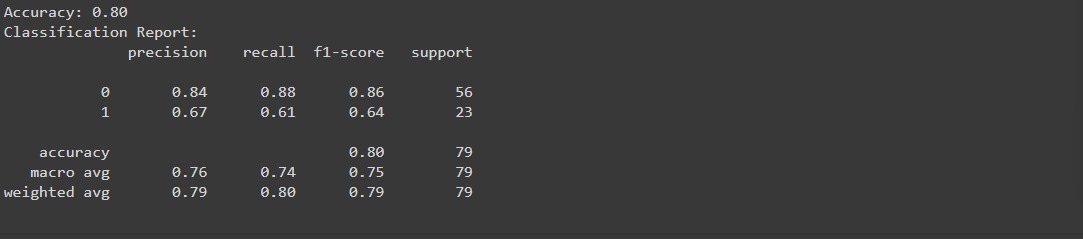
# Print the results

print(f'Accuracy: {accuracy:.2f}')

print('Classification Report:')

print(report)

**Output: -**

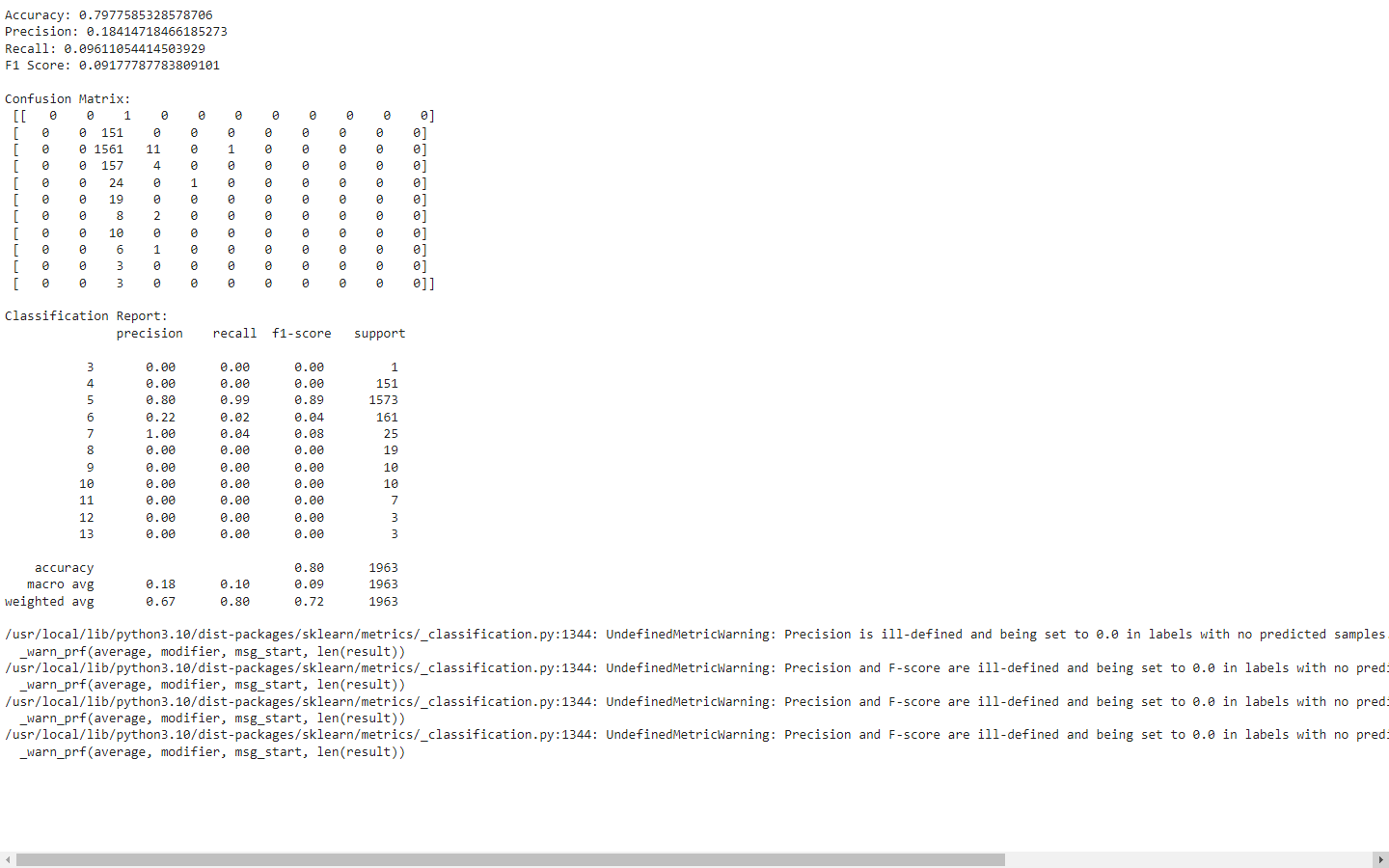
****

**Random Forest Algorithm Dataset-2: National Health and Nutrition Examination Survey (NHANES)**

**Code:**

import pandas as pdfrom sklearn.model\_selection import train\_test\_splitfrom sklearn.ensemble import RandomForestClassifierfrom sklearn.metrics import accuracy\_score, precision\_score, recall\_score, f1\_score, confusion\_matrix, classification\_report# Load the datasetfile\_path = '/content/random.csv'df = pd.read\_csv(file\_path)# Impute all numerical columns with meanfor col in df.select\_dtypes(include=['float64', 'int64']).columns: df[col].fillna(df[col].mean(), inplace=True)# Encode categorical variables (if any)# Assuming 'Gender' is categoricaldf['Gender'] = df['Gender'].astype('category').cat.codes# Convert 'Breast\_fed' to categorical if it is not alreadydf['GlycoHemoglobin'] = df['GlycoHemoglobin'].astype('int')# Ensure there are no missing values leftassert df.isnull().sum().sum() == 0, "There are still missing values in the dataset"# Split the data into features and target# Assuming 'Breast\_fed' is the target variableX = df.drop(columns=['GlycoHemoglobin'])y = df['GlycoHemoglobin']# Check unique values in the target to ensure it's categoricalprint(y.unique())# Split the dataset into training and testing setsX\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)# Train the Random Forest modelmodel = RandomForestClassifier(n\_estimators=100, random\_state=42)model.fit(X\_train, y\_train)# Make predictions on the test sety\_pred = model.predict(X\_test)# Evaluate the modelaccuracy = accuracy\_score(y\_test, y\_pred)precision = precision\_score(y\_test, y\_pred, average='macro')recall = recall\_score(y\_test, y\_pred, average='macro')f1 = f1\_score(y\_test, y\_pred, average='macro')conf\_matrix = confusion\_matrix(y\_test, y\_pred)class\_report = classification\_report(y\_test, y\_pred)# Print the evaluation metricsprint("Accuracy:", accuracy)print("Precision:", precision)print("Recall:", recall)print("F1 Score:", f1)print("\nConfusion Matrix:\n", conf\_matrix)print("\nClassification Report:\n", class\_report)

Output:



**Random Forest Algorithm Dataset-3: Diabetes 130-US Hospitals for Years 1999-2008**

**Code: -**

# Step 1: Import necessary libraries

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import accuracy\_score, classification\_report

from sklearn.impute import SimpleImputer

# Step 2: Load your dataset

df = pd.read\_csv('cleaned1.csv')

# Print column names to verify

print("Column names in the dataset before get\_dummies:")

print(df.columns)

# Step 3: Preprocess your data

# Convert categorical features to numerical using get\_dummies

df = pd.get\_dummies(df)

# Print column names again after get\_dummies

print("Column names in the dataset after get\_dummies:")

print(df.columns)

# Combine 'readmitted' columns into a single binary column

df['readmitted'] = df[['readmitted\_<30', 'readmitted\_>30']].max(axis=1)

# Drop the original 'readmitted' columns

df.drop(columns=['readmitted\_<30', 'readmitted\_>30', 'readmitted\_NO'], inplace=True)

# Step 4: Separate features and target variable

X = df.drop(columns='readmitted')

y = df['readmitted']

# Fill missing values with mean

imputer = SimpleImputer(strategy='mean')

X = imputer.fit\_transform(X)

# Step 5: Split the dataset into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Step 6: Train the Random Forest model

rf = RandomForestClassifier(n\_estimators=100, random\_state=42)

rf.fit(X\_train, y\_train)

# Step 7: Make predictions and evaluate the model

y\_pred = rf.predict(X\_test)

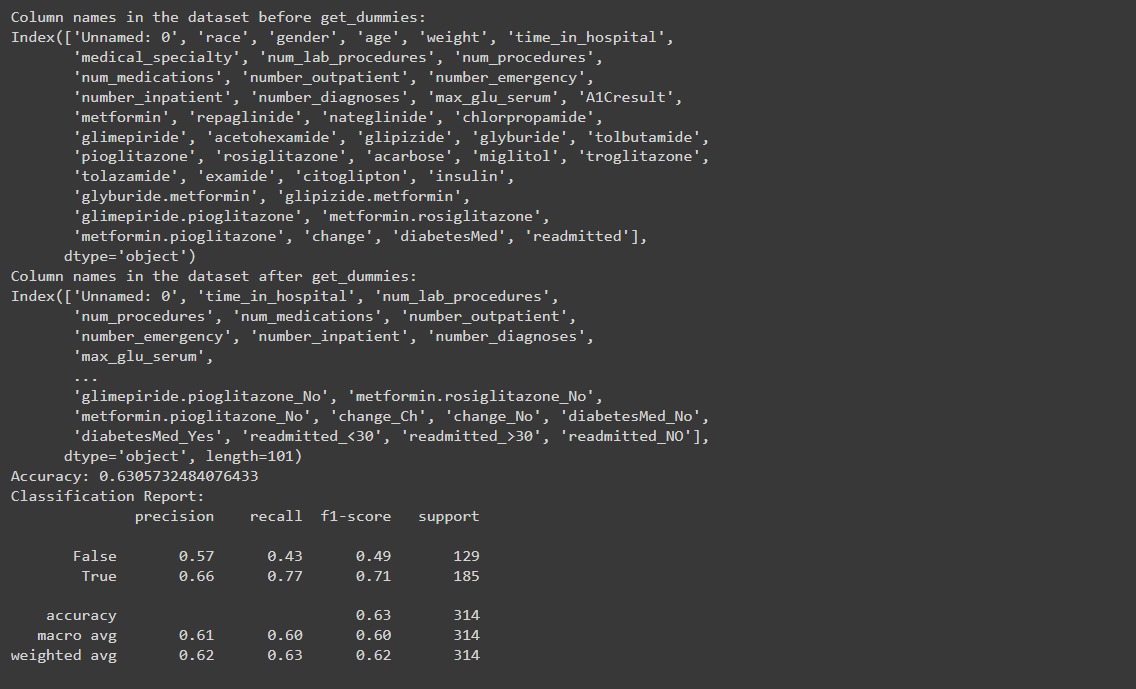
accuracy = accuracy\_score(y\_test, y\_pred)

report = classification\_report(y\_test, y\_pred)

print(f"Accuracy: {accuracy}")

print("Classification Report:")

print(report)

**Output: -**

**SVM (Support Vector Machine) Algorithm**

**Dataset-1: Pima Indians Diabetes Dataset (PIDD)**

**Code: -**

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.svm import SVC

from sklearn.metrics import classification\_report, accuracy\_score

from sklearn.preprocessing import StandardScaler

# Load the dataset

df = pd.read\_csv('cleaned2.csv')

# Drop the unnamed column

df = df.drop(columns=['Unnamed: 0'])

# Define features (X) and target (y)

X = df.drop(columns=['Outcome'])

y = df['Outcome']

# Split the data into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Feature scaling

scaler = StandardScaler()

X\_train = scaler.fit\_transform(X\_train)

X\_test = scaler.transform(X\_test)

# Create an SVM classifier

svm = SVC(kernel='linear', random\_state=42)

# Train the classifier

svm.fit(X\_train, y\_train)

# Make predictions

y\_pred = svm.predict(X\_test)

# Evaluate the model

accuracy = accuracy\_score(y\_test, y\_pred)

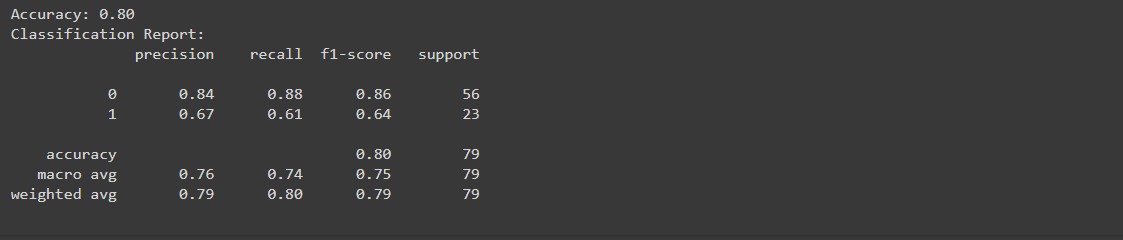
report = classification\_report(y\_test, y\_pred)

# Print the results

print(f'Accuracy: {accuracy:.2f}')

print('Classification Report:')

print(report)

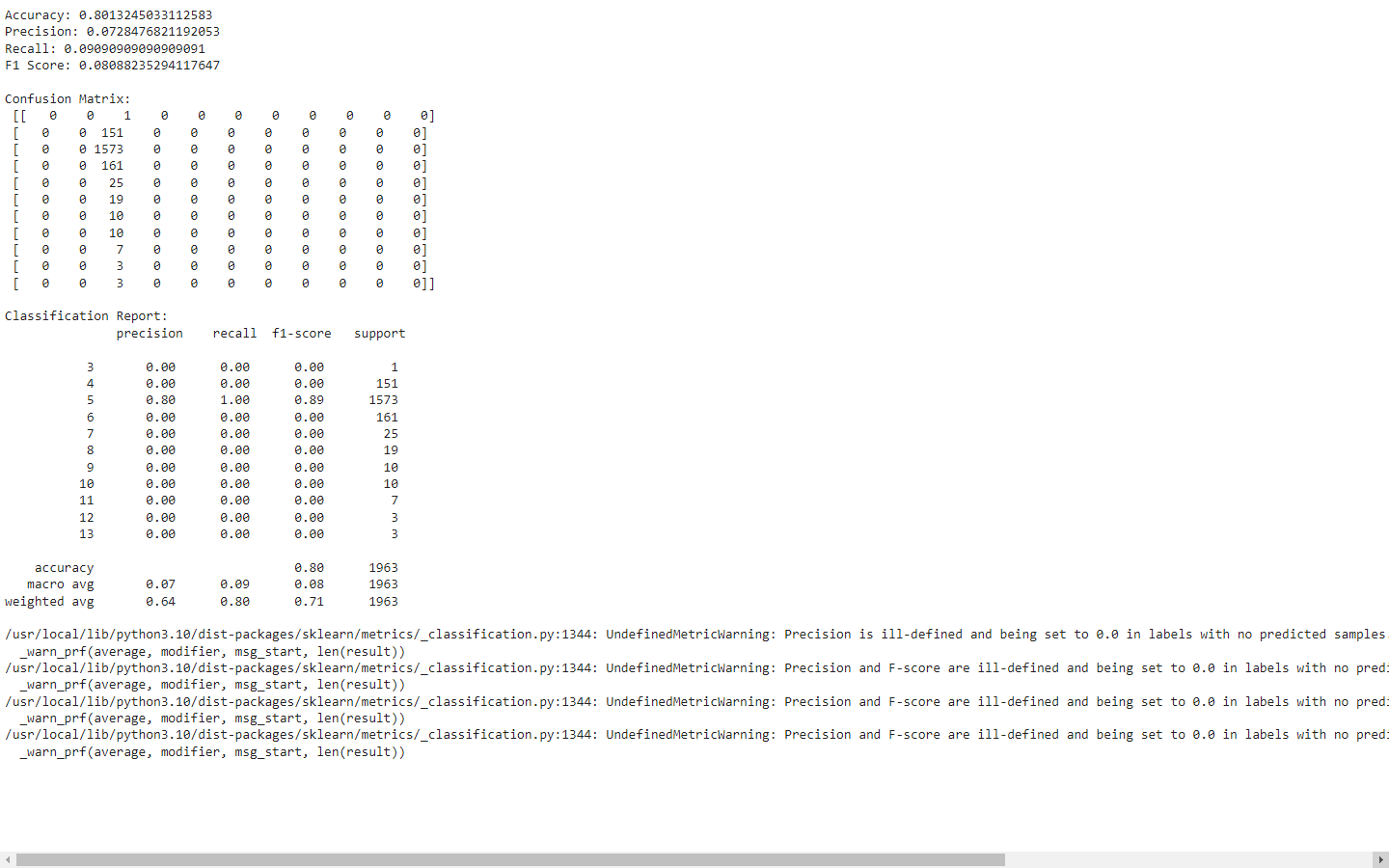
**Output: -**

**SVM (Support Vector Machine) Algorithm Dataset -2: National Health and Nutrition Examination Survey (NHANES)**

**Code:-**

import pandas as pdfrom sklearn.model\_selection import train\_test\_splitfrom sklearn.svm import SVCfrom sklearn.metrics import accuracy\_score, precision\_score, recall\_score, f1\_score, confusion\_matrix, classification\_report# Load your datadata = pd.read\_csv("/content/random.csv")# Define features (independent variables) and target variablefeatures = data.drop("GlycoHemoglobin", axis=1) # Replace "GlycoHemoglobin" with your actual target column nametarget = data["GlycoHemoglobin"]# Split data into training and testing setsX\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)# Create and train the logistic regression modelmodel = SVC(random\_state=42)model.fit(X\_train, y\_train)# Make predictions on the testing sety\_pred = model.predict(X\_test)# Evaluate model performance (e.g., accuracy score)#from sklearn.metrics import accuracy\_scoreaccuracy = accuracy\_score(y\_test, y\_pred)precision = precision\_score(y\_test, y\_pred, average='macro')recall = recall\_score(y\_test, y\_pred, average='macro')f1 = f1\_score(y\_test, y\_pred, average='macro')conf\_matrix = confusion\_matrix(y\_test, y\_pred)class\_report = classification\_report(y\_test, y\_pred)# Print the classification reportprint("Accuracy:", accuracy)print("Precision:", precision)print("Recall:", recall)print("F1 Score:", f1)print("\nConfusion Matrix:\n", conf\_matrix)print("\nClassification Report:\n", class\_report)

Output:-



**SVM (Support Vector Machine) Algorithm   
Dataset-3: Diabetes 130-US Hospitals for Years 1999-2008**

**Code: -**

# Step 1: Import necessary libraries

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.impute import SimpleImputer

from sklearn.preprocessing import StandardScaler

from sklearn.svm import SVC

from sklearn.metrics import accuracy\_score, classification\_report

# Step 2: Load your dataset

df = pd.read\_csv('cleaned1.csv')

# Print column names to verify

print("Column names in the dataset before get\_dummies:")

print(df.columns)

# Step 3: Preprocess your data

# Convert categorical features to numerical using get\_dummies

df = pd.get\_dummies(df)

# Print column names again after get\_dummies

print("Column names in the dataset after get\_dummies:")

print(df.columns)

# Combine 'readmitted' columns into a single binary column

df['readmitted'] = df[['readmitted\_<30', 'readmitted\_>30']].max(axis=1)

# Drop the original 'readmitted' columns

df.drop(columns=['readmitted\_<30', 'readmitted\_>30', 'readmitted\_NO'], inplace=True)

# Step 4: Separate features and target variable

X = df.drop(columns='readmitted')

y = df['readmitted']

# Fill missing values with mean

imputer = SimpleImputer(strategy='mean')

X = imputer.fit\_transform(X)

# Standardize the features

scaler = StandardScaler()

X = scaler.fit\_transform(X)

# Step 5: Split the dataset into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Step 6: Train the SVM model

svm = SVC(kernel='linear', random\_state=42)

svm.fit(X\_train, y\_train)

# Step 7: Make predictions and evaluate the model

y\_pred = svm.predict(X\_test)

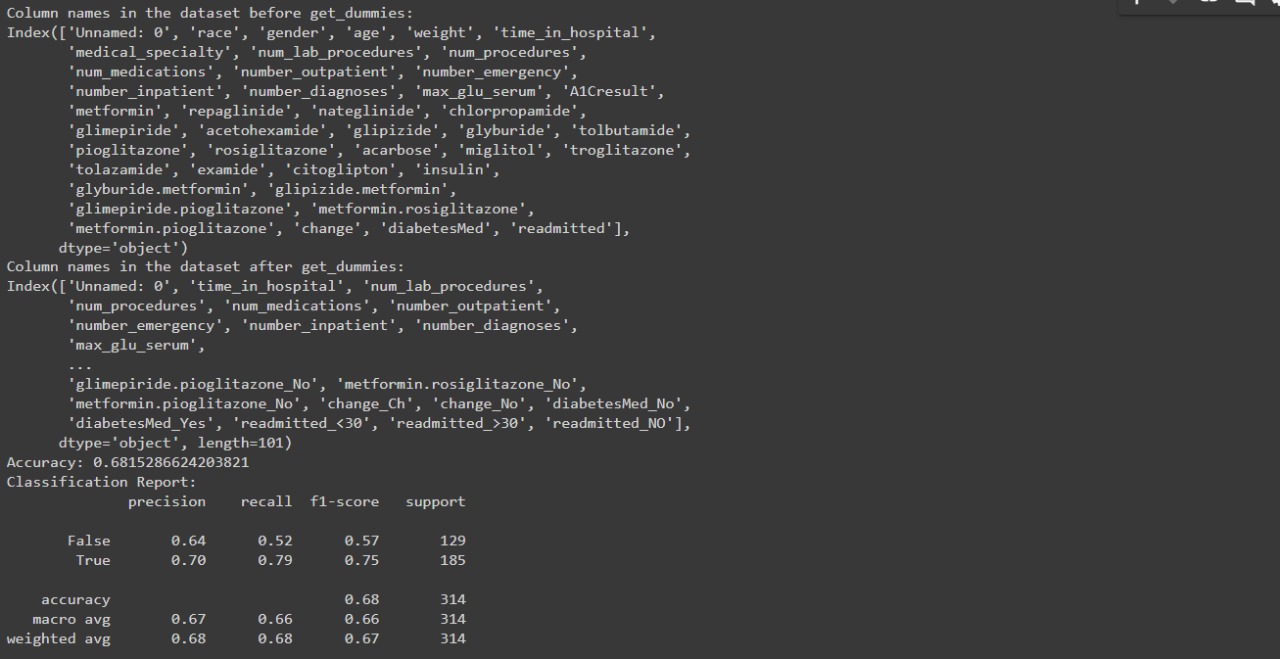
accuracy = accuracy\_score(y\_test, y\_pred)

report = classification\_report(y\_test, y\_pred)

print(f"Accuracy: {accuracy}")

print("Classification Report:")

print(report)

**Output: -**

**Logistic Regression Algorithm Dataset-1: Pima Indians Diabetes Dataset (PIDD)**

**Code: -**

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import classification\_report, accuracy\_score

from sklearn.preprocessing import StandardScaler

# Load the dataset

df = pd.read\_csv('cleaned2.csv')

# Drop the unnamed column

df = df.drop(columns=['Unnamed: 0'])

# Define features (X) and target (y)

X = df.drop(columns=['Outcome'])

y = df['Outcome']

# Split the data into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Feature scaling

scaler = StandardScaler()

X\_train = scaler.fit\_transform(X\_train)

X\_test = scaler.transform(X\_test)

# Create a Logistic Regression model

lr = LogisticRegression(random\_state=42)

# Train the model

lr.fit(X\_train, y\_train)

# Make predictions

y\_pred = lr.predict(X\_test)

# Evaluate the model

accuracy = accuracy\_score(y\_test, y\_pred)

report = classification\_report(y\_test, y\_pred)

# Print the results

print(f'Accuracy: {accuracy:.2f}')

print('Classification Report:')

print(report)

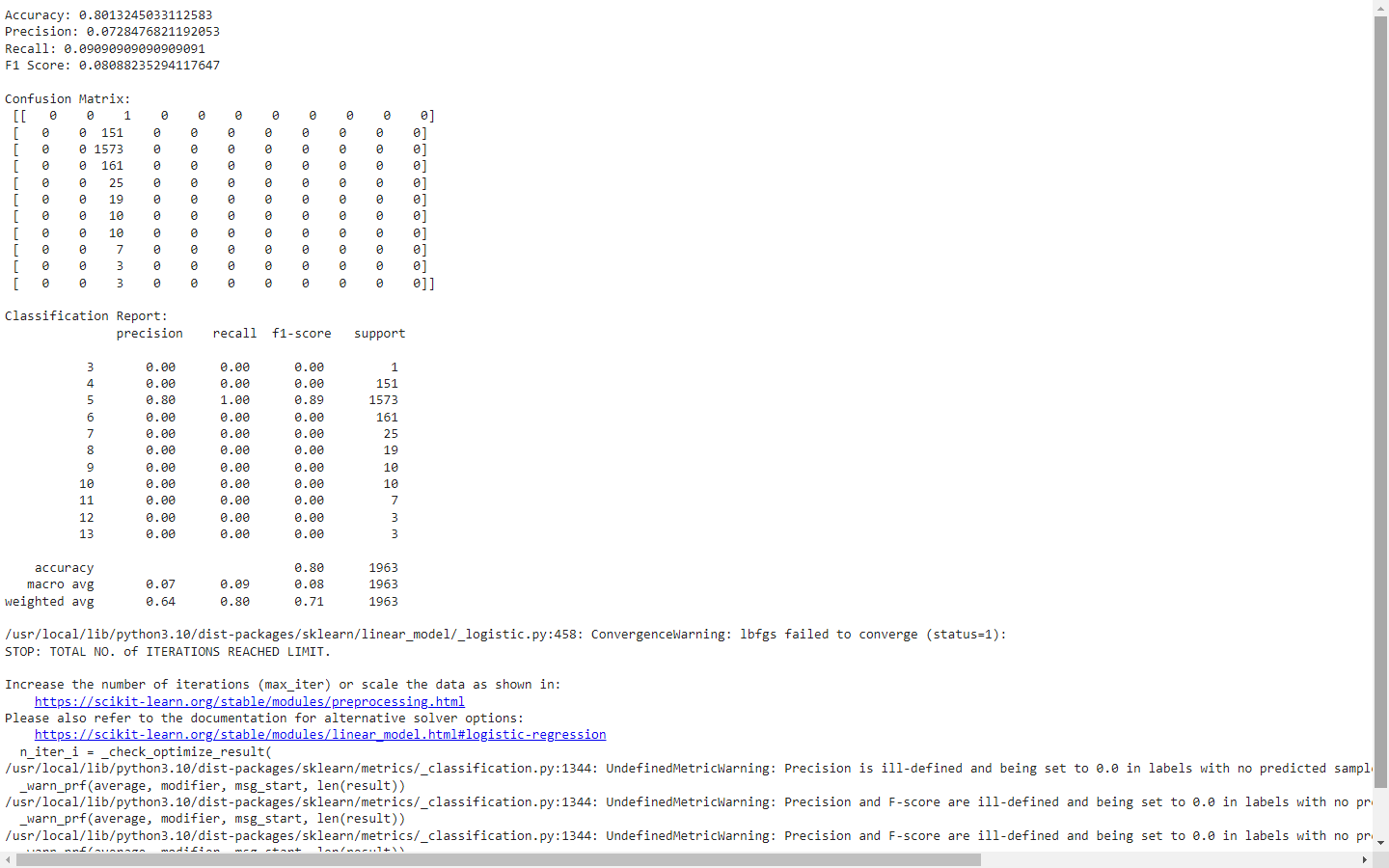
**Output: -**

**Logistic Regression Algorithm Dataset-2: National Health and Nutrition Examination Survey (NHANES)**

**Code:**

import pandas as pdfrom sklearn.model\_selection import train\_test\_splitfrom sklearn.linear\_model import LogisticRegressionfrom sklearn.metrics import accuracy\_score, precision\_score, recall\_score, f1\_score, confusion\_matrix, classification\_report# Load your datadata = pd.read\_csv("/content/random.csv")# Define features (independent variables) and target variablefeatures = data.drop("GlycoHemoglobin", axis=1) # Replace "GlycoHemoglobin" with your actual target column nametarget = data["GlycoHemoglobin"]# Split data into training and testing setsX\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)# Create and train the logistic regression modelmodel = LogisticRegression(random\_state=42)model.fit(X\_train, y\_train)# Make predictions on the testing sety\_pred = model.predict(X\_test)# Evaluate model performance (e.g., accuracy score)#from sklearn.metrics import accuracy\_scoreaccuracy = accuracy\_score(y\_test, y\_pred)precision = precision\_score(y\_test, y\_pred, average='macro')recall = recall\_score(y\_test, y\_pred, average='macro')f1 = f1\_score(y\_test, y\_pred, average='macro')conf\_matrix = confusion\_matrix(y\_test, y\_pred)class\_report = classification\_report(y\_test, y\_pred)# Print the classification reportprint("Accuracy:", accuracy)print("Precision:", precision)print("Recall:", recall)print("F1 Score:", f1)print("\nConfusion Matrix:\n", conf\_matrix)print("\nClassification Report:\n", class\_report)

Output:-



**Logistic Regression Algorithm**

**Dataset-3: Diabetes 130-US Hospitals for Years 1999-2008**

**Code:-**

# Step 1: Import necessary libraries

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.impute import SimpleImputer

from sklearn.preprocessing import StandardScaler

from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import accuracy\_score, classification\_report

# Step 2: Load your dataset

df = pd.read\_csv('cleaned1.csv')

# Print column names to verify

print("Column names in the dataset before get\_dummies:")

print(df.columns)

# Step 3: Preprocess your data

# Convert categorical features to numerical using get\_dummies

df = pd.get\_dummies(df)

# Print column names again after get\_dummies

print("Column names in the dataset after get\_dummies:")

print(df.columns)

# Combine 'readmitted' columns into a single binary column

df['readmitted'] = df[['readmitted\_<30', 'readmitted\_>30']].max(axis=1)

# Drop the original 'readmitted' columns

df.drop(columns=['readmitted\_<30', 'readmitted\_>30', 'readmitted\_NO'], inplace=True)

# Step 4: Separate features and target variable

X = df.drop(columns='readmitted')

y = df['readmitted']

# Fill missing values with mean

imputer = SimpleImputer(strategy='mean')

X = imputer.fit\_transform(X)

# Standardize the features

scaler = StandardScaler()

X = scaler.fit\_transform(X)

# Step 5: Split the dataset into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Step 6: Train the Logistic Regression model

lr = LogisticRegression(random\_state=42, max\_iter=1000)

lr.fit(X\_train, y\_train)

# Step 7: Make predictions and evaluate the model

y\_pred = lr.predict(X\_test)

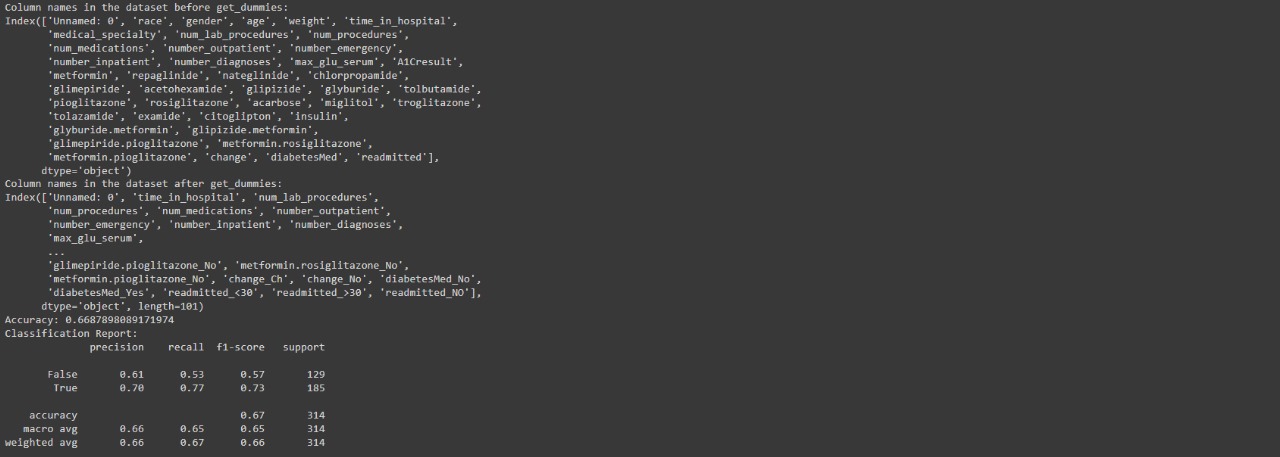
accuracy = accuracy\_score(y\_test, y\_pred)

report = classification\_report(y\_test, y\_pred)

print(f"Accuracy: {accuracy}")

print("Classification Report:")

print(report)

**Output: -**

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No** | **Random Forest Algorithm**  **Accuracy** | **SVM (Support Vector Machine) Algorithm Accuracy** | **Logistic Regression Algorithm Accuracy** |
| **Dataset-1: Pima Indians Diabetes Dataset (PIDD)** | 80% | 80% | 81% |
| **Dataset-2: National Health and Nutrition Examination Survey (NHANES)** | 79% | 80% | 80% |
| **Dataset-3: Diabetes 130-US Hospitals for Years 1999-2008** | 63% | 68% | 66% |

Figure 12 : Table for accuracy Values of Different Datasets

**Chapter-5**

**CONCLUSION**

Each model demonstrated strengths and weaknesses based on the specific characteristics of the datasets. The Logistic Regression Algorithmshowed the highest accuracy with the Pima Indians Diabetes Dataset, indicating its suitability for smaller, more defined datasets. Random Forest, despite its lower accuracy on the Diabetes 130-US Hospitals for Years 1999-2008 dataset, offered valuable insights into high-dimensional health data. Linear Regression was effective in predicting outcomes in the Pima Indians Diabetes Dataset (PIDD), though its performance relies on linear assumptions.

**Future Work**

Future work will focus on optimizing these models, particularly addressing the limitations observed with Random Forest, to improve overall performance and provide deeper insights into diabetes management. Exploring other algorithms and applying them on datasets .Exploring the future technology trends in Healthcare field for prediction of Diabetics and other Diseases.

**Chapter-6**

**INTERNSHIP**

**Hardhik:-**

Completed a 12-week internship as a Frontend Developer at Servcrust, creating multiple applications using Angular.

* Created a website with a user information form
* Converted HTML into Angular by separating it into components and enabling communication through data passing.
* Created an orders table which shows the information about the present state of the order

**Harshit:-**

I recently completed a 12-week internship as a Backend Developer at Servcrust, where I worked with Eclipse on various tasks related to Spring project development and API implementation.

* My work involved creating backend functionalities, integrating databases
* Implementing microservices architecture. I also worked on
* Microservices implementations, focusing on designing scalable and modular applications.

**Azhar:**

I recently completed an 8-week internship at Pragmatiq Systems, aimed at learning and exploring workflow and decision automation using Camunda 8.

* During the internship, I gained hands-on experience with BPMN, DMN, and CMMN, understanding how to model structured, rule-based, and event-driven workflows.
* I learned to use the Camunda Modeler to create and deploy process models, and explored integration techniques using REST APIs, Java SDKs, and external task workers.
* The internship helped me understand real-world use cases such as onboarding workflows, order management, and service escalation, and how Camunda enables automation for these scenarios.

**P.Vaishnavi:**

Completed a 12-week internship as a Data Engineer at Servcrust, where I gained hands-on experience in data engineering practices within a fast-paced, real-world environment.

* Extracted data to build metrics or create dashboards, tracking success toward KPIs.
* Created visualizations using tools like Plotly and Metabase to communicate findings to non-technical stakeholders.
* Created graphs and charts detailing data analysis results. Completed day-to-day duties accurately and efficiently

**Yashaswini:**

Completed a 12-week internship as a Full Stack Developer at Servcrust, worked on developing and managing web applications, including a Movie App, where I built key features using HTML, CSS, JavaScript, Node.js, Express.js, and MongoDB.

* I implemented CRUD operations, designed RESTful APIs, and performed efficient data handling using both SQL and MongoDB.
* Designing and implementing RESTful APIs.
* Performing database operations such as insertion, updates, and data retrieval using both SQL and NoSQL databases.
* Handling user authentication and implementing secure login mechanisms.

**References & Links**

Dataset-

<https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008>

<https://www.kaggle.com/datasets/jimschacko/10-years-diabetes-dataset/data>

<https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/data>

<https://www.kaggle.com/datasets/cdc/national-health-and-nutrition-examination-survey>

Papers-

<https://www.sciencedirect.com/science/article/pii/S1877050920308024#cebibsec1>

<https://link.springer.com/article/10.1007/s11042-023-16407-5#article-info>

Images –

<https://link.springer.com/article/10.1007/s11042-023-16407-5/figures/6>