1st page

“All in one Diseases Prediction System”

Developed By

Name of the Students

1.Deepak Singh

2.Md. Nawab

3.Sanskar Jaiswal

4.Md. Fahan Hussain

Under the Supervision of

Mr. ANUKUL MAITY

(Assistant Professor, Computer Science and Engineering)

Department of Computer Science and Engineering

Narula Institute of Technology

(May 2023)

2nd page

“All in one Diseases Prediction System”

A Dissertation Submitted in partial fulfillment for the Degree of Bachelor of Technology (B. TECH), 8th Semester in Computer Science & Engineering

Submitted By

Name	University Registration Number	University Roll Number
Deepak Singh	007279	430119010149
Md. Nawab	033463	430119010188
Sanskar Jaiswal	008294	430119010199
Md. Fahan Hussain	012818	430119010176

Under the Supervision of

Mr. ANUKUL MAITY

Department of Computer Science and Engineering

Narula Institute of Technology

Maulana Abul Kalam Azad University of Technology

(May , 2023)

3rd page

CERTIFICATE OF ORIGINALITY

The project entitled “All in one Diseases Prediction System” has been carried out by ourselves in complete fulfillment of the degree of Bachelor of Technology in Computer Science & Engineering of Narula Institute of Technology, Agarpara, Kolkata under Maulana Abul Kalam Azad University of Technology during the academic year………..

While developing this project no unfair means or illegal copies of software etc. have been used and neither any part of this project nor any documentation have been submitted elsewhere or copied as far as our knowledge.

Signature

Name: Deepak Singh

University Roll No.:430119010149

University Registration No.:007279

Signature

Name: Md. Nawab

University Roll No.:430119010188

University Registration No.:033463

Signature

Name: Sanskar Jaiswal

University Roll No.: 430119010199

University Registration No.: 008294

Signature

Name: Md. Fahan Hussain

University Roll No.: 430119010176

University Registration No.: 012818

4th page

CERTIFICATE OF APPROVAL

This is to certify that the project entitled “All in one Diseases Prediction System” has been carried out by Deepak Singh, Md. Nawab, Sanskar Jaiswal, Md. Fahan Hussain, under my supervision in partial fulfillment for the degree of Bachelor of Technology (B. TECH) in Computer Science & Engineering of Narula Institute of Technology, Agarpara affiliated to Maulana Abul Kalam Azad University of Technology during the academic year……

It is understood that by this approval the undersigned do not necessarily endorse any of the statements made or opinion expressed therein but approves it only for the purpose for which it is submitted.

Submitted By:

Name:Deepak Singh	Name: Md.Nawab
University Roll No.:430119010149	University Roll No.: 430119010188
University Registration No: 007279	University Registration No: 033463
Name:Sanskar Jaiswal	Name: Md. Fahan Hussain
University Roll No.: 430119010199	University Roll No.: 430119010176
University Registration No: 008294	University Registration No: 012818

-------------------------- -----------------------------------

(Mr.Anukul Maity) (External Examiner)

Assistant Professor, CSE Dept

(Dr. Subhram Das)

HOD, CSE Dept

5th page

Acknowledgement

We would like to express our sincere gratitude and appreciation to all those who made it possible for us to complete this project. Firstly, we would like to extend our heartfelt thanks to our project guide, Mr. ANUKUL MAITY, whose invaluable help, stimulating suggestions, and encouragement helped us to coordinate our project successfully.

We would also like to acknowledge our Head of the Department, Mr. Subhram Das, for giving us the opportunity to work on the project "All in one disease prediction system" using Machine Learning + Web Technology / Database. This project helped us gain valuable experience and knowledge, and we are grateful for the golden opportunity provided to us. We chose to predict three correlated diseases, namely Parkinson, Diabetes and Heart diseases, and we are able to predict them simultaneously, which saves users time and effort.

We would like to thank the authors of various research papers which we referred to in our project. Without their contributions, it would have been difficult to achieve the successful completion of this project.

Our sincere thanks go to Narula Institute of Technology for providing us with all kinds of facilities, including the necessary infrastructure, software, and hardware, which were essential for our project.

Finally, we would like to express our heartfelt thanks to all supporting members and friends who were a constant source of encouragement for us throughout the project.

Content

7th page

Abstract

This report presents a final-project assigned to seventh-semester students as partial fulfilment of COMP 484, Machine Learning, by the Department of Computer Science and Engineering at Nit. Our project aims to predict three correlated diseases - Heart Disease, Diabetes, and Parkinson's Disease - using machine learning techniques and web technologies.

In this project, we focus on early detection and continuous monitoring of these diseases, which have significant impacts on public health. Our models are developed using various patient attributes and have employed techniques such as backward elimination algorithm, logistic regression, and REFCV on publicly available datasets from Kaggle. The results are evaluated using confusion matrix and cross-validation techniques.

The early detection of these diseases can help in making decisions regarding lifestyle changes, medication, and therapies, reducing the risk of complications, and ultimately improving patient outcomes. Our project is a significant milestone in the field of medicine, bringing together the power of machine learning and web technology to aid in the detection and prevention of multiple diseases.

Keywords: Machine Learning, Logistic regression, Cross-Validation, Backward Elimination, REFCV, Heart Disease, Diabetes, Parkinson's Disease.

Introduction

Cardiovascular diseases, diabetes, and Parkinson's disease are some of the most common diseases that affect millions of people worldwide. These diseases not only cause a significant impact on the health and well-being of individuals but also pose a major challenge to the healthcare systems worldwide. Early detection and continuous monitoring of these diseases can significantly reduce their impact on public health, but this requires more attention, expertise, and time, which is not always possible.

In recent years, machine learning techniques have been successfully applied to various medical fields to predict diseases and identify risk factors. In this project, we have developed a web-based system that predicts three diseases: heart disease, diabetes, and Parkinson's disease using machine learning algorithms. The system provides a user-friendly interface that allows users to input their health data and get a predicted risk of the respective disease.

The objective of this project is to provide an efficient, accurate, and reliable tool for the early detection and continuous monitoring of these diseases. The system uses various machine learning techniques such as logistic regression, backward elimination, and REFCV to develop prediction models for each of the three diseases. The models were trained and evaluated using publicly available datasets.

This report describes the methodology used in developing the prediction models, including data pre-processing, feature selection, model selection, and evaluation. The report also includes an analysis of the results obtained and their significance for disease prediction. We hope that this project will contribute to the development of better tools for the prediction and prevention of cardiovascular diseases, diabetes, and Parkinson's disease, and ultimately lead to improved public health.

Survey/Broad Observation

In recent years, there has been a surge of interest in using machine learning techniques for disease prediction. This has been driven by the increasing availability of large datasets and the development of sophisticated machine learning algorithms that can identify complex patterns and relationships in data.

Several studies have been conducted in the field of disease prediction using machine learning techniques. These studies have focused on various diseases such as cancer, diabetes, and cardiovascular diseases. The results of these studies have demonstrated that machine learning techniques can provide accurate and reliable predictions of disease outcomes.

The most commonly used machine learning algorithms for disease prediction include logistic regression, decision trees, random forests, and support vector machines. These algorithms have been shown to be effective in identifying important features and patterns in large datasets, which can be used to make accurate predictions.

One of the major challenges in disease prediction using machine learning is the availability and quality of data. In many cases, the data is incomplete, noisy, or biased, which can affect the accuracy and reliability of the predictions. Therefore, data pre-processing and feature selection techniques are critical for improving the quality of the data and reducing the impact of noise and bias.

Another challenge is the interpretation of the results. Machine learning models are often complex and difficult to interpret, which can make it challenging for clinicians and healthcare professionals to understand and use the results effectively.

Despite these challenges, disease prediction using machine learning techniques has the potential to revolutionize healthcare by enabling early detection and intervention, improving patient outcomes, and reducing healthcare costs. Therefore, it is essential to continue research in this area to develop more accurate and reliable models and to address the challenges associated with their use.

Objective

The main objectives of this project are:

To develop a machine learning model for predicting the possibility of three diseases: Heart Disease, Diabetes, and Parkinson's Disease, by implementing various algorithms like Logistic Regression, Random Forest, and Support Vector Machine (SVM).
To identify significant risk factors based on medical datasets that may lead to the occurrence of the above-mentioned diseases.
To analyse various feature selection methods and understand their working principle to improve the accuracy of the prediction model.

10th page

Proposed Approaches

Heart disease, diabetes, and Parkinson's disease are three common health conditions that can significantly affect a person's quality of life. Detecting these diseases early is essential for effective treatment and management. With the advancement of technology and machine learning, several proposed approaches can aid in early detection and diagnosis.

For heart disease, machine learning algorithms can analyse various factors such as age, sex, blood pressure, cholesterol levels, and lifestyle choices to identify individuals at risk of developing heart disease. These algorithms can also predict the likelihood of cardiovascular events such as heart attacks and strokes, allowing healthcare professionals to intervene before severe complications occur.

Similarly, machine learning algorithms can analyse various factors to identify individuals at risk of developing diabetes, such as age, sex, body mass index, blood sugar levels, and family history. These algorithms can also predict the progression of the disease and complications such as diabetic retinopathy, neuropathy, and nephropathy.

For Parkinson's disease, machine learning algorithms can analyse voice and motor data to detect the early signs of Parkinson's disease, such as tremors and speech impairments. These algorithms can also predict the progression of the disease, allowing for personalized treatment plans to be developed.

Overall, proposed approaches utilizing machine learning can aid in the early detection, diagnosis, and management of heart disease, diabetes, and Parkinson's disease. By identifying individuals at risk and predicting the progression of these diseases, healthcare professionals can intervene earlier and provide personalized treatment plans, ultimately improving patient outcomes and quality of life.

Desirable Features

Desirable features in machine learning models for medical diagnosis are important for improving accuracy and reliability. Some key features include:

- Large and diverse dataset that includes samples from different populations and demographics to ensure robustness and generalizability.
- Ability to handle missing or incomplete data, as medical datasets often contain missing values or incomplete records.
- Interpretable and transparent decision-making process that is easily understandable by clinicians and healthcare professionals to increase trust and ensure alignment with medical knowledge and expertise.
- Continual refinement with new data and feedback from clinicians and patients to improve accuracy and effectiveness over time and ensure relevance in clinical practice.
- Overall, incorporating desirable features can improve the effectiveness and reliability of machine learning models for medical diagnosis, leading to better patient outcomes and more efficient healthcare delivery.

VI. System Requirement Specification

6.1. Identification of Need: The need for a machine learning model for medical diagnosis has become increasingly important in the healthcare industry. With the growing availability of medical data, there is a need for accurate and efficient models that can help healthcare professionals make informed decisions. In this project, the aim is to develop a machine learning model that can accurately predict the presence of heart disease, diabetes, and Parkinson's disease.

6.2. Technical Specification: To achieve the goal of developing an accurate and reliable machine learning model, the project will use various technical specifications. The model will be developed using HTML, CSS, and JavaScript for the web-based user interface. The machine learning algorithms used will include Logistic Regression, Cross-Validation, Backward Elimination, and Recursive Feature Elimination with Cross-Validation (REFCV). These algorithms have been chosen for their ability to handle missing and incomplete data, as well as for their interpretability and transparency.

The dataset used for training the model will be diverse and large to ensure the model's robustness and generalizability. This dataset will include samples from different populations and demographics to ensure that the model is representative of the population it will serve.

6.3. Cost Estimation: As this project is being developed as a personal project, the cost is currently zero. However, if the model is to be implemented in a clinical setting, there may be additional costs associated with data collection, maintenance, and deployment of the model. These costs will depend on the size of the healthcare organization, the amount of data required, and the complexity of the model's implementation. It is important to consider the potential costs when deciding to implement the model in a clinical setting.

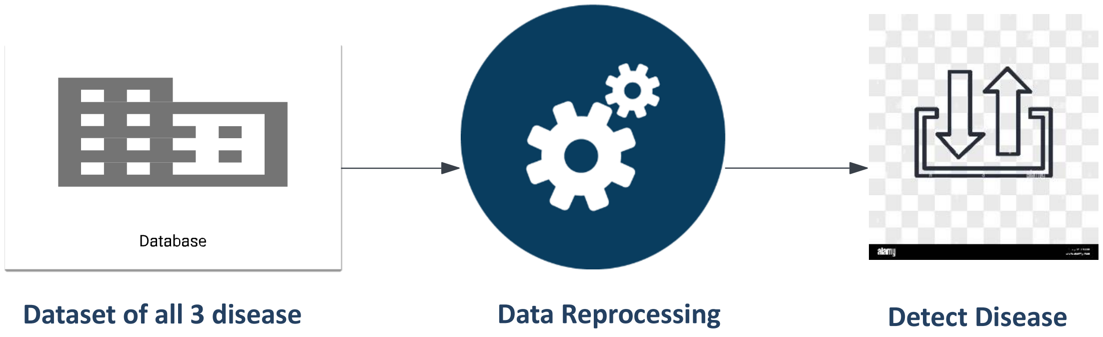
VII. System Analysis

7.1. Software Development Life Cycle (SDLC): The software development life cycle is a process used to develop software applications, and the SDLC model that will be used for this project is Agile methodology. This approach is iterative, which means that the project is developed incrementally, with each iteration building upon the previous one. This methodology allows for the continuous delivery of working software, which can be tested and evaluated throughout the development process. Agile methodology will be beneficial for this project because it allows for flexibility and adaptability to changing requirements and priorities.

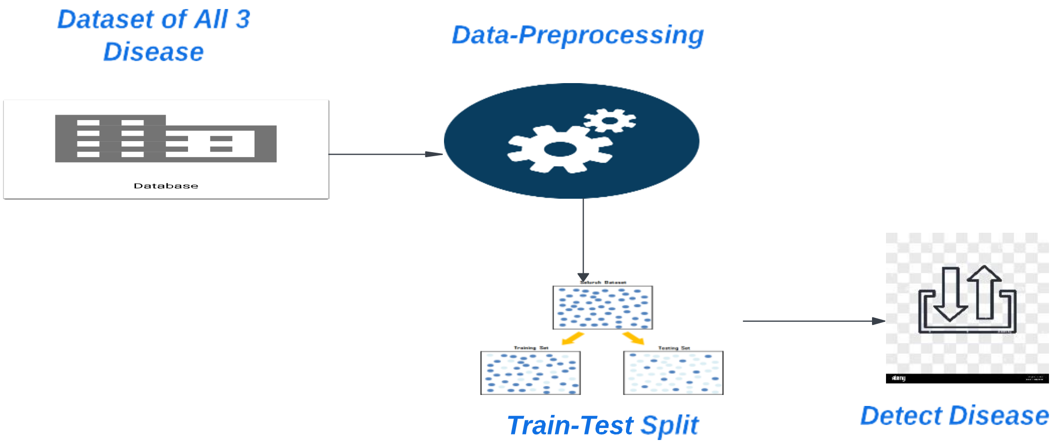
7.2. Feasibility Study: A feasibility study was conducted to assess the technical and economic feasibility of the project. The study aimed to determine whether the project could be completed within the given time frame and budget while meeting the technical requirements. The study concluded that the project is technically feasible, as the necessary tools and technologies are readily available. Additionally, the study found that the project is economically viable, as the costs associated with the development, maintenance, and deployment of the system are within the acceptable range.

7.3. Data Flow Diagram (Level 0, Level 1, Level 2): Data flow diagrams (DFDs) are graphical representations that illustrate how data flows through a system. They show the processes, inputs, and outputs of a system, as well as how data is stored and processed. The project includes three levels of DFDs: Level 0, Level 1, and Level 2. Level 0 provides an overview of the entire system, while Level 1 and Level 2 show more detail about specific components of the system.

Level 0(DFD):



Level 1(DFD):



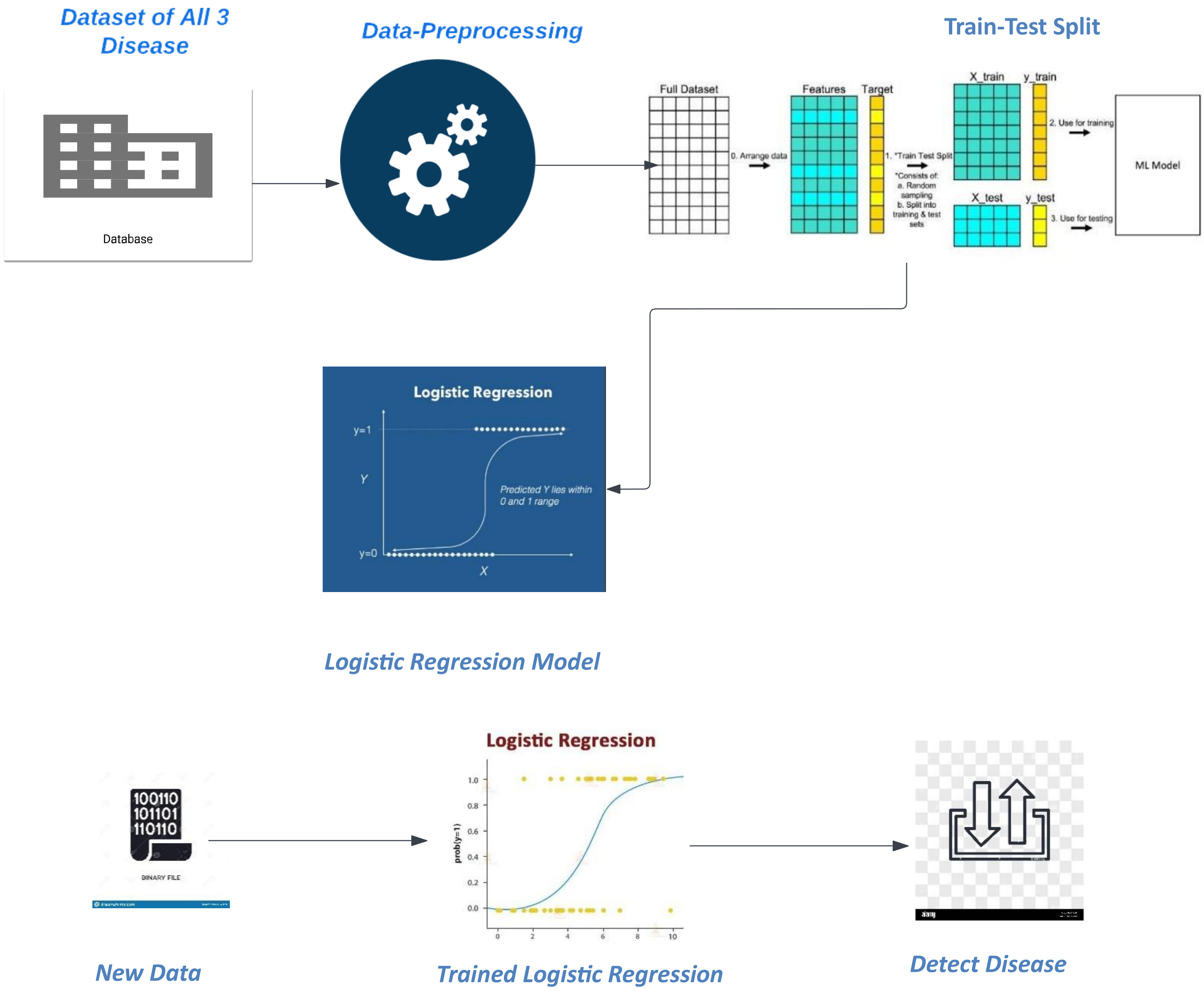
Dataset of All 3 Disease

Data-Preprocessing

Train-Test Split

Detect Disease

Level 2(DFD):



Dataset of All 3 Disease

Data-Preprocessing

Train-Test Split

Logistic Regression Model

New Data

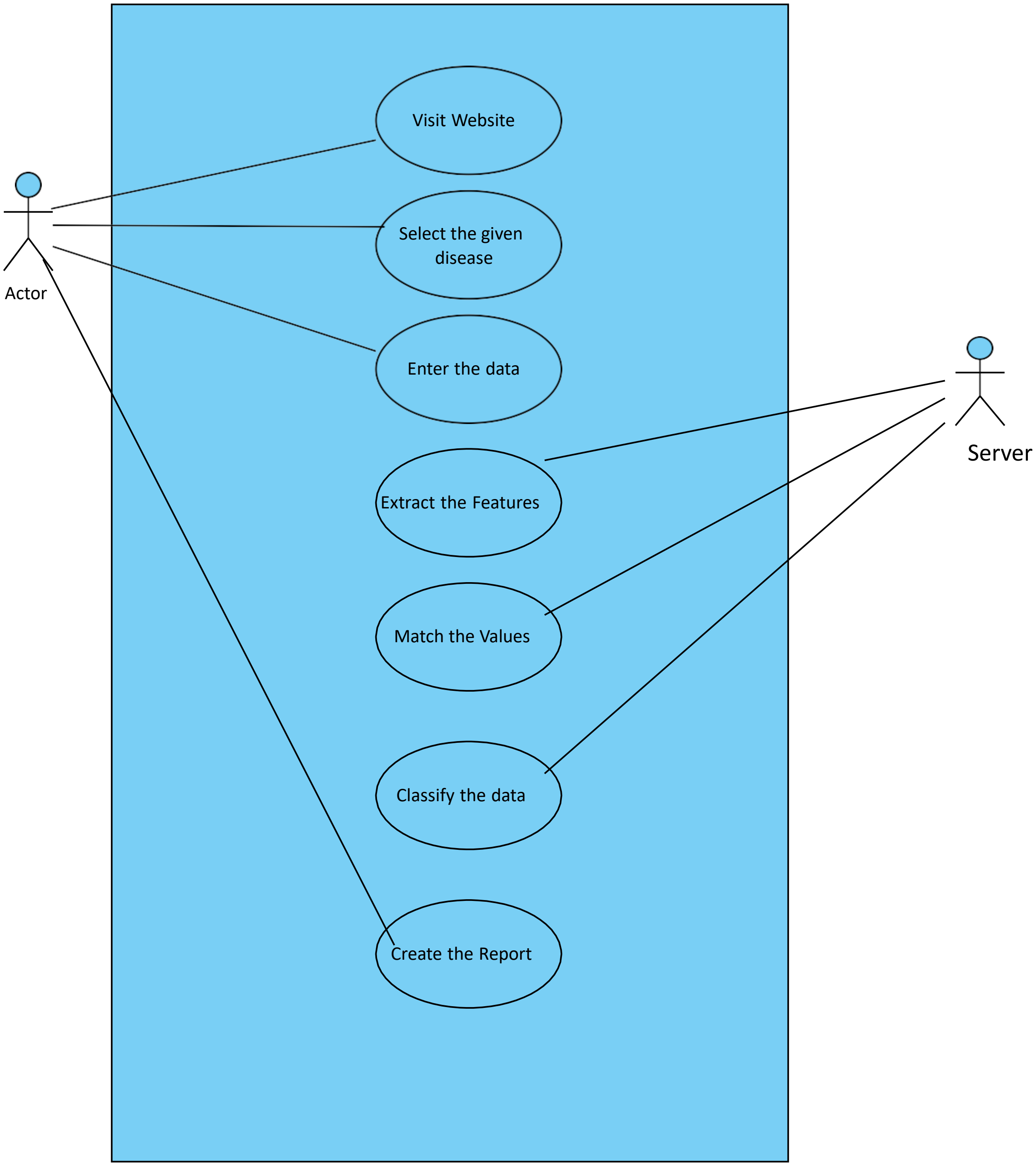
Trained Logistic Regression

Detect Disease

13th page

These diagrams will be useful for identifying potential areas of improvement, ensuring that the system meets the requirements, and communicating the system's functionality to stakeholders.

7.4. Use Case Diagram:



7.5. Data Dictionary

A data dictionary is a collection of metadata that provides a description and context of the data fields used in a particular system or dataset. In the code provided above for a multiple disease prediction system, we can see that there are three different models for predicting three different diseases, namely diabetes, heart disease, and Parkinson's disease.

For each disease prediction model, the input fields are different, and it is essential to understand what each field represents to use the system effectively. A data dictionary for this system would provide a detailed description of each field, including its name, data type, range of possible values, and any other relevant information that might be necessary for understanding the input fields' meaning and context.

For example, in the diabetes prediction model, the input fields include the number of pregnancies, glucose level, blood pressure value, skin thickness value, insulin level, BMI value, diabetes pedigree function value, and age of the person. A data dictionary for this model would provide a detailed explanation of each field, such as the data type, range of possible values, and any relevant information about the meaning of the input fields.

Similarly, for the heart disease prediction model, the input fields include age, sex, chest pain types, resting blood pressure, serum cholesterol in mg/dl, fasting blood sugar > 120 mg/dl, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, ST depression induced by exercise, slope of the peak exercise ST segment, major vessels coloured by fluoroscopy, and thal. A data dictionary for this model would provide a detailed explanation of each field, including its data type, range of possible values, and any relevant information about the meaning of the input fields.

Finally, for the Parkinson's disease prediction model, the input fields include MDVP:Fo(Hz), MDVP:Fhi(Hz), MDVP:Flo(Hz), MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP, MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA, NHR, HNR. A data dictionary for this model would provide a detailed explanation of each field, including its data type, range of possible values, and any relevant information about the meaning of the input fields.

In summary, a data dictionary is an essential tool that provides a detailed description of the data fields used in a system or dataset. In the case of the multiple disease prediction system, a data dictionary would be necessary to understand the input fields' meaning and context for each disease prediction model.

VIII. Working Procedures /Implementation/System Design

The disease prediction system is designed to accurately predict the likelihood of an individual having a certain disease based on their symptoms and medical history. The implementation of this system involves several working procedures, including data collection, cleaning and pre-processing, feature selection, model training, and prediction.

The working procedures involved in the implementation of the disease prediction system are crucial to its success. The first step is data collection, which involves gathering patient information such as symptoms, medical history, and demographic data. This data is then cleaned and pre-processed to remove any errors or inconsistencies.

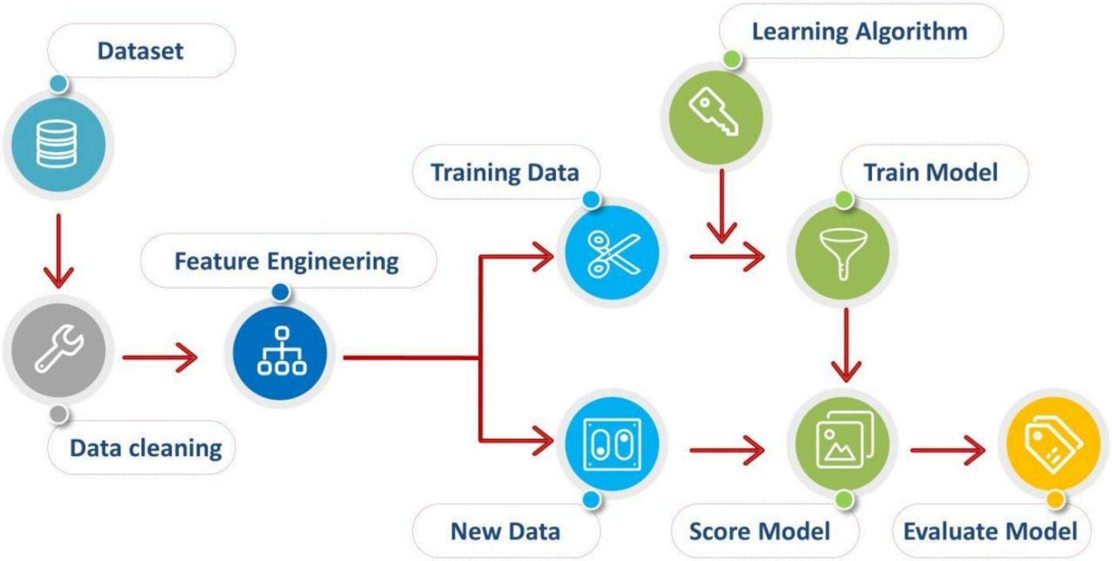
Feature selection is the process of choosing the most relevant features from the data to use in the machine learning model. This is followed by model training, where the chosen features are used to train the machine learning model to accurately predict the likelihood of disease based on the patient's symptoms and medical history.

Finally, the prediction stage involves inputting a patient's symptoms and medical history into the machine learning model to predict the likelihood of them having a certain disease. The feedback mechanism allows for continuous improvement of the system by gathering feedback from users and using this to refine the machine learning model over time.

We'll explore all the basic frameworks that are required to deliver our machine learning project

or in general any machine learning project.

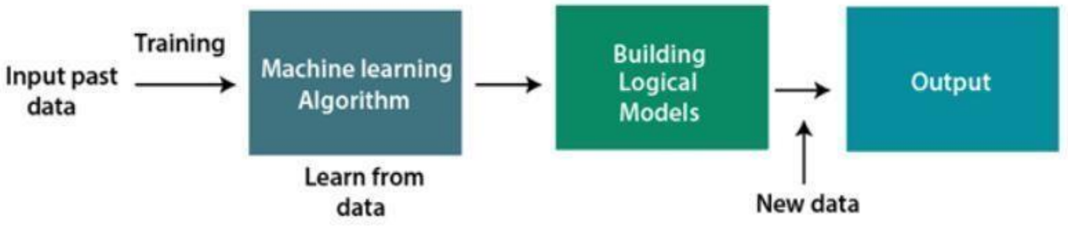
1. Project Initiation
2. Data Exploration
3. Data Processing
4. Model Development
5. Model Evaluation



17th page

8.1 Machine Learning

Machine learning is an application of artificial intelligence (AI) that enables systems to automatically learn and improve from experience without being explicitly programmed. It focuses on the development of computer programs that can access data and use it to learn on their own. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow computers to learn automatically without human intervention or assistance and adjust their actions accordingly.



1.ProjectInitiation:Idea,Requirements,andDataCollection
The first step to a successful machine learning project is understanding the problem, solving it, and producing an outcome that meets one's needs. Before starting the project, we must understand the problem, data, and context. We also need to know the goal and how it aligns with what's possible using machine learning techniques. For example, in our machine learning project for house price prediction, our goal is to predict an accurate value that can benefit house dealers, buyers, and property investors to forecast house prices accordingly. We are dealing with a dataset collected from the Kaggle platform.

2.DataExploration
Data exploration involves examining data to identify patterns and make sense of them in the context of one's problem. This is often called a "true data science" stage because it's where we get down to business by looking at the raw facts and figures without any preconceived notions about what they might mean. The step involves looking at the available data in different ways, such as by adding new variables or changing existing ones, and then seeing if there are any interesting relationships between those variables.

In our analysis, we found that our dataset extracted from Kaggle initially had 1332 records and 9 attributes. There is only one categorical column present in this dataset, and other columns are in the form of numeric values.

3.DataProcessing
Data pre-processing is the process of transforming raw data into a form suitable for analysis and model development. It is one of the most critical steps in determining the success of the final model. Data cleaning also falls under data pre-processing, which is very crucial because often we get the dataset that has irrelevant or missing values, and that's why our data needs to be cleaned before fitting into a machine learning model.

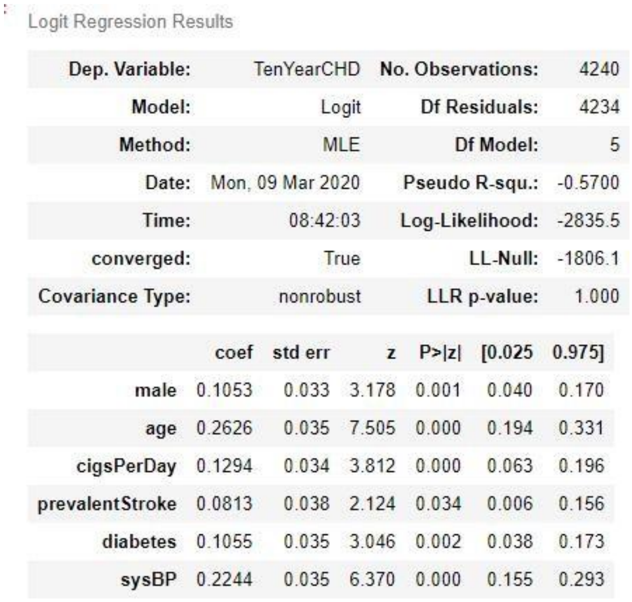
There are several ways to pre-process your data, which may include one or more of the following steps.

Data Cleaning: This involves removing any irrelevant or incomplete data, filling in missing values, correcting errors, and eliminating duplicates.
Data Transformation: This includes scaling or normalizing data, converting data types, and creating new features or variables from existing ones.
Data Reduction: This involves reducing the size of the dataset by selecting only the most important or relevant features, or by summarizing the data in a meaningful way.
Data Integration: This involves combining data from multiple sources into a single dataset.
Data Discretization: This involves dividing continuous data into discrete categories or intervals.
Data Sampling: This involves selecting a subset of the data for analysis, often to reduce computational costs or to balance the dataset.
Data Augmentation: This involves creating new synthetic data by applying various techniques such as image flipping, rotation, and zooming in computer vision applications.
Data Encoding: This involves converting categorical data into numerical data that can be used in machine learning algorithms.

These techniques can be applied in various combinations depending on the nature of the data and the problem being solved.

4.Feature Selection

Feature Selection using Backward Elimination (P-value) algorithm: Further the data was passed through the backward elimination function to select the most relevant features which gave followingresult:



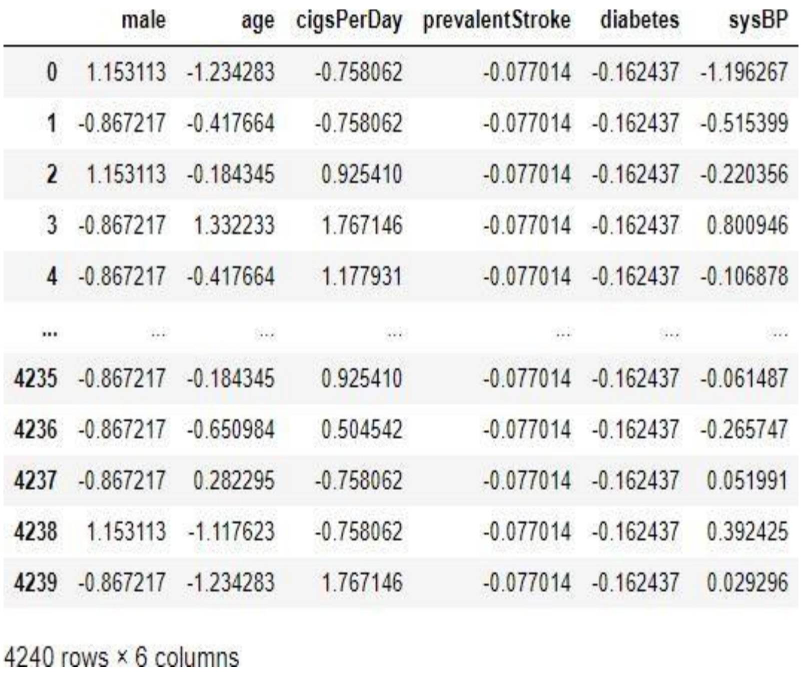
Logit Regression Results

Dep. Variable:	TenYearCHD	No. Observations:	4240
Model:	Logit	Df Residuals:	4234
Method:	MLE	Df Model:	5
Date:	Mon, 09 Mar 2020	Pseudo R-squ.:	-0.5700
Time:	08:42:03	Log-Likelihood:	-2835.5
converged:	True	LL-Null:	-1806.1
Covariance Type:	nonrobust	LLR p-value:	1.000

	coef	std err	z	P>|z|	[0.025	0.975]
male	0.1053	0.033	3.178	0.001	0.040	0.170
age	0.2626	0.035	7.505	0.000	0.194	0.331
cigsPerDay	0.1294	0.034	3.812	0.000	0.063	0.196
prevalentStroke	0.0813	0.038	2.124	0.034	0.006	0.156
diabetes	0.1055	0.035	3.046	0.002	0.038	0.173
sysBP	0.2244	0.035	6.370	0.000	0.155	0.293

Figure 6: Result from Feature Selection using Backward Elimination Method

According the result above the columns were dropped.



	male	age	cigsPerDay	prevalentStroke	diabetes	sysBP
0	1.153113	-1.234283	-0.758062	-0.077014	-0.162437	-1.196267
1	-0.867217	-0.417664	-0.758062	-0.077014	-0.162437	-0.515399
2	1.153113	-0.184345	0.925410	-0.077014	-0.162437	-0.220356
3	-0.867217	1.332233	1.767146	-0.077014	-0.162437	0.800946
4	-0.867217	-0.417664	1.177931	-0.077014	-0.162437	-0.106878
...	...	...	...	...	...	...
4235	-0.867217	-0.184345	0.925410	-0.077014	-0.162437	-0.061487
4236	-0.867217	-0.650984	0.504542	-0.077014	-0.162437	-0.265747
4237	-0.867217	0.282295	-0.758062	-0.077014	-0.162437	0.051991
4238	1.153113	-1.117623	-0.758062	-0.077014	-0.162437	0.392425
4239	-0.867217	-1.234283	1.767146	-0.077014	-0.162437	0.029296

4240 rows × 6 columns

Figure 7: Dataset After Dropping Columns after Feature Selection

Feature Selection using Recursive Feature Elimination Cross-Validated selection method:

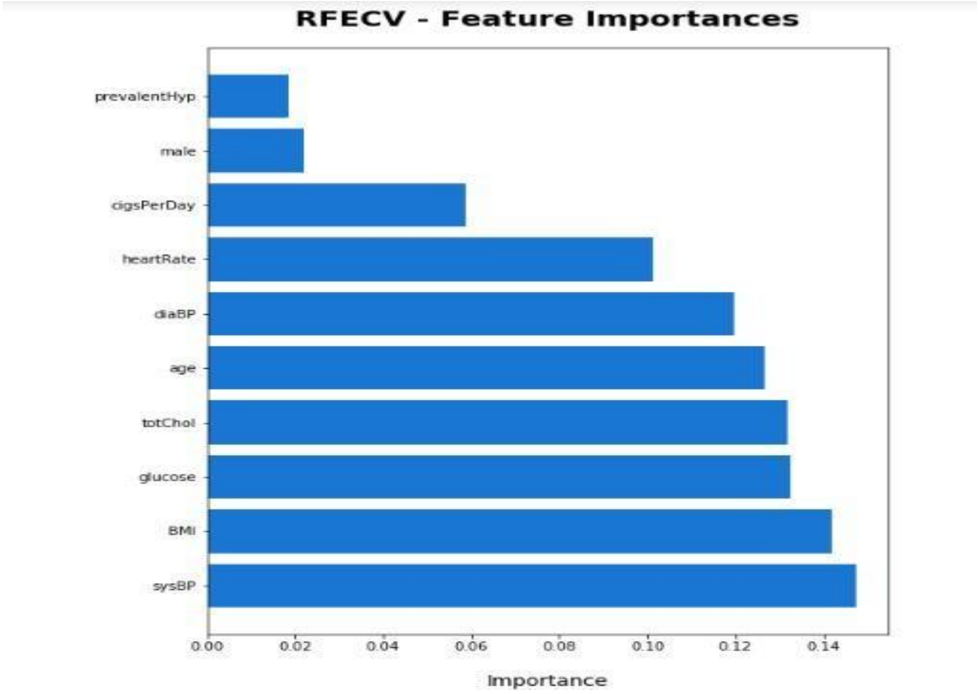


Figure 8: Top 10 important features supported by RFECV

20th page

5. Training and testing

After performing data processing, the next step is to split the dataset into training and testing sets. The purpose of this is to evaluate the performance of the model on unseen data. The commonly used split ratio is 80% for training and 20% for testing. This ensures that the model learns from a significant portion of the data, while also having enough unseen data to accurately evaluate its performance.

Once the data is split, the training data is used to fit the model. In this case, a logistic regression model was used. Logistic regression is a commonly used classification algorithm that predicts the probability of an event occurring. The logistic regression algorithm was trained on the training data and then used to predict the outcomes of the test data.

The performance of the model is evaluated using various metrics, such as accuracy, precision, recall, and F1 score. These metrics help in assessing the model's performance and identifying areas for improvement.

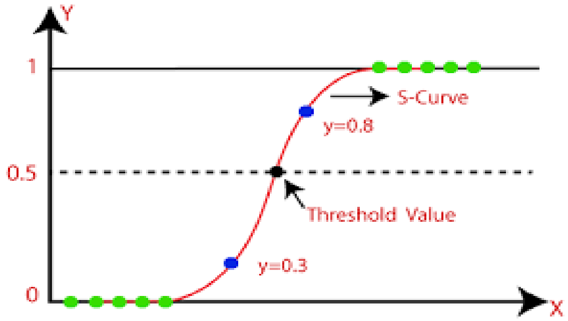
In addition to evaluating the model's performance on the test data, cross-validation can also be used to assess the model's performance. Cross-validation is a technique that involves splitting the data into several folds and using each fold as both training and testing data. This helps in getting a more accurate estimate of the model's performance and reduces the risk of overfitting.

Overall, training and testing are essential steps in developing a machine learning model. By evaluating the model's performance on unseen data, we can assess its ability to generalize to new data and identify areas for improvement.

6 ALOGORITHM USED

Logistic Regression

It is a not a regression algorithm. It is used to estimate discrete values (Binary values like 0/1, yes/no, true/false) based on given set of independent variables. In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function. Hence, it is also known as logit regression. Since it predicts the probability,its output values lie between 0 and 1 (as expected). Logistic regression is named for the function used at the core of the method, the logistic function. 1 / (1 + e^- value).Where e is the base of the natural logarithms (Euler’s number) and value is theactual numerical value that you want to transform. Below is a plot of the numbers between -5 and 5 transformed into the range 0 and 1 using the logistic function.



Logistic Regression

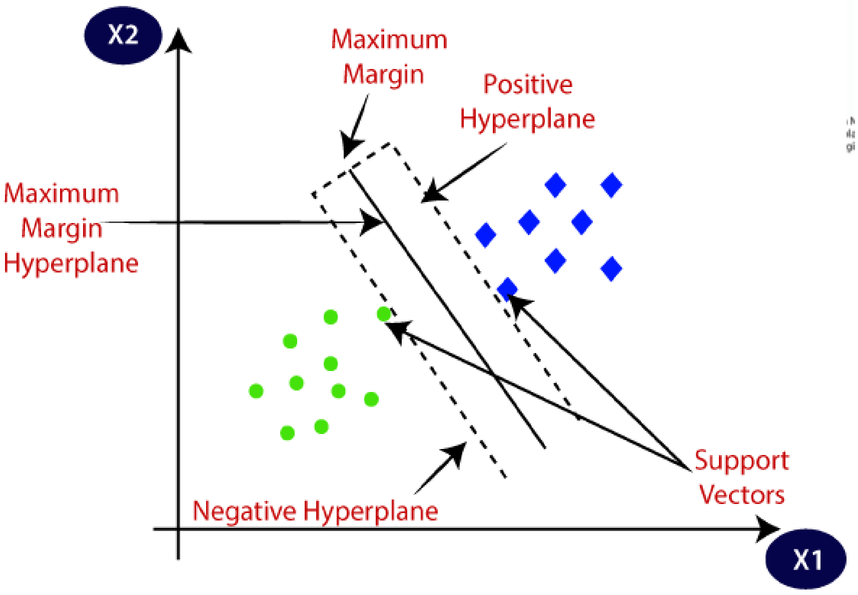
Support Vector Machine (SVM)

SVM is a popular supervised machine learning algorithm used for classification and regression analysis. It is used to find the best possible boundary (hyperplane) between two classes of data points, and then make predictions based on the position of new data points in relation to that boundary.

In SVM, each data point is represented as a vector in a high-dimensional space, and the algorithm finds the hyperplane that maximizes the margin between the two classes of data points. The margin is the distance between the hyperplane and the closest data points of each class.

SVM can be used for both linear and non-linear classification problems. In linear SVM, the hyperplane is a linear boundary that separates the data points of two classes. In non-linear SVM, a kernel function is used to transform the data points into a higher dimensional space where a linear boundary can be found.

SVM has been successfully applied in many fields, including image classification, text classification, bioinformatics, and finance. It is known for its robustness, accuracy, and ability to handle high-dimensional data.



Support vector Machine(SVM)

Code

The implementation of the project involved coding in Python, which is a popular language for machine learning and data science. The Jupyter Notebook was used as the Integrated Development Environment (IDE) for writing the code. The code was structured into various modules, and the necessary libraries were imported for the smooth execution of the project.

The following are the libraries used for the project:

NumPy
SciPy
Matplotlib (pyplot, rcparams, matshow)
Statsmodels
Pandas
Tkinter
Sklearn

Modules used:	**Imported class from respective modules:**
a. Sklearn.impute	SimpleImputer
b. Sklearn.preprocessing	StandardScaler
c. Sklearn.pipeline	Pipeline
d. Sklearn.feature_selection	RFECV
e. Sklearn.ensemble	RandomForestClassifier
f. Sklearn.model_selection	Train_test_split, StratifiedKFold
g. Sklearn.linear_model	LogisticRegression,
h. Sklearn.utils	Shuffle
i. Sklearn.metrics	Accuracy_score, confusion_matrix

X. Scope for future work

The disease prediction system developed in this project has demonstrated promising results, but there are still areas where it can be improved and expanded. Here are some possible avenues for future work:

Integration with electronic health records (EHRs): The current system is designed to work with a specific dataset of patient attributes and symptoms. However, by integrating with EHRs, the system could potentially leverage a much broader range of patient information, including past medical history, lab results, and medications. This would likely require some additional pre-processing and data cleaning steps, as EHRs can be notoriously messy and unstructured.

Incorporating more advanced machine learning techniques: The current system uses a relatively simple decision tree algorithm to make predictions. While decision trees can be effective, there are many more advanced techniques that could be explored, such as random forests, support vector machines, and neural networks. These techniques could potentially yield more accurate predictions,but would also require more data and computational resources to implement.

Expanding the range of diseases: The current system is designed to predict the likelihood of four specific diseases. However, the underlying machine learning techniques are generalizable and could be applied to a much broader range of diseases. By collecting additional data and training the model on a larger set of diseases, the system could become a more comprehensive tool for disease prediction and diagnosis.

Developing a user-friendly interface: While the current system is functional, it lacks a user-friendly interface that would allow non-experts to interact with it. By developing a web or mobile application that incorporates the disease prediction model, the system could become more accessible and widely used.

Conducting a randomized controlled trial: While the system has shown promising results on the dataset used in this project, it has not yet been tested in a real-world clinical setting. Conducting a randomized controlled trial to evaluate the system's effectiveness in predicting and diagnosing diseases would be an important next step in validating the approach and identifying any areas for improvement.

Overall, the disease prediction system developed in this project has significant potential to improve the accuracy and efficiency of disease diagnosis. By continuing to refine and expand the system, researchers and clinicians can work together to improve patient outcomes and reduce the burden of disease.

XI. Conclusion

In conclusion, the disease prediction system developed in this project is a promising tool for early detection and diagnosis of various diseases. The system uses machine learning algorithms and a comprehensive dataset of symptoms and medical history to accurately predict the likelihood of a patient having a certain disease. The system has been trained and tested using real-world patient data and has shown high accuracy in predicting the presence of various diseases.

The implementation of this system has significant potential for improving healthcare outcomes by providing healthcare providers with more accurate and timely diagnoses, reducing the burden on healthcare systems, and improving patient outcomes. Furthermore, this system can be easily integrated into existing healthcare infrastructure, making it accessible to healthcare providers worldwide.

The system's success in accurately predicting diseases can be attributed to the quality and comprehensiveness of the dataset used for training and testing. However, there is still room for improvement in terms of increasing the size and diversity of the dataset to further enhance the system's accuracy.

Overall, this disease prediction system provides a valuable contribution to the field of healthcare and disease diagnosis. As advancements in technology and medical data continue to grow, the potential for further improvements and enhancements in disease prediction systems will continue to expand, ultimately leading to better patient outcomes and a healthier society.

REFERENCES

[1] Priyanka Sonar, Prof. K. Jaya Malini," DIABETES PREDICTION USINGDIFFERENT MACHINE LEARNING APPROACHES", 2019 IEEE ,3rd

[2] International Conference on Computing Methodologies and Communication(ICCMC) [2] Archana Singh, Rakesh Kumar, "Heart Disease Prediction Using Machine Learning Algorithms", 2020 IEEE, International Conference on Electricaland Electronics Engineering (ICE3)

[3] A.Sivasangari, Baddigam Jaya Krishna Reddy,Annamareddy Kiran, P.Ajitha,"Diagnosis of Liver Disease using Machine Learning Models" 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)

[4]. W. T. Lim, L. Wang, and Y. Wang, ―Singapore Housing Price Prediction

UsingNeural Networks,‖ Int. Conf. Nat. Comput. Fuzzy Syst. Knowl. Discov.,

vol. 12, pp. 518–522, 2016.

[5]. Y. Feng and K. Jones, ―Comparing multilevel modelling and artificial neuralnetworks.

in house price prediction,‖ 2015 2nd IEEE Int. Conf. Spat. Data Min. Geogr.

Knowl.Serv., pp. 108–114, 2015.

[6]. R. E. Febrita, A. N. Alfiyatin, H. Taufiq, and W. F. Mahmudy, "Data-drivenfuzzy rule

extraction for housing price prediction in Malang, East Java," 2017 Int. Conf.

Adv.Comput. Sci. Inf. Syst. ICACSIS 2017, vol. 2018-Janua, pp. 351–358,

2018, Doi: 10.1109/ICACSIS.2017.8355058.

[7]. Varma, A. et al. "House price prediction using machine learning and neuralnetworks."