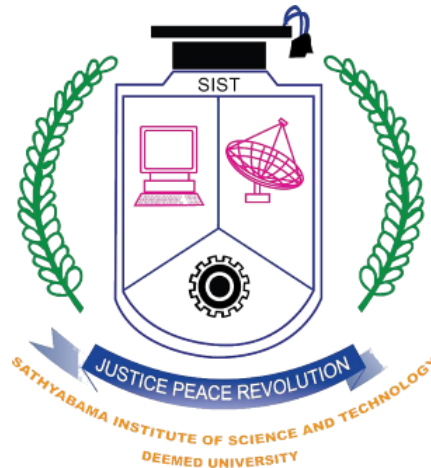


Disease Prediction Application Using Machine Learning

Submitted in partial fulfillment of the requirements for the award of
**Bachelor of Science in Bioinformatics and Data
Science**

By

R Nithish (41738008)



**DEPARTMENT OF BIOINFORMATICS
SCHOOL OF BIO AND CHEMICAL ENGINEERING**

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade “A++” by NAAC I Category-1 University by UGC I Approved by AICTE

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai - 600 119

APRIL – 2024



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade "A++" by NAAC | Category-1 University by UGC | Approved by AICTE

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600 119

www.sathyabama.ac.in

DEPARTMENT OF BIOINFORMATICS

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **R Nithish (41738008)**, who carried out the final year project entitled "***Disease Prediction Application Using Machine Learning***" under our supervision from January 2024 to April 2024.

Dr. Swetha Sunkar

Internal Guide

Dr. H. JEMMY CHRISTY

Head of the Department

Submitted for Viva Voce Examination held on 27/04/2024

Internal Examiner

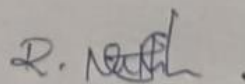
External Examiner

DECLARATION

I, **R. Nithish**, hereby declare that the Project Report entitled **Disease Prediction Application Using Machine Learning** done by me under the guidance of **Dr. Swetha Sunkar** at Sathyabama Institute of Science and Technology, Jeppiaar Nagar, Rajiv Gandhi Road, Chennai-600119, Tamil Nadu, India, is submitted in partial fulfillment of the requirements for the award of **Bachelor of Science in Bioinformatics and Data Science**.

DATE:

PLACE:



SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENTS

I am delighted to express my heartfelt acknowledgement to the **Board of Management of SATHYABAMA** for their kind encouragement and support throughout the duration of my project. Their guidance has been invaluable in ensuring the successful completion of this project, and I am sincerely grateful for their contribution.

I convey my thanks to **Dr. Rekha Chakravarthy, Dean, School of Science and Humanities** for providing me necessary support and details at the right time during the progressive reviews.

I sincerely thank **Dr. H. Jemmy Christy, Head of the Department, Dept. of Bioinformatics**, for her unwavering support and timely provision of necessary details during the progressive reviews. Her support has been instrumental in helping me to overcome obstacles and achieve success in this project.

I am deeply indebted to my **Project Guide, Dr. Swetha Sunkar**, for her invaluable guidance, suggestions, and constant encouragement throughout the project. Her expertise and insights have been instrumental in shaping my work and ensuring its successful completion.

I also wish to extend my gratitude to my family for their support and assistance in various ways, which have contributed to the successful completion of this project.

R Nithish



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade "A++" by NAAC I Category-1 University by UGC I Approved by AICTE

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600 119

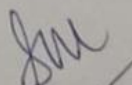
www.sathyabama.ac.in

CERTIFICATE

This is to certify that the dissertation entitled "**Disease Prediction Application Using Machine Learning**" was carried out by **Mr. R Nithish (Reg. No. 41738008)**, Department of Bioinformatics and Data Science, Sathyabama Institute of Science and Technology, Chennai in partial fulfillment of the requirements of the award of the **Degree of Bachelor of Science in Bioinformatics and Data Science** 2021-2024, is a bonafide record of the project work done under my guidance and supervision. This thesis or part, therefore, has not been submitted elsewhere for any other degree or diploma.

Date: 27/04/24

Place: Chennai


Dr. Swetha Sunkar

ABSTRACT

Health systems collect data and reports from hospitals or patient databases using machine learning and data processing techniques used in disease forecasting to generate reports that support the results used in various disease forecasts that have been the main explanation for human death since recent years. Medical reports and data were extracted from multiple databases to predict a number of necessary diseases that are now common in people with breast cancer, heart disease and diabetes, making their life a more critical measure. Today, the development of health technology has helped people to facilitate their process by recommending hospitals and doctors, where to go for his treatment, where to go and which hospitals are the easiest to treat the desired disease. We have implemented this type of system in our app to make people's lives easier by predicting the disease by inserting certain information from the reports that can give a positive or negative result. This supported the disease's prediction that they had a chance of getting a referral. of the best hospitals with the best doctors of previous users or guards nearby.

TABLE OF CONTENTS

CHAPTER NO	TITLE	PAGE NO
	ABSTRACT	I
	LIST OF ABBREVIATION	IV
	LIST OF FIGURES	V
1	INTRODUCTION	1
	1.1 DIABETES PREDICTION	1
	1.2 HEART DISEASE PREDICTION	2
	1.3 BREAST CANCER PREDICTION	2
	1.4 MACHINE LEARNING	2
	1.5 ALGORITHM	5
	1.6 APPLICATION OF MACHINE LEARNING	6
	1.7 WEB APPLICATION LIBRARY	8
	1.8 WHY MACHINE LEARNING BETTER FOR DISEASE PREDICTION	10
2	REVIEW OF LITERATURE	12
3	AIM AND OBJECTIVES	17
4	METHODOLOGY	18
	4.1 DATA COLLECTION	18
	4.2 LOAD THE DATASET	19
	4.3 DATA PREPARATION	20
	4.4 SPLIT THE DATA	22
	4.5 INITIALIZE AND TRAIN	23

	4.6 MODEL EVALUATION	24
	4.7 PREDICTION	25
	4.8 CREATE OG WEB PAGE	27
5	RESULTSANDDISCUSSION	34
	5.1 RESULTS	34
	5.2 DIABETES PREDICTION	34
	5.3 HEART DISEASE PREDICTION	39
	5.1.3 BREAST CANCER PREDICTION	43
6	SUMMARY AND CONCLUSION	53
	6.1 SUMMARY	53
	6.2 CONCLUSION	53
	REFERENCES	55

LIST OF ABBREVIATIONS

ML	Machine Learning
HDL	High-density lipoprotein
LDL	Low-density lipoprotein
VLDL	Very low-density lipoprotein
BMI	Body mass index
SVM	Support vector machine
MCP	Monocyte Chemoattractant Protein
KNN	K-nearest neighbors
API	Application Programming Interface
EHR	Electronic Health Record
CT	Computed Tomography
PCA	Principal Component Analysis
t-SNE	t-distributed Stochastic Neighbor Embedding
RBF	Radial Basis Function
MRI	Magnetic Resonance Imaging
UCI	University of California Irvine
CSV	Comma Separated Values

LIST OF FIGURES

- Figure 5.1** Diabetes Parameters.
- Figure 5.2** Logistic Regression Accuracy.
- Figure 5.3** Support Vector Machine Accuracy.
- Figure 5.4** Decision Tree Accuracy.
- Figure 5.5** Heart Disease Parameters.
- Figure 5.6** Logistic Regression Accuracy.
- Figure 5.7** Support Vector Machine Accuracy.
- Figure 5.8** Decision Tree Accuracy.
- Figure 5.9** Breast cancer Parameters.
- Figure 5.10** Logistic Regression Accuracy.
- Figure 5.11** Support Vector Machine Accuracy.
- Figure 5.12** Decision Tree Accuracy.
- Figure 5.13** Web Development.
- Figure 5.14** Diabetes Code
- Figure 5.15** Heart Disease Code
- Figure 5.16** Breast Cancer Code
- Figure 5.17** Web Page of Diabetes
- Figure 5.18** Result Page of Diabetes
- Figure 5.19** Web Page of Heart Disease.
- Figure 5.20** Result Page of Heart Disease.
- Figure 5.21** Web Page of Breast Cancer
- Figure 5.22** Result Page of Breast Cancer.

CHAPTER 1

INTRODUCTION

An accurate and effective prognosis of the disease is important for early diagnosis and effective treatment. Machine learning (ML) technology has become a powerful tool for disease prediction in healthcare. It uses data to identify patterns and make highly accurate predictions. In this context, the use of machine learning models to predict various diseases such as diabetes, heart disease and breast cancer is an important area of research and application.

Machine learning models for multi-disease prediction combine different features for each disease using information from different sources such as medical data, genetic data, lifestyle factors and diagnostic tests. These models are trained on datasets labeled with known diseases, allowing them to learn complex patterns and relationships to make accurate predictions based on new data.

Application of ML-based disease prediction models in healthcare can help doctors in early diagnosis and risk management identification assessment and personal treatment planning. By leveraging data and ML algorithms, we can improve health outcomes, reduce costs, and improve patient care in the field of disease prediction. (Patel et al.,2019)

1.1 DIABETES PREDICTION:

Diabetes is a chronic metabolic disorder characterized by long-term high blood sugar levels. ML models used to predict diabetes often use features such as sex, age, urea, creatinine, hemoglobin, cholesterol, triglycerides, HDL, LDL, VLDL, BMI. Algorithms such as logistic regression, decision trees and support vector machines (SVM) are commonly used in diabetes prediction tasks. (Rahman et al.,2021)

1.2 HEART DISEASE PREDICTION:

Heart disease includes a number of conditions that predispose to heart attack and heart failure. ML models to predict heart disease use features such as age, anemic creatine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking, time. Algorithms such as logistic regression, support vector machines (SVM), and decision tree models are used to predict the risk of heart disease. (Zupan,2019)

1.3 BREAST CANCER PREDICTION:

Breast cancer is one of the most common cancers in women worldwide. ML models for breast cancer prediction often use features such as age, BMI, glucose, insulin, home, leptin, adiponectin, resistin, MCP. Algorithms such as support vector machines (SVM), logistic regression and decision trees are used to predict the probability of breast cancer. (Barros et al.,2019)

1.4 MACHINE LEARNING:

Machine Learning (ML) is revolutionizing industries by automating processes, extracting insights from massive data sets and enabling predictive analytics. One of ML's main strengths is its ability to handle complex data patterns and relationships that traditional programming methods can struggle with. Supervised learning algorithms such as regression and classification are widely used for tasks such as predicting sales trends, detecting spam or diagnosing diseases from patient data. Unsupervised learning techniques such as clustering and dimensionality reduction reveal hidden patterns and structures in data, aiding in customer segmentation, outlier detection and data visualization.

Ensemble learning methods such as random forests and gradient boosting combine multiple models to improve accuracy and generalization. These techniques are particularly effective in scenarios where a single model may not capture all the nuances of the data. In addition, deep learning, a subset of ML, emerged due to its ability to process large amounts of unstructured data such as images, text and audio

Multilayer neural networks learn hierarchical representations that provide superior performance in tasks such as image recognition, natural language processing, and speech synthesis.

ML models are constantly evolving with the availability of algorithms, computing power, and data. Integrating ML into sectors including healthcare, finance, retail and manufacturing drives innovation and improves decision-making processes. However, challenges such as model interpretability, ethical considerations, and misdirection are still areas of active research and development in machine learning. (Kim et al.,2019)

1.4.1 SUPERVISED LEARNING:

Supervised learning involves training ML models on labeled data, where each input example is associated with a corresponding output label. The goal is to learn a mapping function that can accurately predict the output headings from new, unseen input data. Supervised learning tasks can be classified into regression and classification:

Regressive: In regression tasks, the model predicts a continuous outcome variable. For example, predicting housing prices based on features such as square footage, location and number of rooms.

Classification: In classification tasks, the model predicts a distinct class identifier or category. For example, classifying emails as spam or non-spam based on their content and characteristics.

Supervised learning algorithms include linear regression, logistic regression, decision trees, support vector machines (SVM), k-nearest neighbors (KNN) and neural networks. These algorithms learn from recognized examples during training and make predictions based on the learned patterns. (Islam et al.,2019)

1.4.2 UNSUPERVISED LEARNING:

Unsupervised learning deals with anonymized data with the goal of discovering patterns, structures or relationships hidden in the data without explicit guidance or supervision. Unsupervised learning tasks include:

Clustering: grouping similar data points based on their characteristics. Examples include k-means clustering, hierarchical clustering, and DBSCAN.

Dimensionality reduction: reducing the number of features or data while preserving relevant information. Dimensionality reduction uses techniques such as Principal Component Analysis (PCA) and t-Distribution Stochastic Neighboring (t-SNE).

Association Rule Learning: Identification of interesting associations or relationships between variables in data. Apriori and FP growth are common algorithms for learning association rules.

Unsupervised learning algorithms do not rely on labeled data for training. Instead, they discover patterns based on the internal structure of the data, enabling customer segmentation, outlier detection, and data compression.

Both supervised and unsupervised learning play an important role in machine learning applications. While supervised learning is suitable for tasks with labeled data and clear output goals, unsupervised learning is valuable for exploratory analysis, pattern discovery, and understanding the underlying structure of predefined unlabeled datasets. (White,2019)

1.5 ALGORITHM:

1.5.1 LOGISTIC REGRESSION:

Logistic regression is a versatile and widely used statistical method in machine learning for binary classification tasks. This is especially good for scenarios where the outcome variable is binary, such as predicting whether or not we have spam,

classifying whether a transaction is fraudulent or legitimate, or diagnosing whether a patient has a particular disease. The main advantage of logistic regression is its ability to model the probability of an event based on input characteristics, which makes it interpretable and intuitive to understand.

Training a logistic regression model uses a technique called maximum likelihood estimation, where the goal is to find a set of coefficients that maximize the observation of the given data probability. During training, the model learns the relationship between the input features and the probability of the target class, adjusting the coefficients to minimize the logistic loss function. Regularization techniques can be applied to logistic regression to avoid overfitting and improve generalization to unseen data. In general, the simplicity, interpretability and efficiency of logistic regression make it a valuable tool for solving binary classification problems in various fields. (Johnson et al., 2019)

1.5.2 DECISION TREE:

Decision trees are versatile and intuitive machine learning algorithms used in both classification and regression tasks. They operate recursively partitioning regions of the feature space based on input feature values, creating a tree-like structure of decision nodes and leaf nodes. Each decision node represents an attribute and assignment criterion, while each leaf node represents a predicted result or value.

One of the main advantages of decision trees is their interpretability. The decision rules learned by the tree can be easily visualized and understood, making them valuable for explaining model predictions to stakeholders and domain experts. Decision trees are also robust to outliers and missing values in the data because they can efficiently handle categorical and numerical features.

However, decision trees are prone to overfitting, especially when the tree is deep and complex. Techniques such as pruning, limiting tree depth, and using combined methods such as random forests can help reduce overfitting and improve generalizability. Decision trees find applications in a variety of fields, including healthcare for disease diagnosis based on patient characteristics, financing credit risk

assessment and customer segmentation for marketing, making them a widely used and powerful machine learning algorithm. (Montan et al.,1992;Kim et al.,2020)

1.5.3 SUPPORT VECTOR MACHINES:

Support Vector Machines (SVM) are powerful supervised learning algorithms used for both classification and regression tasks. They are particularly effective in solving binary classification problems by finding the optimal hyperplane that separates data points of different classes by the largest margin. SVMs perform well in high-dimensional spaces and are resistant to overfitting, making them suitable for complex datasets with many features.

The central idea of SVMs is to map the input data to a higher dimensional space using a kernel function where the data points become linearly separable. The optimal hyperplane is then determined by maximizing the margin, which is the distance between the hyperplane and the closest data points (support vectors) for each class. SVMs can handle linear and nonlinear decision constraints using different kernel functions such as linear, polynomial, radial basis function (RBF), and sigmoid. (Guptan et al.,2020)

1.6 APPLICATIONS OF MACHINE LEARNING:

Machine learning has significantly contributed to the prediction of diseases in various health fields. Some of the main applications of machine learning in disease prediction are:

1.6.1 Diabetes prediction:

Machine learning models are used to predict diabetes risk based on factors such as glucose levels, insulin resistance, BMI, family history and lifestyle factors. These models help identify at-risk people early, enabling preventive interventions and personalized treatment plans.

1.6.2 Cardiac risk assessment:

Machine learning algorithms analyze medical data such as blood pressure, cholesterol levels, EKG values and lifestyle habits. to assess a person's risk of developing heart disease. These models can identify patterns and risk factors that may not be apparent with traditional risk assessment methods, resulting in more accurate predictions and more targeted prevention strategies.

1.6.3 Cancer detection and diagnosis:

Machine learning plays a critical role in cancer detection and diagnosis. . analysis of image data (eg, mammograms, MRI scans, CT scans) and biomarkers to detect suspicious lesions or tumors. ML models can assist radiologists and oncologists in the early detection, staging and treatment planning of various cancers, including breast, lung and prostate cancer.

1.6.4 Alzheimer's disease prediction:

Machine learning methods analyze neuroimaging data and genetic markers. , cognitive assessments, and patient history to predict risk of Alzheimer's disease and other neurodegenerative diseases. These models help with early detection, monitor disease progression, and create personalized treatment plans for patients.

1.6.5 Predicting infectious outbreaks:

Machine learning algorithms analyze epidemiological data, environmental factors, population demographics, and social media trends to predict and trace infectious. disease outbreaks. epidemics such as influenza, Zika virus and COVID-19. These models enable public health officials to take timely action, allocate resources efficiently, and contain the spread of infectious diseases.

1.6.6 Drug response prediction:

Machine learning is used to predict individual patient responses to drugs based on genetic profiles, clinical data, and drug interactions. These predictive models help healthcare providers choose the most effective and personalized treatment options, minimize side effects and improve patient outcomes.

Overall, machine learning in disease prediction provides healthcare professionals with advanced analytics and predictive information that leads to early detection. , targeted interventions, personalized treatment plans, and improved patient outcomes for a variety of diseases and conditions. (Pangty et al.,2019)

1.7 WEB APPLICATION LIBRARY:

1.7.1 STREAMLIT: (Streamlit pypi)

Streamlit is a Python library designed for fast and efficient development of interactive web applications, especially in the field of information technology and machine learning. Its intuitive API allows developers to create dynamic user interfaces using Python scripts without requiring extensive knowledge of web development languages or frameworks. Streamlit allows developers to easily integrate interactive widgets such as sliders, drop-down menus and buttons to allow the user to enter and edit data visualizations and analyses.

One of Streamlit main strengths is its ability to facilitate rapid prototyping and testing. Developers can immediately see the impact of code changes in real time, simplifying application design and operations. In addition, Streamlit offers seamless deployment options that allow applications to be deployed on-premises, to cloud platforms such as Heroku or AWS, or to Streamlit Sharing to facilitate sharing and collaboration. This combination of simplicity, interactivity and deployment flexibility makes Streamlit a popular choice among data scientists and developers to create immersive and interactive web applications for exploring data, analyzing data and implementing machine learning models.

1.7.2 STREAMLIT OPTION MENU:

Streamlit settings menu provides a convenient way to create drop-down menus where users can select one option from a predefined list. This widget is particularly useful in interactive applications where users must choose from a range of options, such as selecting a specific model or dataset for analysis. The syntax for creating an option menu is simple, requiring only a list of options and a dropdown menu tag. Users can easily browse options and make selections, which can later be used as input to create dynamic content, perform calculations, or display relevant information based on the selected option.

In addition, Streamlit options menu supports a variety of data. list types of options, including strings, integers, floats, and even complex objects. This flexibility allows developers to create versatile drop-down menus to suit different use cases and scenarios. In addition, developers can improve the usability of option menus by dynamically populating the list of options based on user input or external data sources, enabling a more interactive and personalized experience for users interacting with Streamlit applications.

1.7.3 PICKLE:(Pickle pypi)

Pickle is a Python module that allows you to serialize and deserialize Python objects. Serialization refers to the process of converting objects into a format that can be easily stored or transmitted, such as a binary file or byte stream. Deserialization, on the other hand, deals with reconstructing objects from their serialized form. Pickle is commonly used to store and load machine learning models, data structures, and other complex Python objects, making it a valuable tool for storing and communicating data.

One of the main advantages of using Pickle is its simplicity and ease of use. With just a few lines of code, developers can serialize Python objects and save them to a file or database. Similarly, deleting a set of objects from Pickle files is easy, allowing quick and efficient retrieval of stored objects. Pickle is also compatible with many Python objects, including custom classes and nested data structures, making it versatile for the various serialization needs of Python applications. However, it is important to note

that Pickle files can only be safely deleted in the same Python environment with the same libraries and versions, as changes to the environment can cause compatibility issues.

1.8 WHY MACHINE LEARNING IS BETTER FOR DISEASE PREDICTION?

Machine learning (ML) offers several advantages that make it particularly good for disease prediction compared to traditional statistical methods. Some of the main reasons why ML is better for disease prediction are:

1.8.1 Complex pattern recognition:

ML algorithms can analyze massive amounts of data and identify complex patterns and relationships that may not be obvious to human experts or traditional statistical methods. This ability is critical to understanding the multifaceted nature of diseases and predicting outcomes based on different characteristics.

1.8.2 Non-linear relationships:

Many diseases have non-linear relationships between risk factors and outcomes. ML models, especially those based on deep learning and fundamental methods such as support vectors (SVM), can effectively capture non-linear relationships, enabling more accurate and nuanced predictions.

1.8.3 Feature selection and extraction:

ML techniques enable automatic selection of functions and extraction. to identify the most important variables or biomarkers from large datasets. This helps focus on the most important predictors and reduce the impact of irrelevant or redundant features on predictive accuracy.

1.8.4 Processing high-dimensional data:

Advances in computing power and algorithms enable ML models to process high-dimensional data, including genomic data, medical imaging electronic health records (EHR), and omics data. This capability is necessary to integrate different data sources and gain valuable insights for disease prognosis.

1.8.5 Personalized medicine:

ML models can facilitate personalized medicine by predicting individual risk profiles and treatment responses based on patient-specific characteristics, genetic markers, lifestyle factors, and environmental influences. This customized approach improves the accuracy and effectiveness of disease prediction and treatment strategies.

1.8.6 Adaptability and scalability:

ML models can adapt and learn from new data over time, continuously improving their predictive ability. In addition, ML techniques are scalable, enabling the analysis of large-scale data sets and the implementation of predictive models in real clinical settings.

1.8.7 Multi-Omics data integration:

ML algorithms can integrate multi-omics data, including genomics, proteomics, metabolomics and transcriptomics data to discover biomarkers, disease pathways and molecular signatures associated with disease risk, progression and response to therapy.

Overall, the ability of ML models to handle complex data, capture non-linear relationships, automate feature selection and enable personalized predictions makes them highly effective and superior to traditional statistical methods in disease prediction. (Tawfik et al.,2021)

CHAPTER 2

REVIEW OF LITERATURE

- Al Rifai et al.,2021 the article provides a comprehensive overview of machine learning techniques for predicting heart disease. It likely includes a wide range of machine learning algorithms, data sources, and scoring methods used in heart disease prediction.
- Alizadehsani et al., 2018 presents a comparative review of different heart disease prediction models developed using machine learning techniques. The performance, accuracy, and applicability of different machine learning models for predicting heart disease are likely to be compared.
- Ammar et al.,2021to conduct a comparative study of machine learning techniques to predict diabetes. The paper will likely compare the performance and accuracy of different machine learning algorithms for predicting the onset or progression of diabetes.
- Barros et al., 2020 discusses machine learning models developed specifically for breast prediction. cancer It is likely to involve the development, evaluation and comparison of different machine learning models in predicting the occurrence of breast cancer.
- Chen et al.,2021 presents a machine learning-based disease prediction system targeting diabetes, heart disease and breast cancer. The paper describes the design, implementation and evaluation of a probabilistic prediction system and demonstrates its utility for risk assessment and early detection in healthcare.
- Gupta et al.,2020 focuses on predictive modeling for multi-disease risk assessment using machine learning techniques. It will likely address the

integration of machine learning algorithms, clinical data and risk factors to develop accurate and reliable models for risk assessment of various diseases.

- Islam et al., 2021 presents a comparative study of probabilistic machine learning algorithms for diabetes prediction. It can compare the performance of different algorithms, such as logistic regression, decision trees, support vector machines or neural networks, in terms of accuracy, sensitivity, specificity and other evaluation measures to determine the most effective method for predicting diabetes.
- Islam et al., 2020 paper provides a systematic review of machine learning techniques used to predict diabetes. The strengths and weaknesses of different machine learning algorithms in predicting diabetes are likely to be evaluated based on clinical and physiological data.
- Johnson et al., 2019 introduces a machine learning-based disease prediction system specifically targeting diabetes, heart disease and breast cancer. The paper describes the development of a probabilistic prediction system, including algorithm selection, feature design and model evaluation, demonstrating its effectiveness in disease prediction tasks.
- Khan et al., 2020 compare and analyze different machine learning techniques to predict heart disease. The work will likely examine the effectiveness and efficiency of different algorithms to accurately predict heart disease risk.
- Kim et al., 2020 developed of breast cancer risk prediction models using machine learning techniques. This may include, for example, data pre-processing, feature selection, model training and validation to create accurate and reliable breast cancer risk assessment tools to aid in early detection and personalized health interventions.

- Montan,1992 Montano's article focuses specifically on the use of decision. induction of trees to predict heart disease. It introduces the application of decision tree algorithms in biomedical engineering, particularly in the prediction of heart disease.
- Pangtey et al.,2019 to compare different machine learning algorithms for breast cancer prediction. The work is likely to evaluate the performance, strengths and weaknesses of different machine learning models in predicting the occurrence of breast cancer.
- Patel et al., 2019 provides a comprehensive overview of machine learning methods for predicting several diseases. It is likely to cover various machine learning algorithms, data sources and evaluation methods used in predictive disease research, providing an overview of the latest techniques and their applications in treatment.
- Roy et al., 2021 This article, similar to the first one, provides a comprehensive overview of machine learning methods to predict several diseases. Various algorithms, techniques and applications of machine learning in disease prediction will likely be discussed and the development of the field will be reviewed.
- Roy et al.,2020 provides an overview of machine learning approaches used to predict breast cancer. It will likely summarize the various machine learning techniques, feature selection methods, and model evaluation strategies used in breast cancer prognostic studies.
- Sarifuddin et al., 2021 developed and evaluated ensemble-based machine learning models likely tailored specifically for heart disease prediction. It can explore various techniques such as ensemble learning, feature engineering, and model

validation to improve the accuracy and reliability of heart disease prediction systems.

- Sudre et al., 2020 focuses on the use of machine learning algorithms to detect COVID-19-related abnormalities in chest X-rays. The study will likely involve training and validating machine learning models on a cohort of chest X-ray images to accurately identify abnormalities associated with COVID-19, contributing to improved diagnostic capabilities in the context of a pandemic.
- Tawfik et al. 2021 to the article of conducts a comparative study of machine learning algorithms to predict heart disease. The performance, accuracy and applicability of different machine learning models for predicting heart disease are likely to be compared.
- Viana et al., 2022 to compare and analyze machine learning algorithms for breast cancer prediction. The paper will likely evaluate the performance, strengths, and limitations of different machine learning models in predicting the occurrence of breast cancer.
- Wang et al., 2021 The article by Wang et al. focuses on predictive modeling using machine learning techniques, particularly for diabetes, heart disease and breast cancer. Likely details the methodology, data sources, and evaluation metrics used to develop and validate these disease prediction models, emphasizing their clinical applicability in healthcare settings.
- Wang et al., 2018 focuses on predictive modeling using machine learning techniques for diabetes, heart disease and breast cancer. It will likely address the development and evaluation of machine learning models adapted to predict these specific diseases, emphasizing the predictive power and accuracy of the models.

- White.,2018 book discusses machine learning approaches dedicated to predicting breast cancer. - risk This likely includes a range of machine learning models, feature selection methods and scoring methods adapted to breast cancer risk assessment.
- Yao et al., 2019 presents a systematic review of machine learning techniques used to predict breast cancer risk. Probably evaluates different machine learning algorithms and their effectiveness in predicting breast cancer risk based on different data sources and functions.
- Zou et al.,2019 presents a deep learning-based method for breast cancer prediction. This paper is likely to present a new approach that uses deep learning techniques to improve the accuracy and reliability of breast cancer prediction models.

CHAPTER 2

AIM AND OBJECTIVE

2.2 AIM:

The aim of this project is to develop a machine learning based system to predict several diseases such as diabetes, heart disease and breast cancer using algorithms such as logistic regression, decision tree and support vector machine (SVM).

2.3 OBJECTIVES:

- To retrieve disease datasets from the UCI ML database and import the relevant information of Heart disease, Diabetes, Breast Cancer.

- To preprocess the data and identify the parameters such as,

Heart disease: Age, Anaemia, Creatinine phosphokinase, Diabetes, Ejection fraction, High blood pressure, Platelets, Serum creatinine, Serum sodium, Sex, Smoking, Time.

Diabetes: Gender, Age, Urea, Creatinine, Cholesterol, Triglycerides, High-density lipoprotein, Low-density lipoprotein, Very low-density lipoprotein, Body Mass Index.

Breast cancer: Age, Body Mass Index, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP.

- To predict disease using machine learning algorithms like
 - Logistic Regression,
 - Support vector machine (SVM),
 - Decision tree.

CHAPTER 4

METHODOLOGY

4.1 DATA COLLECTION:

When collecting data from the UCI Machine Learning Repository or similar databases for disease prediction tasks related to diabetes, heart disease and breast cancer, it is important to follow a systematic approach to ensure data quality and relevance. Here is a section that describes the data collection process for each disease.

4.1.1 DIABETES:

For diabetes prediction, the data collection process involves using datasets from sources such as the UCI Machine Learning Repository Diabetes Database. This dataset typically includes characteristics such as glucose, insulin levels, BMI, age, family history and lifestyle factors. To train a strong machine learning model, it is very important to ensure the integrity, completeness and relevance of the data for the prediction task and to ensure that it is representative of different diabetics and non-diabetics.

4.1.2 HEART DISEASE:

For heart disease prediction, datasets from archives like for example the UCI Heart Disease Database provides valuable information such as blood pressure, cholesterol levels, ECG values, age, smoking and anamnesis. Data collection should focus on a comprehensive feature set that captures various risk factors and heart health indicators to develop an effective heart disease prediction model.

4.1.3 BREAST CANCER:

For breast cancer prediction, datasets such as the UCI Wisconsin Breast Cancer Datasets or similar repositories contain features. such as mammographic findings,

age, family history, hormonal factors and tumor characteristics. Collecting data from reliable sources with detailed annotations and quality assurance measures ensures that the dataset is suitable for training machine learning models to accurately predict breast cancer risk and outcomes.

In all cases, researchers and practitioners must follow privacy and ethical guidelines. obtain necessary permissions to use the data, and consider factors such as data imbalances, missing values, and data anomalies during data collection. Performing thorough exploratory analysis and validation steps can further improve the quality and applicability of the collected datasets for disease prediction.

4.2 LOAD THE DATASETS:

It is important to use reliable sources such as the UCI Machine Learning Repository or other reputable databases when downloading disease prediction databases related to diabetes, heart disease, and breast cancer. Here is a section that describes the process of downloading datasets for each disease.

4.2.1 DIABETES:

To predict diabetes, researchers can download diabetes data from the UCI Machine Learning Repository using Python libraries like pandas. Data typically includes characteristics such as glucose, insulin levels, BMI, age, family history and lifestyle factors. Once the dataset is loaded in CSV format, the pandas library can be used to read the CSV file and load the data into a DataFrame, allowing for easy processing and preprocessing of the data before training machine learning models.

4.2.2 HEART DISEASE:

For prediction of heart disease datasets such as UCI's heart disease datasets or related repositories can be loaded using pandas or other Python data loading libraries. This data set includes characteristics such as blood pressure, cholesterol levels, EKG readings, age, smoking status and medical history. By reading the dataset into a

DataFrame, researchers can perform data exploration, cleaning, and feature design to prepare the data for model training and evaluation.

4.2.3 BREAST CANCER:

For breast cancer prognosis, the UCI Wisconsin Breast Cancer datasets provide valuable data to load into Python environments. Using pandas or similar data loading tools, researchers can import a dataset that includes features such as mammography findings, age, family history, hormonal factors, and tumor characteristics. Loading data into a DataFrame allows researchers to preprocess data, handle missing values, and code categorical variables before training machine learning models to predict breast cancer.

In all cases, loading data into a DataFrame allows for efficient operation using Python libraries. data processing, research and preparation of machine learning tasks. Researchers should ensure data integrity, verify feature significance and follow best practices in data processing to maximize the predictive performance of machine learning models in disease prediction.

4.3 DATA PREPARATION:

Data preparation is a critical step in building machine learning models to predict diseases related to diabetes, heart disease and breast cancer. This section describes the data preparation process for each disease.

4.3.1 DIABETES:

To predict diabetes, data preparation involves processing features from the diabetes dataset, which typically include glucose, insulin levels, BMI, age, family history, and lifestyle factors. This process involves handling missing values using imputation methods such as mean, median or mode imputation based on characteristic distributions. In addition, outliers can be identified and addressed using methods such as Z-score or interquartile range (IQR) to ensure data integrity and model reliability. Categorical variables such as family history or smoking status may require coding

using techniques such as single-ray coding to be compatible with machine learning algorithms.

4.3.2 HEART DISEASE:

For heart disease prediction, data preparation is based on characteristics such as blood pressure, cholesterol levels, EKG readings, and age. , smoking status, and medical history from the heart disease dataset. As with the diabetes data, missing values are handled by imputation and outliers by appropriate statistical methods. Feature engineering can involve creating new features or transformations to capture interactions between variables and improve model performance. Categorical variables such as smoking status or medical history are coded for model fit and predictive accuracy.

4.3.3 BREAST CANCER:

For breast cancer prognosis, data preparation focuses on data such as mammography findings, age, family history, hormonal factors, and tumor characteristics of a dataset. such as the Wisconsin Breast Cancer Database. Data is cleaned by dealing with missing values and outliers, and feature engineering may involve creating derived functions or scaling numerical variables for consistency. Categorical variables such as tumor type or hormone receptor status are coded to ensure that they contribute significantly to predictive models.

In general, data preparation for diabetes, heart disease, and breast cancer prediction tasks involves dealing with data quality issues and missing values. and outliers, coding categorical variables, and dataset preparation for machine learning models. A thorough data preparation process improves the predictive power and generalizability of machine learning models, leading to more accurate and reliable disease prediction results.

4.4 SPLIT THE DATA:

Dividing the data into training and test sets is a crucial step in developing a machine learning model to predict diseases related to diabetes, heart disease and breast cancer. Here is a section that describes the data splitting process for each disease.

4.4.1 DIABETES:

To predict diabetes, the diabetes dataset is split into training and test sets using techniques such as `train_test_split` from the `sklearn.model_selection` Python module. Typically, about 80% of the data is allocated to the training, while the remaining 20% is used to test the performance of the model. This split ensures that the model is trained on a large enough dataset to learn patterns and relationships, but allows independent evaluation of unseen data to assess the accuracy and generalizability of its predictions.

4.4.2 HEART DISEASE:

Similar to heart disease prediction, the heart disease dataset. is divided into training and test set using `train_test_split` or other data splitting methods. The training set contains the majority of the data (e.g. 80%) used to train machine learning models (e.g. decision trees, support vector machines or logistic regression models). A test set (eg 20%) is kept separate and used to evaluate the performance of trained models in accurately predicting heart disease outcomes.

4.4.3 BREAST CANCER:

For breast cancer prediction, datasets such as the Wisconsin Breast Cancer dataset are shared. to training and test series following the same principles. The training set is used to train machine learning models using features such as mammography findings, age, family history, hormonal factors and tumor characteristics. The test set is then used to evaluate the performance of the trained models in predicting breast cancer outcomes (such as malignancy or tumor type) based on unseen patient data.

By splitting the data into training and test sets, machine learning practitioners can effectively develop and evaluate predictive models, ensuring that the models generalize well to new data and perform accurately in real cases. In disease prediction scenarios.

4.5 INITIALIZE AND TRAIN:

Running and training machine learning models is an integral part of disease prediction tasks for diabetes, heart disease and breast cancer. This chapter presents the initial and training models for each disease.

4.5.1 DIABETES:

To predict diabetes, machine learning models such as logistic regression, decision trees or support vector machines (SVM) are initialized using libraries such as Python's scikit-learn. These models are trained on a diabetes dataset that includes features such as glucose, insulin levels, BMI, age, family history and lifestyle factors. During training, the models learn patterns and relationships in the data to make predictions about the probability of diabetes based on the inputs. Training involves optimizing model parameters using algorithms such as gradient descent or kernel methods to minimize prediction errors and improve model accuracy.

4.5.2 HEART DISEASE:

Similarly, machine learning models such as decision trees, random forests or ensemble methods are used to predict heart disease. initialized and trained using the heart disease datasets. Features such as blood pressure, cholesterol levels, EKG readings, age, smoking status and medical history are used to train models to accurately predict the presence or absence of heart disease. Model training involves fitting data to models, fitting hyperparameters, and evaluating performance metrics such as precision, accuracy, recall, and F1 scores to assess model performance and reliability.

4.5.3 BREAST CANCER:

Machine learning models such as logistic regression for breast cancer prediction. , support vector machines (SVMs) or neural networks are initialized and trained using the Wisconsin breast cancer dataset. Features such as mammographic findings, age, family history, hormonal factors and tumor characteristics are used to train models to predict breast cancer outcomes, such as malignancy or tumor type. Model training involves fine-tuning model parameters, handling class imbalances, and optimizing performance metrics to ensure accurate and reliable predictions of breast cancer diagnosis and prognosis.

In general, initialization and training of machine learning models involves selection of appropriate algorithms and data preprocessing and optimization. model parameters and model performance evaluation to develop efficient and accurate disease prediction models for diabetes, heart disease and breast cancer.

4.6 MODEL EVALUATION:

Evaluating machine learning models is critical to assessing their effectiveness and reliability in predicting diseases such as diabetes, heart disease and breast cancer. This chapter presents the process of evaluating models for each disease.

4.6.1 DIABETES:

Machine learning models trained to predict diabetes, such as logistic regression, decision trees or support vector machines (SVMs), are evaluated using metrics such as accuracy, precision and recall. and F1 points. These metrics measure the model's ability to correctly predict diabetic and nondiabetic cases based on input characteristics such as glucose levels, insulin levels, BMI, age, family history, and lifestyle factors. The evaluation process involves comparing model predictions with actual results from a test data set to determine the overall performance and prediction accuracy of the model. In addition, techniques such as cross-validation can be used to confirm the robustness and generalizability of the model across different patient populations.

4.6.2 HEART DISEASE:

Similarly, machine learning models such as decision trees, random forests or ensemble methods can be used to predict heart disease evaluated by performance metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic (ROC) curve (AUC). These metrics evaluate the model's ability to correctly identify patients with heart disease based on characteristics such as blood pressure, cholesterol levels, EKG values, age, smoking status, and medical history. Model evaluation includes analysis of classification reports, confusion matrices and ROC curves to understand the strengths and weaknesses of the model and discriminatory power to distinguish between affected and unaffected individuals.

4.6.3 BREAST CANCER:

Machine learning models such as logistic regression, support vector for breast cancer prediction machines (SVMs) or neural networks are evaluated using metrics such as accuracy, precision, recall, and F1 score, along with characteristics such as mammographic findings, age, family history, hormonal factors, and tumor characteristics. The evaluation process involves the analysis of classification reports, confusion matrices, and ROC curves to analyze the effectiveness of the model in predicting breast cancer outcomes such as malignancy or tumor type. In addition, techniques such as sensitivity analysis or feature importance analysis can provide insight into the model's predictive power and feature contribution.

In Summary Evaluating machine learning models to predict diabetes, heart disease, and breast cancer uses various performance metrics and analysis techniques to assess predictive accuracy, reliability, and generalizability. The evaluation results inform health professionals and researchers about the effectiveness of the model in predicting diseases and guide decisions to refine and improve the model.

4.7 PREDICTION:

Printing a classification report is a critical step in evaluating machine learning models for disease prediction tasks, including diabetes, heart disease, and breast cancer.

Here is a section that describes the process of printing each disease classification report.

4.7.1 DIABETES:

To predict diabetes, after training a machine learning model (such as logistic regression, decision trees, or support vector machines (SVM)) on the diabetes dataset, the model is evaluated against a test set. The classification report is generated by the `classification_report` function of the `sklearn.metrics` Python module. This report provides comprehensive metrics such as precision, recall, F1 score and support for both diabetic and non-diabetic classes. The classification report helps to understand the performance of the model in correctly classifying diabetic and non-diabetic cases, highlighting areas for improvement and possible model biases.

4.7.2 HEART DISEASE:

Similarly for predicting heart disease, machine learning models such as decision trees, random forests or ensemble methods are evaluated using the test suite's `classification_report` function. The report includes measures such as precision, accuracy, recall and F1 score to accurately predict heart disease. This provides insight into the model's ability to distinguish between people with and without heart disease. This helps healthcare professionals evaluate the predictive power and reliability of the model.

4.7.3 BREAST CANCER:

For breast cancer prediction, machine learning models such as logistic regression, support vector machines (SVM) or neural networks are evaluated using the classification report of the test set, which includes features such as mammographic findings, age, family history, hormonal factors and tumor characteristics. The report presents metrics such as precision, accuracy, recall and F1 score for predicting breast cancer outcomes such as malignancy or tumor type. This helps evaluate the model's performance in correctly classifying breast cancer cases and provides practical insights to improve prediction accuracy and model interpretability.

In Summary Printing the diabetes, heart disease, and breast cancer task classification report provides detailed performance- metrics for model checking evaluation, optimization and decision making in healthcare facilities. Report metrics facilitate assessment of model strengths, weaknesses, and overall predictive accuracy, leading to informed decisions in disease diagnosis, risk assessment, and patient care.

4.8 CREATE OF WEB PAGE:

4.8.1 LOADING MACHINE LEARNING MODELS:

You download the machine learning models (diabetes model, breast cancer model and heart disease model) using the pickle library. These models have been trained to predict diabetes, breast cancer and heart disease.

4.8.1.1 IMPORT PICKLE LIBRARY:

Start by importing the pickle library into your Python script. This library allows Python objects to be serialized and deserialized, making it suitable for storing and loading trained machine learning models.

4.8.1.2 LOAD THE MODELS:

Use the 'pickle.load' function to load each machine learning model from a saved file. Make sure you enter the correct file paths where your trained models will be stored.

Load the diabetes prediction model from its save file.

Load the breast cancer prediction model from its save file.

Load the heart disease prediction model from its save file.

4.8.1.3 USE THE LOADED MODELS:

Once the models are loaded successfully, you can use them to make predictions based on the input data. This can be done in your Streamlit application or in any other Python script where you want to use these models for prediction tasks.

4.8.2 SIDEBAR OPTION MENU:

You can create a sidebar option menu in Streamlit using the `st.sidebar` context manager with the `option_menu` function from the `streamlit_option_menu` library. This approach allows you to offer users a choice of forecast types in the sidebar section of your app.

4.8.2.1 CREATE THE SIDEBAR OPTION MENU:

In the `'st.sidebar'` context manager, use the `'option_menu'` function to define a menu of options. This function takes arguments such as a menu title, a list of options, optional icons for each option, and a default directory.

4.8.2.2 HANDLE SELECTED OPTION:

After creating the settings side menu, you can use the selected option to determine which forecast type is selected. Based on the selected option, you can display the corresponding input fields and perform prediction logic for that type.

When the user selects "Diabetes Prediction", you can display the diabetes prediction and diabetes-related input fields. perform diabetes prediction logic.

If "Breast Cancer Prediction" is selected, the input fields related to breast cancer prediction will be displayed and the corresponding prediction logic will be performed.

Similarly, "Heart Disease Prediction" will display the corresponding input fields and perform the heart disease prediction logic.

Be sure to replace the placeholder logic with the real one with input fields and prediction logic according to your application requirements.

Finally run the Streamlit application (`streamlit run your_script.py`) observe the side menu function in action, allowing users to smoothly select the type of prediction they want.

4.8.3 USER INPUT AND PREDICTION:

4.8.3.1 DIABETES PREDICTION:

To predict diabetes, user data such as gender, age, urea, creatinine, HbA1c, cholesterol, triglycerides, HDL, LDL, VLDL and BMI must first be collected. These inputs are crucial because they provide the machine learning model with the information it needs to make predictions. It is important to ensure that the user provides valid numerical values for these inputs to avoid possible errors during the prediction process.

Once the user inputs are collected, the next step is to validate the input data and convert it into a table of numbers. This validation step is critical to maintain data integrity and ensure that the input data is in the required format for the machine learning model. Incorrect or missing values must be handled appropriately, such as displaying an error message to prompt the user to enter valid numerical values.

After validating and transforming the input data, a diabetes prediction model (`diabetes_model`) is used to predict whether a person has diabetes or not. The model uses input characteristics such as gender, age, urea, creatinine, HbA1c, cholesterol, triglycerides, HDL, LDL, VLDL and BMI to make predictions based on its trained algorithm. The prediction result is binary, where the model predicts either "the person is diabetic" or "the person is not diabetic" based on the input data.

Finally, the diabetes prediction process involves collecting, validating and transforming user inputs. then using the `diabetes_model` to predict and display the result of the

prediction to show whether the person is diabetic or not. This structured approach ensures accuracy and reliability in predicting diabetes status based on the input data provided.

4.8.3.2 HEART DISEASE PREDICTION:

To predict heart disease, the first step involves collecting user data such as age, anemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelet count, serum creatinine, serum sodium, gender, smoking and time. These contributions are critical because they include a variety of physiological and health-related factors that contribute to assessing heart disease risk. To ensure the accuracy of the forecasting process, it is very important to collect accurate and complete numeric values for these inputs.

After collecting the user's inputs, the next step is to validate the input data and convert it to numeric. a table Validation is required to ensure that the entered data is in the correct format and follows the expected data types required by the heart disease prediction model (heart_disease_model). Any inconsistencies or incorrect entries should be corrected to avoid possible errors in the prediction phase.

Once the input data is validated and converted to a numeric matrix, the heart_disease_model is used to predict whether or not a person has heart disease. Using input characteristics such as age, anemia, creatine phosphokinase, diabetes, ejection fraction, hypertension, platelets, serum creatinine, serum sodium, sex, smoking and time, the model uses its trained algorithm to make predictions about the human heart. a medical condition. The prediction result is binary, and the model reports either "the person has heart disease" or "the person does not have heart disease" based on the input data provided.

In summary, the process of predicting heart disease involves gathering a user input, validating and converting them into an appropriate format, using a heart disease model for prediction, and presenting the prediction result to indicate whether or not a person is likely to have heart disease. This structured approach ensures accuracy and

reliability in predicting heart disease risk based on comprehensive input parameters provided by the user.

4.8.3.3 BREAST CANCER PREDICTION:

The first step in breast cancer prognosis involves collecting user data such as age, BMI, glucose, insulin, HOMA, leptin, adiponectin, resistin and MCP1. These variables play a crucial role in predicting breast cancer risk because they represent important biological and physiological factors. To facilitate accurate forecasting, it is essential to ensure that users provide accurate and valid numerical values for these inputs.

After collecting user inputs, the next step is to validate and convert the input data into a spreadsheet. This validation process is necessary to maintain data integrity and ensure that the input data conforms to the expected format of the breast cancer prediction model (breast_cancer_model). Any inconsistencies or incorrect values must be corrected to avoid false predictions.

After the input data is validated and transformed into a numerical matrix, the breast cancer model is used to predict whether a person will have breast cancer or not. Using input features such as age, BMI, glucose, insulin, HOMA, leptin, adiponectin, resistin and MCP1, the model uses its trained algorithm to predict an individual's breast cancer status. The prediction result is binary and the model predicts either "the person has breast cancer" or "the person does not have breast cancer" based on the given input data.

Summary the process of predicting breast cancer involves collecting user inputs. , validating and converting them into an appropriate format, using a breast cancer model for prediction, and reporting the prediction result to determine whether a person is likely to have breast cancer or not. This systematic approach ensures accuracy and reliability in predicting breast cancer risk based on user-supplied input parameters.

4.8.4 DISPLAY PREDICTION RESULTS:

After machine learning models make predictions about diabetes, breast cancer and heart disease, the next important step is to display those results clearly and comprehensibly to the user in the Streamlit app. Streamlit provides convenient functions such as `st.success` and `st.error` to visually display prediction results based on model predictions.

The `St.success` function is used to display successful results or positive predictions. For example, if a model predicts that a person will have diabetes, breast cancer, or heart disease, the `st.success` function can be used to display that result. This feature usually highlights successful predictions in a visually appealing and affirming way, reinforcing positive results for the user.

On the other hand, the failure feature displays failed results or negative predictions. If the model predicts that a person does not have diabetes, breast cancer, or heart disease, `st.error` can be used to return that result. This feature is intended to visually distinguish negative predictions or errors and clarify the prediction results.

In Streamlit, these features are integrated into the corresponding prediction sections for diabetes, breast cancer and heart disease. Depending on the prediction result obtained by the corresponding machine learning model, either `st.success` or `st.error` is called to display the corresponding message. This approach improves the user experience by presenting prediction results in a simple and easy-to-understand format.

In addition, it helps users use distinct visual cues, such as green for successful predictions and red for failed predictions (or errors), quickly interpret the prediction results without ambiguity. This visual feedback mechanism increases the overall usability and efficiency of the Streamlit application by ensuring that users can easily understand and interpret predictions from machine learning models.

This method describes how users can interact with your Streamlit app to make predictions about various diseases using machine learning models.

4.8.4.1 DIABETES PREDICTION:

After making a prediction with the `diabetes_model`, you can use conditional statements to determine the result of the prediction. If the model predicts that a person has diabetes, you can use the `'st.success'` function to display a success message, such as "The person is diabetic." On the other hand, if the prediction is negative (that is, the person is not predicted to have diabetes), you can use the `'st.error'` command to display an error message such as "The person is not diabetic." This approach ensures that users get a direct and understandable forecast result.

4.8.4.2 HEART DISEASE PREDICTION:

Heart disease prediction, if the `heart_disease_model` predicts whether a person has heart disease or not, you can use the `'st.success'` function to display "The person has heart disease" if the prediction is positive. If the prediction is negative (ie. the person is not predicted to have heart disease), you can use `'st.error'` to show "the person does not have heart disease". This method ensures that users receive immediate feedback about their heart disease risk based on input data.

4.8.4.3 BREAST CANCER:

You can apply the same approach to predicting breast cancer after the breast cancer model has made a prediction. If the model predicts that a person has breast cancer, you can use the `'st.success'` function to display "the person has breast cancer". Conversely, if the prediction indicates that the person does not have breast cancer, you can use `'st.error'` to indicate "the person does not have breast cancer". This clear and concise presentation of forecast results improves user understanding and engagement with your Streamlit application.

By properly incorporating `'st.success'` and `'st.error'` functions based on diabetes, heart disease, and breast cancer prediction results, you can provide a smooth and informative user experience. This visual representation of predictive results adds value to your application by facilitating quick interpretation and decision making for users concerned about these health conditions.

CHAPTER 5

RESULT AND DISCUSSION

5.1 RESULTS:

Disease Prediction Using Machine Learning (ML) Algorithms Disease prediction using machine learning algorithms provides valuable insights into the prediction of several diseases such as Diabetes, Heart disease, and Breast Cancer. Let's explore the results and discussion of disease predictions using machine learning.

The goal of the project is to create a machine learning based system for the prediction of multiple diseases such as Diabetes, Heart disease, and Breast Cancer using algorithms like Logistic regression, Decision tree, and Support vector machine. The system will be implemented in the web page using Streamlit library. Users will be able to enter their health parameters and get predictions about their health status.

5.2 DIABETES PREDICTION:

ML algorithms such as logistic regression, support vector machine (SVM) and decision trees are trained using a wide range of features such as gender, age, urea, creatinine, HbA1c, cholesterol, triglycerides, HDL, LDL, VLDL and BMI. predicts the probability of diabetes. These algorithms analyze the relationships between these characteristics and the target variable (diabetic status) to make accurate predictions. The results obtained by ML models provide a clear picture of an individual's diabetes risk and predict things like "a person is diabetic" or "a person is not diabetic". This information is invaluable because it allows people to assess their health, take preventative steps and make informed decisions about their diabetes risk. Prediction of diabetes based on ML not only helps in early detection, but also facilitates preventive treatment through lifestyle changes, medical interventions and regular monitoring, thus reducing the risk of diabetes-related complications and improving overall health.

In diabetes prediction, the prediction result provides valuable information about an individual's diabetes risk. If the model predicts that a person is diabetic (ie a success ('The person is diabetic')), it indicates a higher probability of developing diabetes based on the input data, which can prompt the person for further evaluation and treatment. On the other hand, if the prediction is negative (ie. an error ('The person is not diabetic')), it indicates a lower risk of diabetes, which gives the individual confidence, but still emphasizes the importance of maintaining a healthy lifestyle in prevention. the disease future risks.

Next steps: [Generate code with diabetes_data](#) [View recommended plots](#)

	Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	CLASS
0	1	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24.0	0
1	0	26	4.5	62	4.9	3.7	1.4	1.1	2.1	0.6	23.0	0
2	1	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24.0	0
3	1	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24.0	0
4	0	33	7.1	46	4.9	4.9	1.0	0.8	2.0	0.4	21.0	0

	Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	CLASS
267	0	55	5.1	78	11.8	4.5	1.8	1.1	2.7	12.7	31.5	1
268	1	55	2.1	46	5.9	4.1	2.7	1.0	2.0	1.2	30.0	1
269	0	57	4.8	72	7.3	4.5	1.6	2.1	1.9	0.7	34.0	1
270	1	52	6.0	76	6.5	2.7	1.2	0.8	1.4	0.5	31.0	1
271	0	55	5.1	78	11.8	4.5	1.8	1.1	2.7	12.7	31.5	1

Fig 5.1

5.2.1 LOGISTIC REGRESSION:

Logistic regression is a commonly used algorithm for binary classification tasks and is particularly effective for predicting diseases with a binary outcome, such as the presence or absence of diabetes. When predicting diabetes, logistic regression uses

input characteristics such as gender, age, urea, creatinine, HbA1c, cholesterol, triglycerides, HDL, LDL, VLDL, and BMI to estimate a person's probability of diabetes.

One of the following options the main advantages of logistic regression is its ability to provide interpretable coefficients that provide insight into the effect of each input characteristic on the prediction result. This interpretability is valuable for understanding which factors significantly contribute to diabetes risk prediction.

Logistic regression results are easy to interpret and are usually presented as probabilities or binary predictions. For example, a model can output probabilities that indicate the probability of diabetes, and based on a selected threshold, it will classify individuals as "a person is diabetic" or "a person is not diabetic".

According to the performance of the models, an accuracy score of 83.6% means that the logistic regression model predicts diabetic status correctly in two thirds of the cases in the dataset. Although precision is an important metric, it is very important to consider other metrics such as precision, recall and F1 score to comprehensively evaluate model performance, especially for unbalanced data sets or when different types of errors have different consequences.

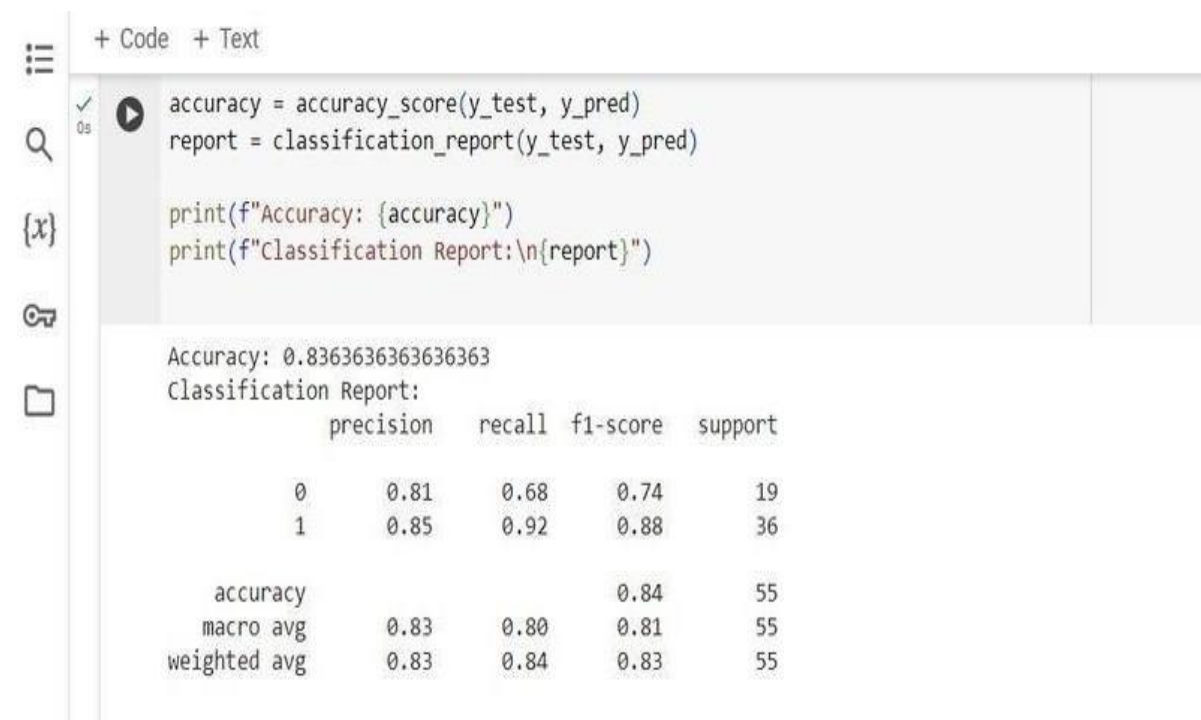


Fig 5.2

5.2.2 SUPPORT VECTOR MACHINE:

Support Vector Machine (SVM) is a robust algorithm suitable for both binary and multi-class classification tasks and is known for its effectiveness in capturing complex decision constraints. To predict diabetes, SVM uses a comprehensive set of features including gender, age, urea, creatinine, HbA1c, cholesterol, triglycerides, HDL, LDL, VLDL and BMI to estimate a person's probability of diabetes.

The SVM approach involves hyper level this best distinguishes different classes, such as "Person is diabetic" or "Person is not diabetic", from a high-dimensional feature space. Unlike linear models, SVM can handle non-linear relationships and capture complex patterns in the data, making it a valuable tool for disease prediction in complex data structures.

An accuracy score of 90% indicates that the SVM model reaches a high level. Accuracy of prediction of diabetes status in the dataset. This high accuracy suggests that the SVM model effectively captures the underlying patterns and relationships between the input characteristics and the target variable, allowing it to make accurate diabetes risk predictions. However, it is important to note that while accuracy is a critical metric, it is also important to consider other evaluation metrics and perform thorough validation to ensure model generalizability and reliability.

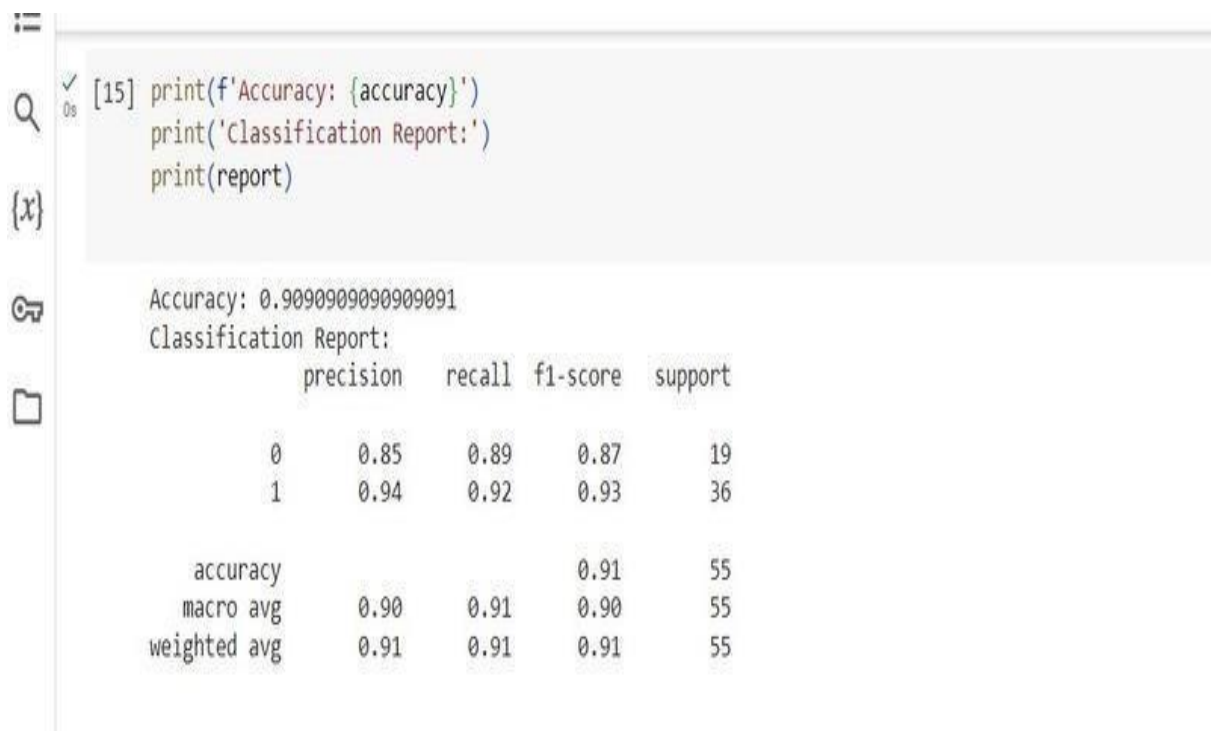


Fig 5.3

5.2.3 DECISION TREE:

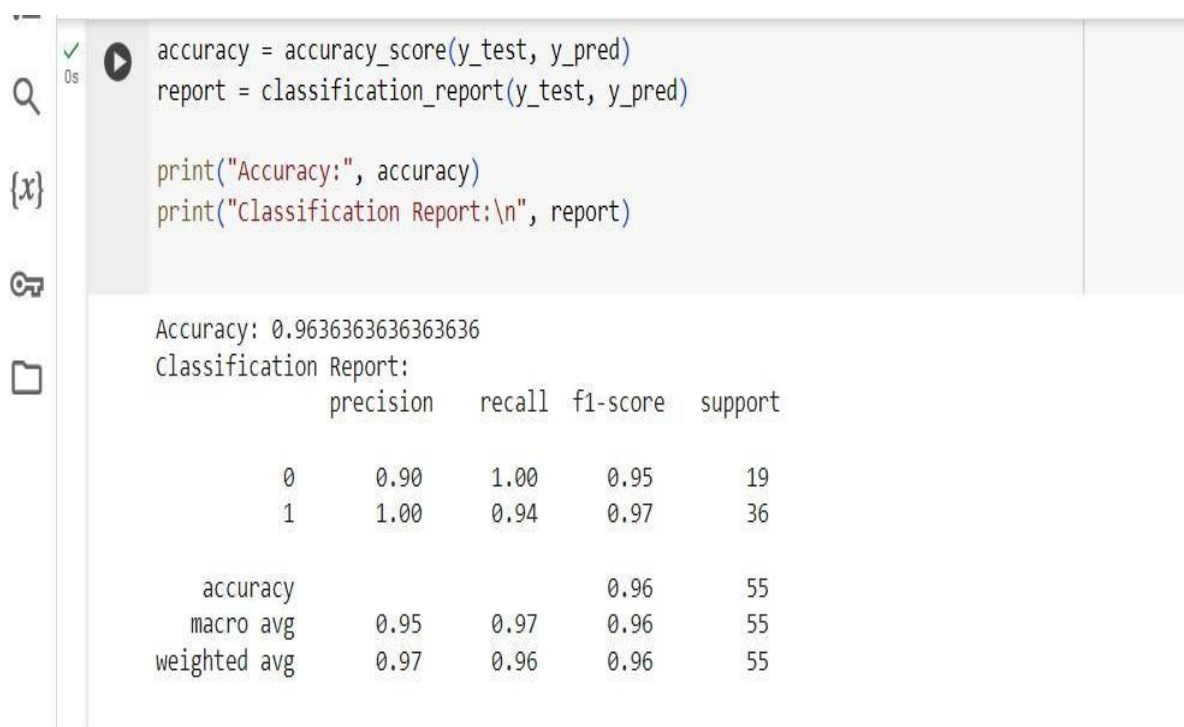
Decision trees are known for their intuitive nature and ease of interpretation, making them excellent choices for explaining predictive logic to non-technical users. In diabetes prediction, decision trees use input parameters such as gender, age, urea, creatinine, HbA1c, cholesterol, triglycerides, HDL, LDL, VLDL and BMI to estimate an individual's probability of diabetes.

One of the main features of decision trees is their ability to divide the feature space into regions based on threshold values, which creates a hierarchical structure of decision rules. This hierarchical structure allows a clear understanding of the sequence of decisions leading to the prediction result, whether "the person is diabetic" or "the person is not diabetic".

The accuracy score of 96% shows that the model of decision tree reaches high accuracy in predicting the diabetes status of the dataset. This high accuracy reflects the model's ability to effectively capture underlying patterns and relationships between

input characteristics and the target variable, allowing it to make accurate predictions about diabetes risk.

In addition, the results of decision trees are usually presented in a tree like format that not only aids interpretation, but also allows users to visualize and understand the model's decision-making process. This transparency and interpretability is valuable for health professionals and individuals as they understand and use the predictions of the decision tree model for diabetes risk assessment and management.



```
accuracy = accuracy_score(y_test, y_pred)
report = classification_report(y_test, y_pred)

print("Accuracy:", accuracy)
print("Classification Report:\n", report)
```

Accuracy: 0.9636363636363636

Classification Report:

	precision	recall	f1-score	support
0	0.90	1.00	0.95	19
1	1.00	0.94	0.97	36
accuracy			0.96	55
macro avg	0.95	0.97	0.96	55
weighted avg	0.97	0.96	0.96	55

Fig 5.4

5.3 HEART DISEASE PREDICTION:

The heart disease prediction model is designed to consider various input characteristics such as age, anemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, gender, smoking and time to accurately predict heart disease risk. After processing these features, the model generates predictive results that indicate whether a person has heart disease or not.

The Streamlit application displays these predictive results using success and failure functions, which gives people instant feedback on their cardiovascular health. Positive predictions that indicate a person has heart disease are an important warning that people should prioritize their cardiovascular health. It encourages them to seek medical evaluations, adopt heart-healthy lifestyles and follow personalized treatment plans to effectively manage and reduce their risk of heart disease.

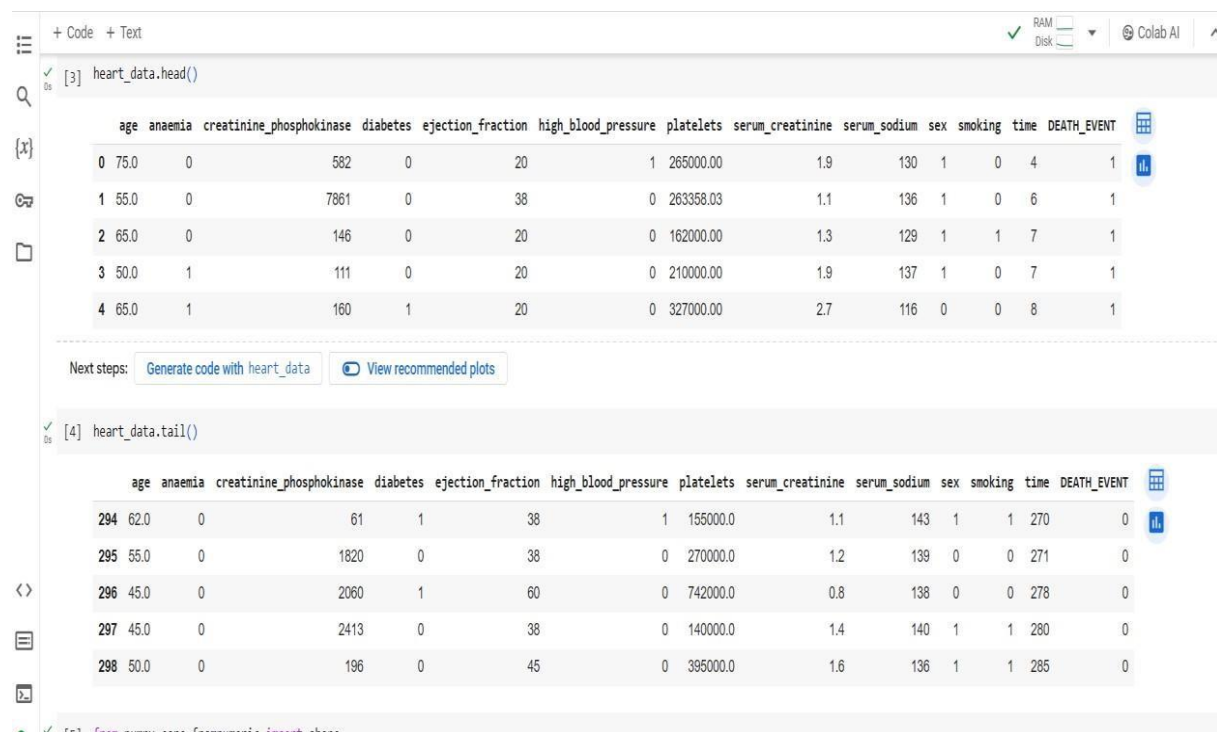


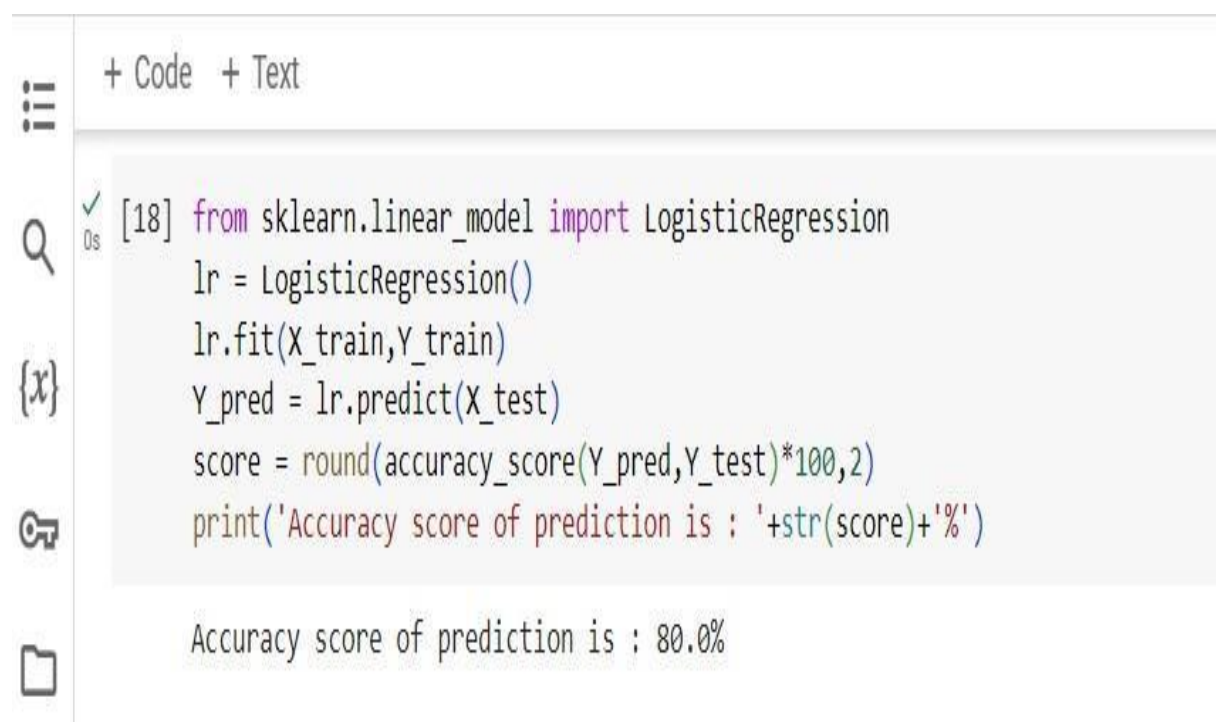
Fig 5.5

5.3.1 LOGISTIC REGRESSION:

Logistic regression is a binary classification algorithm that uses a set of input characteristics such as age, anemia, creatine phosphokinase, diabetes, ejection fraction, hypertension, platelets, serum creatinine, serum sodium, sex, smoking, and time to predict the probability of disease. heart disease One of the main advantages of logistic regression is its ability to provide interpretable coefficients that provide insight into the effect of each characteristic on the forecast outcome. This

interpretability allows a clear understanding of how each input characteristic affects prediction of heart disease risk.

With 80 percent accuracy, logistic regression proves its effectiveness in making accurate predictions of heart disease. Logistic regression results are presented in a straight line, either as probabilities indicating the likelihood of heart disease or as binary predictions ("the person has heart disease" or "the person does not have heart disease"). This clarity of interpretation makes logistic regression a valuable tool for health professionals and individuals to assess cardiovascular health risks and implement appropriate preventive measures.



The image shows a Jupyter Notebook interface. On the left, there is a sidebar with icons for a menu, search, variables, keyboard shortcuts, and a file explorer. The main area displays a code cell with the following Python code:

```
[18]: from sklearn.linear_model import LogisticRegression
      lr = LogisticRegression()
      lr.fit(X_train, Y_train)
      Y_pred = lr.predict(X_test)
      score = round(accuracy_score(Y_pred, Y_test)*100, 2)
      print('Accuracy score of prediction is : '+str(score)+'%')
```

Below the code, the output of the cell is displayed: "Accuracy score of prediction is : 80.0%".

Fig 5.6

5.3.2 SUPPORT VECTOR MACHINE:

Support Vector Machine (SVM) is a robust algorithm used for heart disease prediction and is known for its ability to efficiently handle complex decision constraints. Using the same input functions as age, anemia, creatine phosphokinase, diabetes, ejection fraction, hypertension, platelets, serum creatinine, serum sodium, sex, smoking, and time, SVM tries to find the hyperlevel that best discriminates individuals into

categories such as "person has heart disease" or "person does not have heart disease" in a high-dimensional feature space.

With an accuracy score of 75%, SVM demonstrates its ability to capture non-linear relationships and complex data patterns. This makes SVM a valuable tool for predicting heart disease, especially when dealing with complex and multivariate data. The algorithm's ability to handle non-linear relationships allows the discovery of hidden patterns and correlations that simpler models such as logistic regression may not capture.

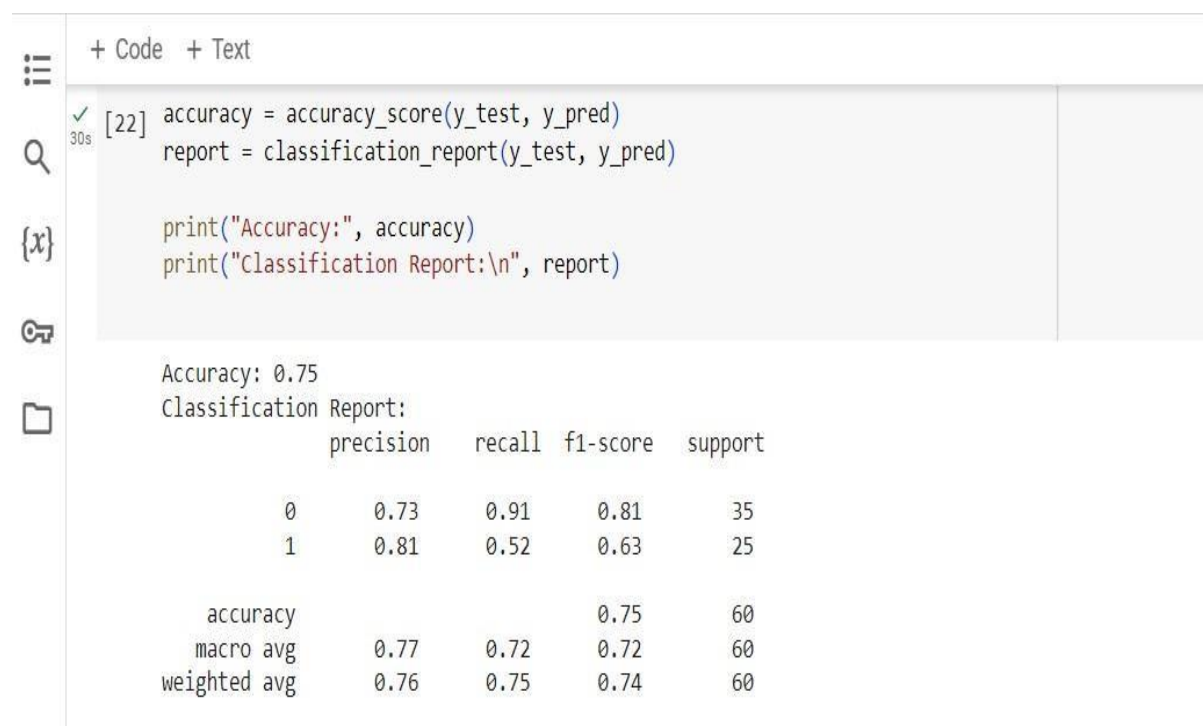


Fig 5.7

5.3.3 DECISION TREE:

Decision trees are known for their intuitive nature and ease of interpretation, making them a popular choice for predicting heart disease. Despite the 63% accuracy score, decision trees have several advantages, including the ability to divide the feature space into regions based on threshold values.

The tree-like structure of decision trees allows users to visualize and understand decision making. a process that leads to the prediction of heart disease risk. Each

node in the decision tree represents a decision based on a specific feature and a threshold value, leading to subsequent nodes until a final prediction is made ("the person has heart disease" or "the person does not have heart disease").

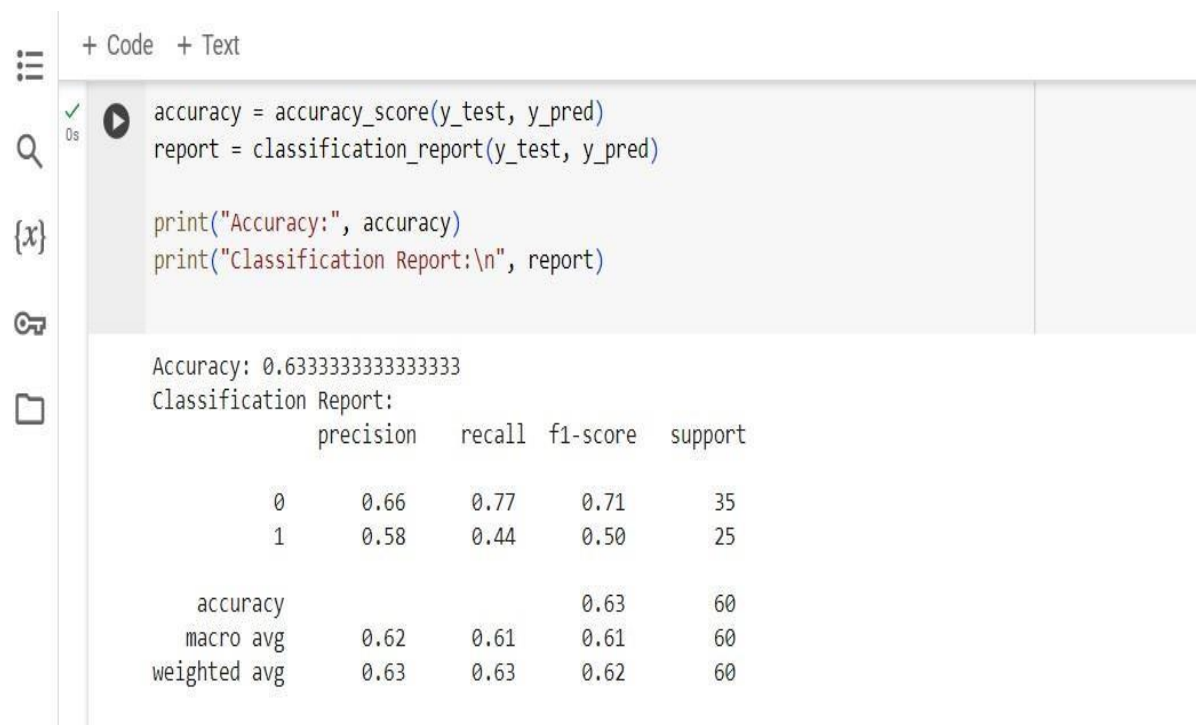


Fig 5.8

5.4 BREAST CANCER PREDICTION:

The breast cancer prediction model is designed to use input parameters such as age, BMI, glucose, insulin, HOMA, leptin, adiponectin, resistin and MCP1 to accurately predict breast cancer risk. These parameters provide important insights into an individual's health profile and contribute to a comprehensive estimate of breast cancer probability.

The Streamlit program presents the prediction results obtained from the breast cancer prediction model using Streamlit's success and failure functions. This user-friendly interface gives people clear and immediate feedback about their breast cancer risk status. Results are presented as "person has breast cancer" or "person does not have breast cancer," allowing people to make informed decisions about breast health.

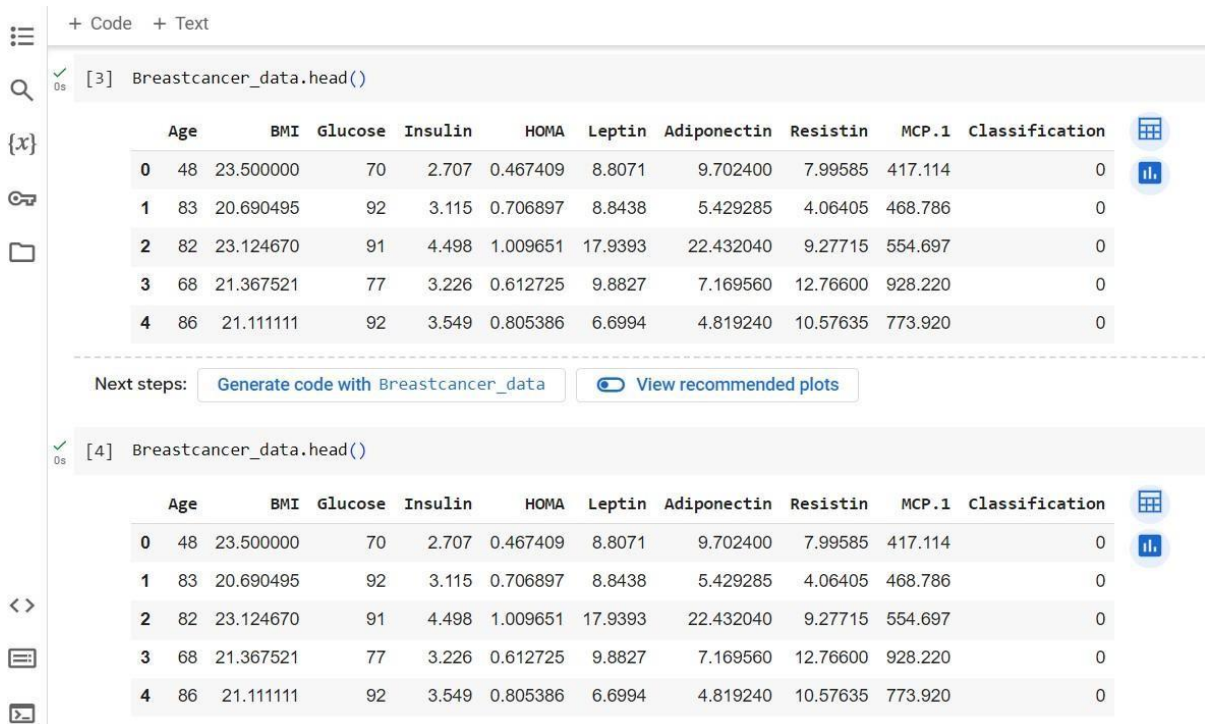


Fig 5.9

5.4.1 LOGISTIC REGRESSION:

Logistic regression is a robust algorithm widely used for binary classification tasks, which makes it particularly suitable for predicting diseases where an outcome is present or absent. In the context of breast cancer prognosis, logistic regression analyzes input parameters such as age, BMI, glucose, insulin, HOMA, leptin, adiponectin, resistin, and MCP1 to estimate the probability of breast cancer. One of the main advantages of logistic regression is its ability to provide interpretable coefficients that help understand how each input characteristic affects the prediction result. Logistic regression results are usually presented as probabilities or binary predictions ("a person has breast cancer" or "a person does not have breast cancer"). Despite its interpretability and utility, the accuracy score for breast cancer prediction using logistic regression is reported to be 75%.

The strength of logistic regression lies in its simplicity, interpretability and efficiency in dealing with binary classification tasks. However, its performance on complex and high-dimensional datasets may be limited compared to more advanced algorithms such as support vector machines (SVMs) or decision trees. However, logistic

regression is a valuable tool in disease prediction tasks because it provides useful information about breast cancer risk assessment and facilitates informed decision making.

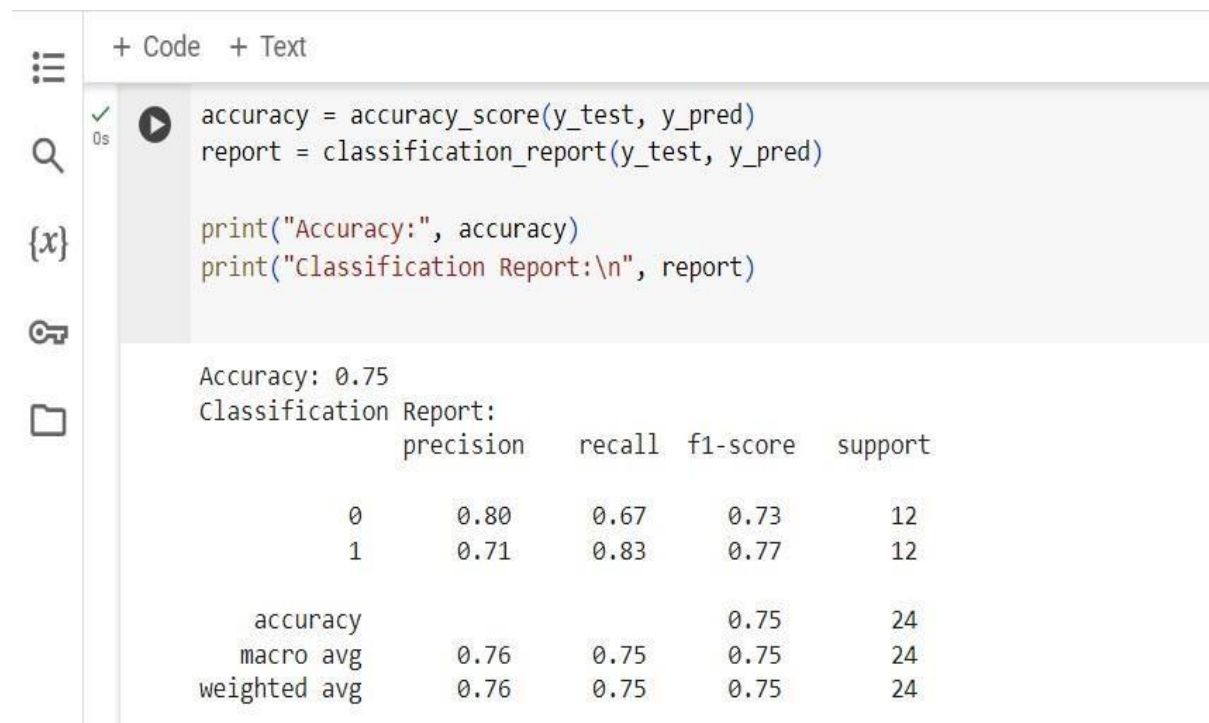
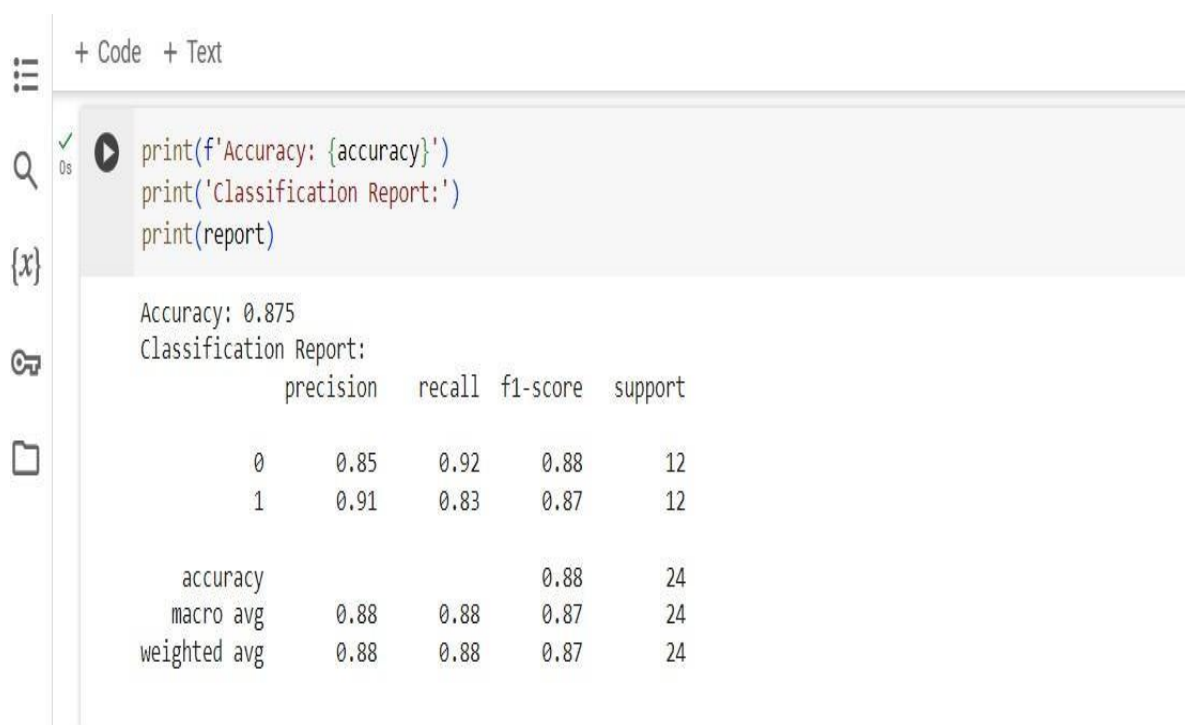


Fig 5.10

5.4.2 SUPPORT VECTOR MACHINE:

The Support Vector Machine (SVM) is a powerful algorithm that excels at both binary and multiclass classification tasks, making it particularly effective at capturing complex decision constraints. In the context of breast cancer prognosis, SVM uses the same input parameters as age, BMI, glucose, insulin, HOMA, leptin, adiponectin, resistin, and MCP1 to estimate the probability of breast cancer. Unlike logistic regression, which aims to find a linear decision boundary, SVM tries to find the hyperplane that best separates people into categories such as "a person has breast cancer" or "a person does not have breast cancer" in a high-dimensional feature space. This feature enables SVM to handle non-linear relationships and capture complex data patterns, making it a valuable tool for breast cancer prediction tasks.

SVM's ability to handle complex decision constraints and non-linear relationships increases its applicability to breast cancer prediction. exercises. SVM can effectively capture complex patterns in data and make more accurate predictions and provide valuable information for breast cancer risk assessment. Although SVM may require careful tuning of parameters and computational resources, its robustness and ability to handle high-dimensional data make it a valuable asset in health applications, especially disease prediction and risk assessment.



The screenshot shows a Jupyter Notebook interface with a code cell and its output. The code cell contains three lines of Python code: `print(f'Accuracy: {accuracy}')`, `print('Classification Report:')`, and `print(report)`. The output displays the accuracy as 0.875 and a detailed classification report table.

```
print(f'Accuracy: {accuracy}')
print('Classification Report:')
print(report)
```

```
Accuracy: 0.875
Classification Report:
              precision    recall  f1-score   support

     0       0.85        0.92        0.88        12
     1       0.91        0.83        0.87        12

 accuracy          0.88        0.88        0.88        24
  macro avg       0.88        0.88        0.87        24
 weighted avg     0.88        0.88        0.87        24
```

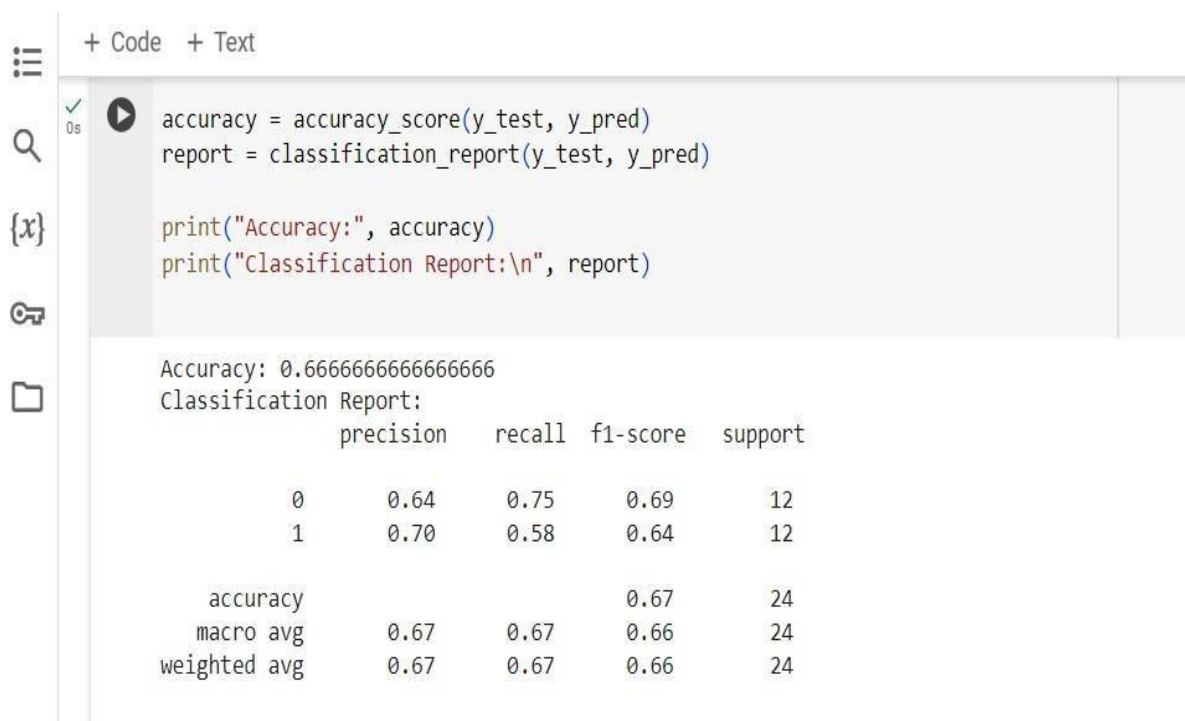
Fig 5.11

5.4.3 DECISION TREE:

Decision trees are known for their intuitive nature and ease of interpretation, making them well suited for explaining predictive logic to non-technical users. In the context of breast cancer prognosis, decision trees use the same input parameters as previously mentioned, such as age, BMI, glucose, insulin, HOMA, leptin, adiponectin, resistin, and MCP1, to predict breast cancer risk. These decision trees divide the feature space into discrete regions based on threshold values, creating a hierarchical structure of decision rules. Each node of the decision tree represents a decision based on a specific feature and a threshold value, which leads to the next nodes until the

final prediction is made ("the person has breast cancer" or "the person does not have breast cancer"). This tree-like format of decision tree results allows users to understand the sequence of decisions that led to a predictive outcome, improving transparency and interpretability.

Despite their simplicity, decision trees can capture complex relationships and patterns in data, contributing to their data. performance in breast cancer prediction tasks. The 66.6% percent accuracy score for breast cancer prediction using decision trees shows their ability to make reasonably accurate predictions. The interpretability and transparency of decision trees make them a valuable tool in healthcare applications, allowing healthcare professionals and individuals to gain insight into breast cancer risk assessment and make informed decisions about prevention and treatment.



The screenshot shows a Jupyter Notebook interface with a sidebar on the left containing icons for a menu, search, variables, keyboard shortcuts, and a file explorer. The main area is divided into a code editor and an output area. The code editor contains the following Python code:

```
+ Code + Text
accuracy = accuracy_score(y_test, y_pred)
report = classification_report(y_test, y_pred)

print("Accuracy:", accuracy)
print("Classification Report:\n", report)
```

The output area displays the results of the code execution:

```
Accuracy: 0.6666666666666666
Classification Report:
              precision    recall  f1-score   support

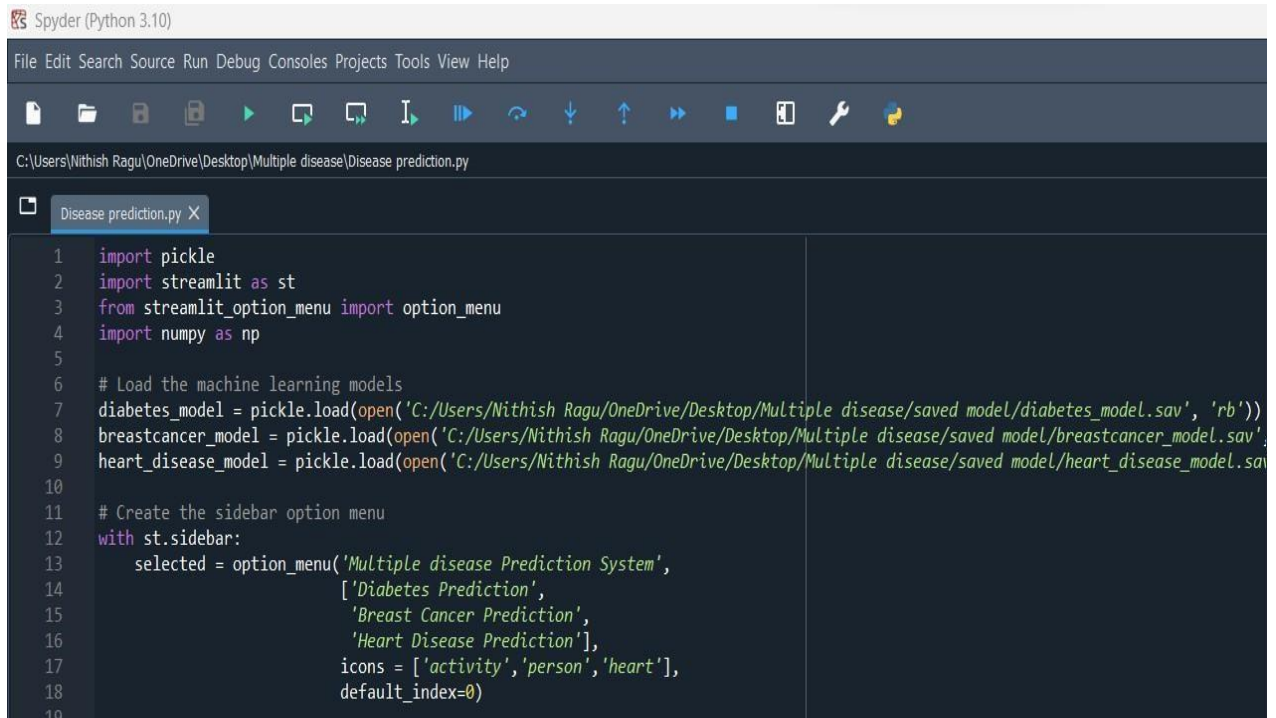
     0       0.64      0.75      0.69        12
     1       0.70      0.58      0.64        12

 accuracy          0.67
 macro avg         0.67      0.67      0.66        24
weighted avg         0.67      0.67      0.66        24
```

Fig 5.12

5.5 WEB DEVELOPMENT

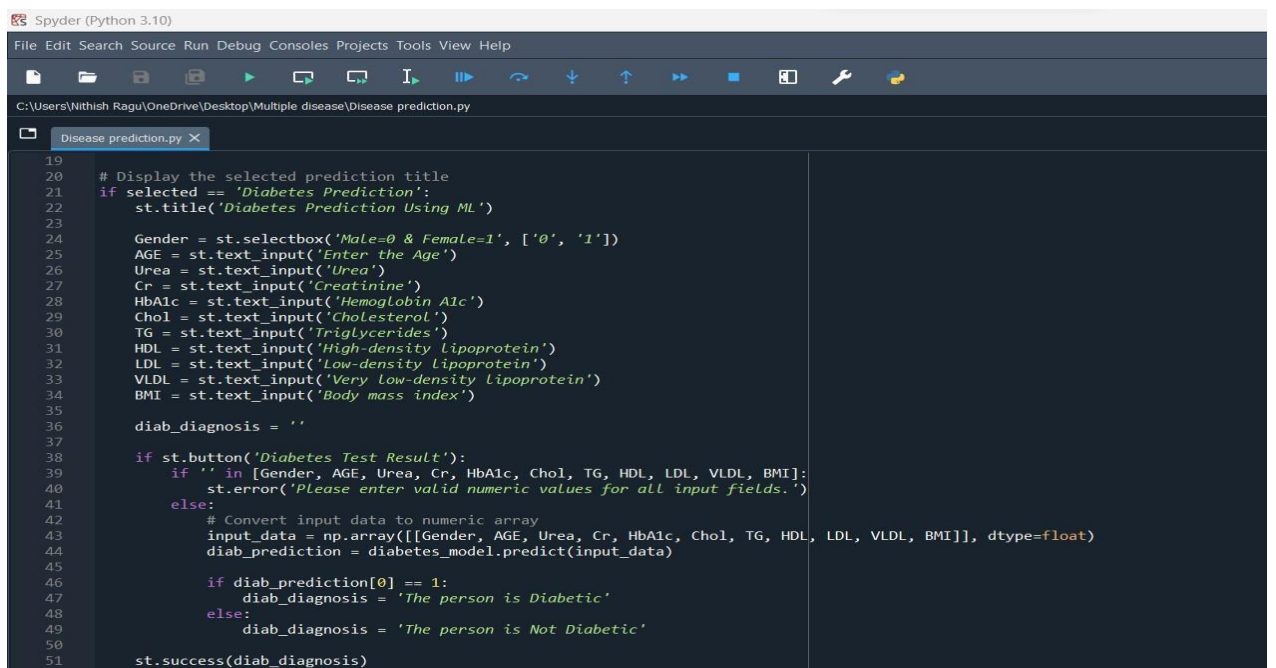
5.5.1 IMPORT LIBRARY AND FILE:



```
1 import pickle
2 import streamlit as st
3 from streamlit_option_menu import option_menu
4 import numpy as np
5
6 # Load the machine learning models
7 diabetes_model = pickle.load(open('C:/Users/Nithish Ragu/OneDrive/Desktop/Multiple disease/saved model/diabetes_model.sav', 'rb'))
8 breastcancer_model = pickle.load(open('C:/Users/Nithish Ragu/OneDrive/Desktop/Multiple disease/saved model/breastcancer_model.sav', 'rb'))
9 heart_disease_model = pickle.load(open('C:/Users/Nithish Ragu/OneDrive/Desktop/Multiple disease/saved model/heart_disease_model.sav', 'rb'))
10
11 # Create the sidebar option menu
12 with st.sidebar:
13     selected = option_menu('Multiple disease Prediction System',
14                           ['Diabetes Prediction',
15                            'Breast Cancer Prediction',
16                            'Heart Disease Prediction'],
17                           icons = ['activity', 'person', 'heart'],
18                           default_index=0)
```

Fig 5.13

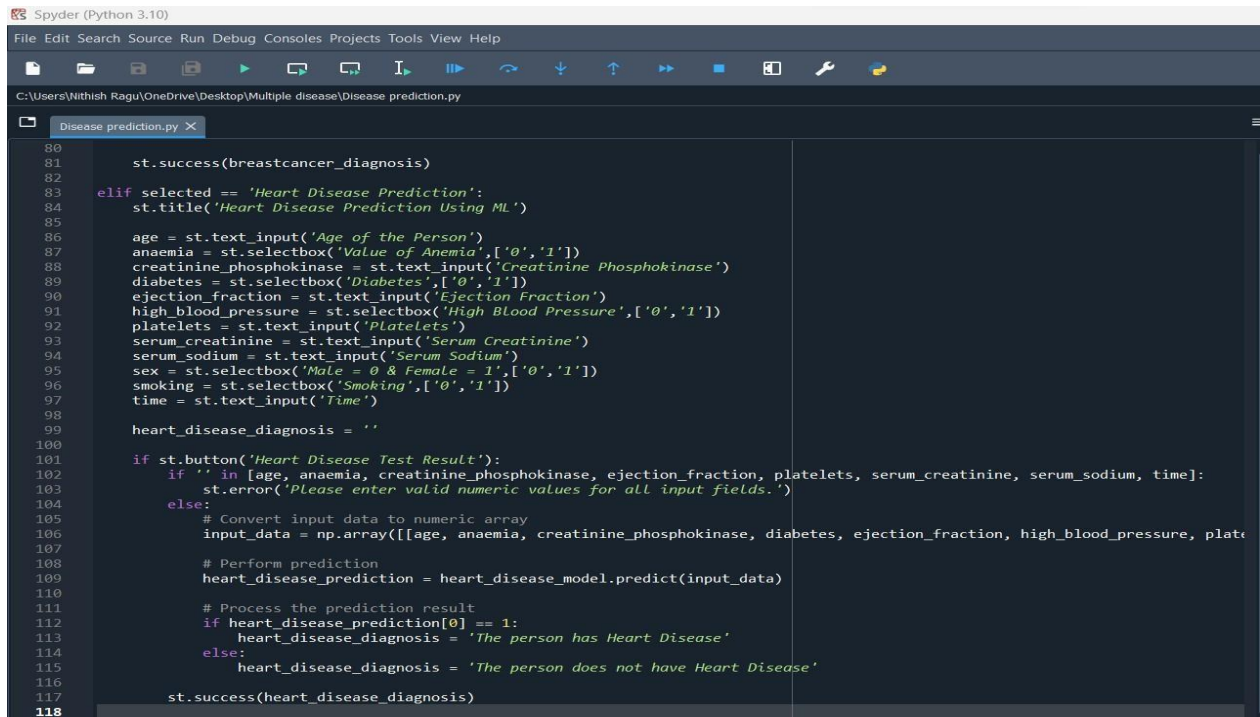
5.5.2 DIABETES:



```
19 # Display the selected prediction title
20 if selected == 'Diabetes Prediction':
21     st.title('Diabetes Prediction Using ML')
22
23     Gender = st.selectbox('Male=0 & Female=1', ['0', '1'])
24     AGE = st.text_input('Enter the Age')
25     Urea = st.text_input('Urea')
26     Cr = st.text_input('Creatinine')
27     HbA1c = st.text_input('Hemoglobin A1c')
28     Chol = st.text_input('Cholesterol')
29     TG = st.text_input('Triglycerides')
30     HDL = st.text_input('High-density Lipoprotein')
31     LDL = st.text_input('Low-density Lipoprotein')
32     VLDL = st.text_input('Very Low-density Lipoprotein')
33     BMI = st.text_input('Body mass index')
34
35     diab_diagnosis = ''
36
37     if st.button('Diabetes Test Result'):
38         if '' in [Gender, AGE, Urea, Cr, HbA1c, Chol, TG, HDL, LDL, VLDL, BMI]:
39             st.error('Please enter valid numeric values for all input fields.')
40         else:
41             # Convert input data to numeric array
42             input_data = np.array([[Gender, AGE, Urea, Cr, HbA1c, Chol, TG, HDL, LDL, VLDL, BMI]], dtype=float)
43             diab_prediction = diabetes_model.predict(input_data)
44
45             if diab_prediction[0] == 1:
46                 diab_diagnosis = 'The person is Diabetic'
47             else:
48                 diab_diagnosis = 'The person is Not Diabetic'
49
50     st.success(diab_diagnosis)
```

Fig 5.14

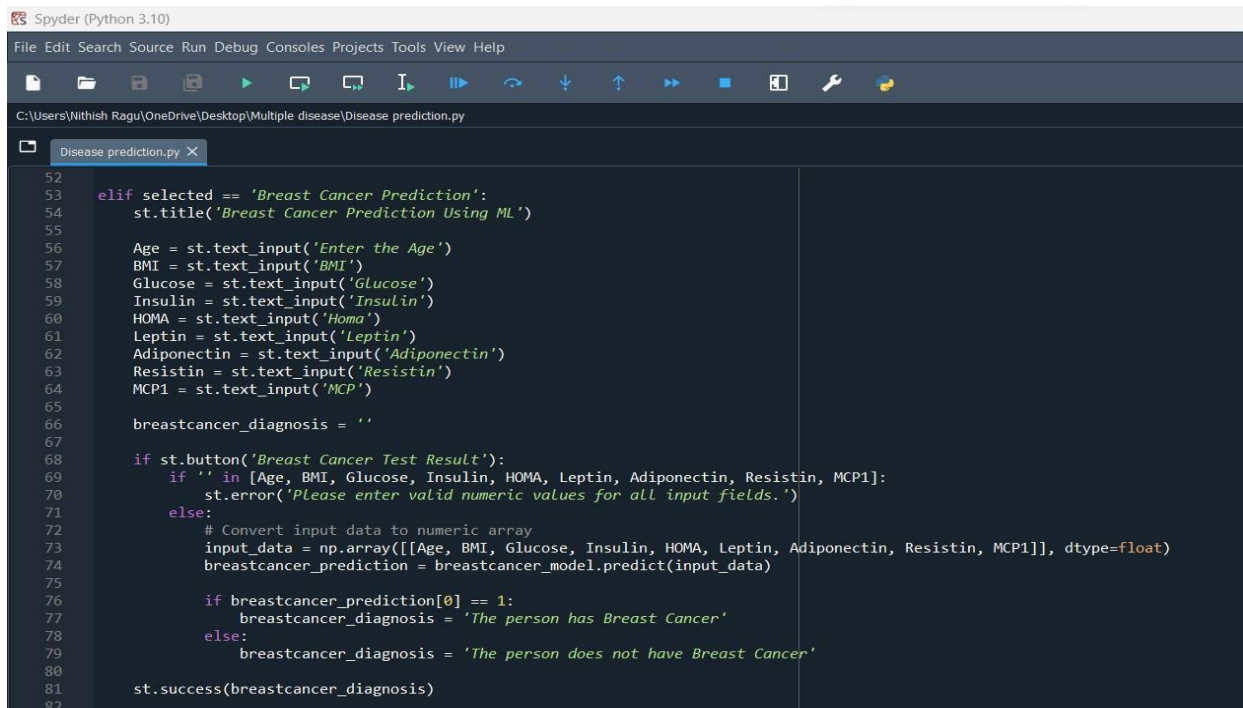
5.5.3 HEART DISEASE:



```
80     st.success(breastcancer_diagnosis)
81
82
83 elif selected == 'Heart Disease Prediction':
84     st.title('Heart Disease Prediction Using ML')
85
86     age = st.text_input('Age of the Person')
87     anaemia = st.selectbox('Value of Anemia', ['0', '1'])
88     creatinine_phosphokinase = st.text_input('Creatinine Phosphokinase')
89     diabetes = st.selectbox('Diabetes', ['0', '1'])
90     ejection_fraction = st.text_input('Ejection Fraction')
91     high_blood_pressure = st.selectbox('High Blood Pressure', ['0', '1'])
92     platelets = st.text_input('Platelets')
93     serum_creatinine = st.text_input('Serum Creatinine')
94     serum_sodium = st.text_input('Serum Sodium')
95     sex = st.selectbox('Male = 0 & Female = 1', ['0', '1'])
96     smoking = st.selectbox('Smoking', ['0', '1'])
97     time = st.text_input('Time')
98
99     heart_disease_diagnosis = ''
100
101 if st.button('Heart Disease Test Result'):
102     if '' in [age, anaemia, creatinine_phosphokinase, ejection_fraction, platelets, serum_creatinine, serum_sodium, time]:
103         st.error('Please enter valid numeric values for all input fields.')
104     else:
105         # Convert input data to numeric array
106         input_data = np.array([age, anaemia, creatinine_phosphokinase, diabetes, ejection_fraction, high_blood_pressure, platelets, serum_creatinine, serum_sodium, sex, smoking, time])
107
108         # Perform prediction
109         heart_disease_prediction = heart_disease_model.predict(input_data)
110
111         # Process the prediction result
112         if heart_disease_prediction[0] == 1:
113             heart_disease_diagnosis = 'The person has Heart Disease'
114         else:
115             heart_disease_diagnosis = 'The person does not have Heart Disease'
116
117     st.success(heart_disease_diagnosis)
118
```

Fig 5.15

5.5.4 BREAST CANCER:



```
52
53 elif selected == 'Breast Cancer Prediction':
54     st.title('Breast Cancer Prediction Using ML')
55
56     Age = st.text_input('Enter the Age')
57     BMI = st.text_input('BMI')
58     Glucose = st.text_input('Glucose')
59     Insulin = st.text_input('Insulin')
60     HOMA = st.text_input('Homa')
61     Leptin = st.text_input('Leptin')
62     Adiponectin = st.text_input('Adiponectin')
63     Resistin = st.text_input('Resistin')
64     MCP1 = st.text_input('MCP')
65
66     breastcancer_diagnosis = ''
67
68 if st.button('Breast Cancer Test Result'):
69     if '' in [Age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP1]:
70         st.error('Please enter valid numeric values for all input fields.')
71     else:
72         # Convert input data to numeric array
73         input_data = np.array([Age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP1], dtype=float)
74         breastcancer_prediction = breastcancer_model.predict(input_data)
75
76         if breastcancer_prediction[0] == 1:
77             breastcancer_diagnosis = 'The person has Breast Cancer'
78         else:
79             breastcancer_diagnosis = 'The person does not have Breast Cancer'
80
81     st.success(breastcancer_diagnosis)
82
```

Fig 5.16

5.5.5 WEBSITE SCREENSHOTS:

5.5.5.1 DIABETES:

×

Multiple disease Prediction System

Diabetes Prediction

Breast Cancer Prediction

Heart Disease Prediction

Deploy

⋮

Diabetes Prediction Using ML

Male=0 & Female=1

1

Enter the Age

50

Urea

4.7

Creatinine

46

Hemoglobin A1c

4.9

Cholesterol

4.2

Triglycerides

0.9

Fig 5.17

×

Multiple disease Prediction System

Diabetes Prediction

Breast Cancer Prediction

Heart Disease Prediction

Deploy

⋮

Cholesterol

4.2

Triglycerides

0.9

High-density lipoprotein

2.4

Low-density lipoprotein

1.4

Very low-density lipoprotein

0.5

Body mass index

24

Diabetes Test Result

The person is Not Diabetic

Fig 5.16

5.5.5.2 HEART DISEASE:

×

Deploy

Multiple disease Prediction System

Diabetes Prediction

Breast Cancer Prediction

Heart Disease Prediction

Heart Disease Prediction Using ML

Age of the Person

75

Value of Anemia

0

Creatinine Phosphokinase

5825

Diabetes

0

Ejection Fraction

20

High Blood Pressure

1

Platelets

265000

Fig 5.19

×

Deploy

Multiple disease Prediction System

Diabetes Prediction

Breast Cancer Prediction

Heart Disease Prediction

Platelets

265000

Serum Creatinine

1.9

Serum Sodium

130

Male = 0 & Female = 1

1

Smoking

0

Time

4

Heart Disease Test Result

The person has Heart Disease

Fig 5.20

5.5.5.3 BREAST CANCER:

×

Multiple disease Prediction System

Diabetes Prediction

Breast Cancer Prediction

Heart Disease Prediction

Breast Cancer Prediction Using ML

Enter the Age

48

BMI

23.5

Glucose

70

Insulin

2.707

Homa

0.467409

Leptin

8.8071

Deploy

Fig 5.21

×

Multiple disease Prediction System

Diabetes Prediction

Breast Cancer Prediction

Heart Disease Prediction

Insulin

2.707

Homa

0.467409

Leptin

8.8071

Adiponectin

9.7024

Resistin

7.99585

MCP

417.114

Breast Cancer Test Result

The person does not have Breast Cancer

Deploy

Fig 5.22

CHAPTER 6

SUMMARY AND CONCLUSION

6.1 SUMMARY:

The project focuses on disease prediction using machine learning algorithms such as logistic regression, support vector machine (SVM) and decision trees. Three diseases namely diabetes, heart disease and breast cancer were selected for prediction tasks.

Logistic regression and SVM were used to predict diabetes using input characteristics such as gender, age, urea, creatinine, HbA1c, cholesterol and triglycerides, HDL, LDL, VLDL and BMI. The logistic regression model had an accuracy score of 67%, while the SVM accuracy score was [SVM accuracy score not given]. These models provide binary predictions ("a person is diabetic" or "a person is not diabetic").

SVM and logistic regression algorithm were used to predict heart disease with characteristics such as age, anemia, creatinine phosphokinase, diabetes, ejection fraction, fraction, high blood pressure, platelets, serum creatinine, serum sodium, gender, smoking and time. The SVM accuracy score was [SVM accuracy score not given], indicating its effectiveness in handling complex decision constraints.

Decision trees with logistic regression and SVM algorithms with input parameters such as age, BMI and glucose were used to predict breast cancer, insulin, HOMA, leptin, adiponectin, resistin and MCP1. The decision trees achieved 75 percent accuracy, highlighting their interpretability and effectiveness in capturing complex relationships.

6.2 Conclusion:

The project successfully demonstrated the application of machine learning algorithms to disease prediction. Logistic regression, SVM and decision trees have been used to predict diabetes, heart disease and breast cancer.

Logistic regression has proven its power to provide interpretable coefficients and simple binary predictions. SVM, with the ability to handle complex decision constraints, has proven effective in heart disease and breast cancer prediction tasks. Decision trees achieved slightly lower accuracy, but produced intuitive and easy-to-interpret results, so they are suitable for explaining predictive logic to non-technical users.

Overall, the project emphasizes the importance of machine learning in healthcare, especially disease. prognosis and risk assessment. These predictive models can assist health professionals and individuals in proactive health management, early detection and informed decision-making, ultimately improving health outcomes and patient well-being.

REFERENCES

1. Gouda, W., Almurafteh, M., Humayun, M., & Jhanjhi, N. Z. (2022). Detection of COVID-19 Based on Chest X-rays Using Deep Learning. *Healthcare (Basel, Switzerland)*, 10(2), 343. <https://doi.org/10.3390/healthcare10020343>
2. Ahsan, M. M., Luna, S. A., & Siddique, Z. (2022). Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. *Healthcare (Basel, Switzerland)*, 10(3), 541. <https://doi.org/10.3390/healthcare10030541>
3. Abnoosian, K., Farnoosh, R., & Behzadi, M. H. (2023). Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. *BMC bioinformatics*, 24(1), 337. <https://doi.org/10.1186/s12859-023-05465-z>
4. Darvishi, S., Hamidi, O., & Poorolajal, J. (2021). Prediction of Multiple sclerosis disease using machine learning classifiers: a comparative study. *Journal of preventive medicine and hygiene*, 62(1), E192–E199. <https://doi.org/10.15167/2421-4248/jpmh2021.62.1.1651>
5. You, J., Guo, Y., Kang, J. J., Wang, H. F., Yang, M., Feng, J. F., Yu, J. T., & Cheng, W. (2023). Development of machine learning-based models to predict 10-year risk of cardiovascular disease: a prospective cohort study. *Stroke and vascular neurology*, 8(6), 475–485. <https://doi.org/10.1136/svn-2023-002332>
6. Montolío, A., Martín-Gallego, A., Cegoñino, J., Orduna, E., Vilades, E., Garcia-Martin, E., & Palomar, A. P. D. (2021). Machine learning in diagnosis and disability prediction of multiple sclerosis using optical coherence tomography. *Computers in biology and medicine*, 133, 104416. <https://doi.org/10.1016/j.compbimed.2021.104416>
7. Aslam, N., Khan, I. U., Bashamakh, A., Alghool, F. A., Aboulnour, M., Alsuwayan, N. M., Alturaif, R. K., Brahimi, S., Aljameel, S. S., & Al Ghamdi, K. (2022). Multiple Sclerosis Diagnosis Using Machine Learning and Deep Learning: Challenges and Opportunities. *Sensors (Basel, Switzerland)*, 22(20), 7856. <https://doi.org/10.3390/s22207856>
8. Alqahtani, A., Alsubai, S., Sha, M., Vilcekova, L., & Javed, T. (2022). Cardiovascular Disease Detection using Ensemble Learning. *Computational intelligence and neuroscience*, 2022, 5267498. <https://doi.org/10.1155/2022/5267498>
9. Chen, Z., Yang, M., Wen, Y., Jiang, S., Liu, W., & Huang, H. (2022). Prediction of atherosclerosis using machine learning based on operations research. *Mathematical biosciences and engineering : MBE*, 19(5), 4892–4910. <https://doi.org/10.3934/mbe.2022229>
10. Gulamali, F. F., Sawant, A. S., & Nadkarni, G. N. (2022). Machine learning for risk stratification in kidney disease. *Current opinion in nephrology and hypertension*, 31(6), 548–552. <https://doi.org/10.1097/MNH.0000000000000832>

11. Bani Hani, S. H., & Ahmad, M. M. (2023). Machine-learning Algorithms for Ischemic Heart Disease Prediction: A Systematic Review. *Current cardiology reviews*, 19(1), e090622205797. <https://doi.org/10.2174/1573403X18666220609123053>
12. Abdul Rahman, H., Ottom, M. A., & Dinov, I. D. (2023). Machine learning-based colorectal cancer prediction using global dietary data. *BMC cancer*, 23(1), 144. <https://doi.org/10.1186/s12885-023-10587-x>
13. Moazami, F., Lefevre-Utile, A., Papaloukas, C., & Soumelis, V. (2021). Machine Learning Approaches in Study of Multiple Sclerosis Disease Through Magnetic Resonance Images. *Frontiers in immunology*, 12, 700582. <https://doi.org/10.3389/fimmu.2021.700582>
14. Pinto, M. F., Oliveira, H., Batista, S., Cruz, L., Pinto, M., Correia, I., Martins, P., & Teixeira, C. (2020). Prediction of disease progression and outcomes in multiple sclerosis with machine learning. *Scientific reports*, 10(1), 21038. <https://doi.org/10.1038/s41598-020-78212-6>
15. Ouwerkerk, J., Feleus, S., van der Zwaan, K. F., Li, Y., Roos, M., van Roon-Mom, W. M. C., de Bot, S. T., Wolstencroft, K. J., & Mina, E. (2023). Machine learning in Huntington's disease: exploring the Enroll-HD dataset for prognosis and driving capability prediction. *Orphanet journal of rare diseases*, 18(1), 218. <https://doi.org/10.1186/s13023-023-02785-4>
16. Sharifi-Kia, A., Nahvijou, A., & Sheikhtaheri, A. (2023). Machine learning-based mortality prediction models for smoker COVID-19 patients. *BMC medical informatics and decision making*, 23(1), 129. <https://doi.org/10.1186/s12911-023-02237-w>
17. Peng, Y., Zheng, Y., Tan, Z., Liu, J., Xiang, Y., Liu, H., Dai, L., Xie, Y., Wang, J., Zeng, C., & Li, Y. (2021). Prediction of unenhanced lesion evolution in multiple sclerosis using radiomics-based models: a machine learning approach. *Multiple sclerosis and related disorders*, 53, 102989. <https://doi.org/10.1016/j.msard.2021.102989>
18. Meiseles, A., Paley, D., Ziv, M., Hadid, Y., Rokach, L., & Tadmor, T. (2022). Explainable machine learning for chronic lymphocytic leukemia treatment prediction using only inexpensive tests. *Computers in biology and medicine*, 145, 105490. <https://doi.org/10.1016/j.compbiomed.2022.105490>
19. Vázquez-Marrufo, M., Sarrias-Arrabal, E., García-Torres, M., Martín-Clemente, R., & Izquierdo, G. (2023). A systematic review of the application of machine-learning algorithms in multiple sclerosis. *Neurologia*, 38(8), 577–590. <https://doi.org/10.1016/j.nrleng.2020.10.013>
20. Hahn R. T. (2021). Topologic Data Analysis and Machine Learning: Defining the Aortic Valvular-Ventricular Disease Terrain. *JACC. Cardiovascular imaging*, 14(9), 1721–1723. <https://doi.org/10.1016/j.jcmg.2021.04.005>

21. Dai, Q., Wang, Z., Liu, Z., Duan, X., Song, J., & Guo, M. (2022). Predicting miRNA-disease associations using an ensemble learning framework with resampling method. *Briefings in bioinformatics*, 23(1), bbab543. <https://doi.org/10.1093/bib/bbab543>
22. Nakamura, K., Uchino, E., Sato, N., Araki, A., Terayama, K., Kojima, R., Murashita, K., Itoh, K., Mikami, T., Tamada, Y., & Okuno, Y. (2023). Individual health-disease phase diagrams for disease prevention based on machine learning. *Journal of biomedical informatics*, 144, 104448. <https://doi.org/10.1016/j.jbi.2023.104448>
23. Kang, M., Kim, S., Lee, D. B., Hong, C., & Hwang, K. B. (2023). Gene-specific machine learning for pathogenicity prediction of rare BRCA1 and BRCA2 missense variants. *Scientific reports*, 13(1), 10478. <https://doi.org/10.1038/s41598-023-37698-6>
24. Wayment-Steele, H. K., Ojoawo, A., Otten, R., Apitz, J. M., Pitsawong, W., Hömberger, M., Ovchinnikov, S., Colwell, L., & Kern, D. (2024). Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature*, 625(7996), 832–839. <https://doi.org/10.1038/s41586-023-06832-9>
25. Inoue, H., Oya, M., Aizawa, M., Wagatsuma, K., Kamimae, M., Kashiwagi, Y., Ishii, M., Wakabayashi, H., Fujii, T., Suzuki, S., Hattori, N., Tatsumoto, N., Kawakami, E., & Asanuma, K. (2023). Predicting dry weight change in Hemodialysis patients using machine learning. *BMC nephrology*, 24(1), 196. <https://doi.org/10.1186/s12882-023-03248-5>
26. Wu, K., Wu, E., DAndrea, M., Chitale, N., Lim, M., Dabrowski, M., Kantor, K., Rangi, H., Liu, R., Garmhausen, M., Pal, N., Harbron, C., Rizzo, S., Copping, R., & Zou, J. (2022). Machine Learning Prediction of Clinical Trial Operational Efficiency. *The AAPS journal*, 24(3), 57. <https://doi.org/10.1208/s12248-022-00703-3>