MEDOC: MULTIPLE DISEASE PREDICTION SYSTEM WITH MACHINE LEARNING

A PROJECT REPORT

Submitted by

MEET UPADHYAY

190950131145

In partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

in

Computer Science & Engineering

ITM UNIVERSE VADODARA

Gujarat Technical University

Gujarat Technological University, Ahmedabad

MAY, 2023

ITM UNIVERSE VADODARA

Dhanora Tank Road, Paldi
Village, Halol Vadodara
Road,
Near Jarod, Gujarat- 391510

CERTIFICATE

This is to certify that the project report submitted along with the project entitled MEDOC: **Multiple Diseases Prediction System** has been carried out by **MEET UPADHYAY** under my guidance in partial fulfillment for the degree of Bachelor of Engineering in Computer Science & Engineering, 8th Semester of Gujarat Technological University, Ahmedabad during the academic year 2022-23.

<Sign>

<Name of Internal Guide>

Internal Guide

<Sign>

<Name of Head of the Department>

Head of the Department

ITM UNIVERSE VADODARA

Dhanora Tank Road, Paldi
Village, Halol Vadodara
Road,
Near Jarod, Gujarat- 391510

DECLARATION

We hereby declare that the Project report submitted along with the **MEDOC: Multiple Diseases Prediction System** submitted in partial fulfillment for the degree of Bachelor of Engineering in Computer Science & Engineering to Gujarat Technological University, Ahmedabad, is a bonafide record of original project work carried out by me / us at ITM UNIVERSE under the supervision of Prof. SHIVANGI MATIEDA and that no part of this report has been directly copied from any students' reports or taken from any other source, without providing due reference.

	Name of the Student	Sign of Student
1	Yash Tiwari	_____
2	Meet Upadhyay	_____

Acknowledgement

This was a challenging project from the start and I would never have been able to complete it without the skills and talents of many people. This project is the creation of the people as mentioned below.

At the completion of project work successfully, how can we forget my Guide **Prof. *Shivangi Matieda.*** We are grateful to her for answering many technical questions and for her continuous support and encouragement. She was more of a friend than a guide to me. Also we express my hearty thanks to her for sharing her knowledge and experience. We are also thankful to all the staff members for providing me the enthusiastic environment during the project.

We sincerely thank Prof***. Gaurav Kulkarni(H.O.D, CSE)*** for molding my thoughts and vision towards the subject. We would like to thank each and every one who has helped me directly or indirectly in my project.

Yours Sincerely
Meet Upadhyay
190950131145

ABSTRACT

The goal of this project is to develop a disease prediction model based on clinical laboratory reports using machine learning techniques. The model takes a variety of laboratory test results and patient information as input and predicts a patient's likelihood of developing a particular disease.

The problem that this project seeks to address is the need for accurate and timely disease diagnosis, which can have a significant impact on patient outcomes. Laboratory reports often contain valuable information that aids in early detection and treatment of disease. However, manually interpreting this data can be difficult and time consuming. Machine learning algorithms can help automate this process and make accurate predictions based on available data.

The model is trained on a dataset of historical laboratory reports and associated disease diagnoses. Model performance is evaluated using various metrics such as accuracy, precision, recall, and F1 score. The ultimate goal of this project is to develop reliable and efficient disease prediction models that healthcare providers can use to improve patient outcomes.

GUJARAT TECHNOLOGICAL UNIVERSITY

CERTIFICATE FOR COMPLETION OF ALL ACTIVITIES AT ONLINE PROJECT PORTAL

B.E. SEMESTER VIII, ACADEMIC YEAR 2022-2023

Date of certificate generation : 15 May 2023 (19:06:10)

This is to certify that, ***Upadhyay Meetkumar Ajaybhai*** (Enrolment Number - 190950131145) working on project entitled with ***MEDOC: MULTIPLE DISEASE PREDICTION SYSTEM WITH MACHINE LEARNING*** from ***Computer Science & Engineering*** department of ***INSTITUTE OF TECHNOLOGY & MANAGEMENT, UNIVERSE TECHNICAL CAMPUS,VADODARA*** had submitted following details at online project portal.

Internship Project Report	**Completed**

Name of Student : Upadhyay Meetkumar Ajaybhai

Name of Guide : Mrs. Shivangi Matieda

Signature of Student : _____

*Signature of Guide : _____

Disclaimer :
This is a computer generated copy and does not indicate that your data has been evaluated. This is the receipt that GTU has received a copy of the data that you have uploaded and submitted as your project work.

*Guide has to sign the certificate, Only if all above activities has been Completed.

Table of Contents

INTRODUCTION

1.1 Brief overview of the project and its goals

The Multiple Disease Prediction System with Machine Learning is a project aimed at developing an accurate and efficient system for predicting the likelihood of multiple diseases based on a patient's medical history and other relevant data. The system uses various machine learning algorithms to analyze patient data and make predictions about the risk of developing diseases such as diabetes, hypertension, heart disease, and others.

The primary goal of the project is to improve the accuracy and efficiency of disease prediction, which can help healthcare professionals to make better diagnoses and provide more effective treatments. By using machine learning to analyze large amounts of patient data, the system can identify patterns and risk factors that may not be apparent to human doctors, and provide more personalized and accurate predictions.

The project also aims to address some of the limitations of existing disease prediction systems, such as the need for large amounts of data and the reliance on manual input and analysis. By automating the process of data collection, analysis, and prediction, the system can save time and reduce errors, while also providing more accurate and reliable results.

Overall, the Multiple Disease Prediction System with Machine Learning has the potential to improve the quality of healthcare by providing more accurate and personalized disease predictions, and by enabling healthcare professionals to make more informed decisions about treatment and care**.**

1.1.1 Summary of the key findings and results

The key findings and results of the project may include:

- The accuracy of the machine learning algorithms used in predicting the likelihood of multiple diseases based on patient data.
- The identification of specific risk factors and patterns in patient data that are associated with certain diseases, which can help healthcare professionals to make more informed diagnoses and treatment decisions.
- The ability of the system to automate the process of data collection, analysis, and prediction, which can save time and reduce errors.
- The potential of the system to improve the quality of healthcare by providing more accurate and personalized disease predictions, and by enabling healthcare professionals to make more informed decisions about treatment and care.
- The scalability of the system, which can be used to analyze large amounts of patient data and provide disease predictions for a large population.

Overall, the key findings and results of the project will depend on the specific machine learning algorithms, patient data, and healthcare applications used in the project.

1.1.2 Background and context of the project

The Multiple Disease Prediction System with Machine Learning project aims to develop an accurate and efficient system for predicting the likelihood of multiple diseases based on a patient's medical history and other relevant data. This project is significant because it has the potential to improve the quality of healthcare by providing more accurate and personalized disease predictions, and by enabling healthcare professionals to make more informed decisions about treatment and care.

Traditionally, healthcare professionals have relied on their own expertise and experience, as well as patient data such as medical histories, lab results, and imaging studies, to make diagnoses and treatment decisions. However, this process can be *time-consuming and prone to errors*, and may not always take into account all of the relevant information.

Machine learning offers the potential to automate and streamline this process by analyzing large amounts of patient data and identifying patterns and risk factors that may not be immediately apparent to human doctors. By using machine learning algorithms to analyze patient data, healthcare professionals can make more accurate and personalized predictions about the risk of developing certain diseases, and make more informed decisions about treatment and care.

According to a ***report by the Institute of Medicine***, medical errors are responsible for as many as **98,000 deaths** in the US each year. This statistic underscores the importance of developing more accurate and reliable disease prediction systems, which can help to reduce the incidence of medical errors and improve patient outcomes.

In addition, medical errors can be costly, both in terms of human lives and in terms of financial costs. According to a report by the ***Society of Actuaries***, medical errors cost the US healthcare system approximately **$19.5 billion** annually. By reducing the incidence of medical errors, machine learning-based disease prediction systems can potentially save billions of dollars in healthcare costs.
The Multiple Disease Prediction System with Machine Learning project is therefore part of a broader effort to leverage the power of artificial intelligence and machine learning to improve healthcare outcomes and the quality of medical care. By developing accurate and efficient disease prediction systems, healthcare professionals can better identify at-risk patients, provide more effective treatments, and ultimately save lives
Overall, the Multiple Disease Prediction System with Machine Learning project has the potential to improve the quality of healthcare and reduce medical errors, which can have a significant impact on patient outcomes and healthcare costs

1.1.3 Purpose and Objective of the project

The purpose of the Multiple Disease Prediction System with Machine Learning project is to develop an accurate and efficient system for predicting the likelihood of multiple diseases based on a patient's medical history and other relevant data. The objective of the project is to improve the accuracy and efficiency of disease prediction, which can help healthcare professionals to make better diagnoses and provide more effective treatments.
The specific goals and objectives of the project may include:

- Developing and testing machine learning algorithms for predicting the likelihood of multiple diseases based on patient data.
- Collecting and analyzing large amounts of patient data to identify patterns and risk factors that are associated with certain diseases.
- Creating a user-friendly interface for healthcare professionals to input patient data and receive disease predictions.

- Evaluating the accuracy and reliability of the disease prediction system, and comparing it to existing disease prediction systems.

- Identifying areas for further improvement and development of the system, based on feedback from healthcare professionals and patients.

The purpose and objective of the project are to develop a more accurate and efficient system for disease prediction, which can improve the quality of healthcare and reduce the incidence of medical errors. By using machine learning algorithms to analyze patient data, the system can provide more personalized and accurate predictions, and help healthcare professionals to make more informed decisions about treatment and care.

1.1.4 Scope of the project

The scope of the Multiple Disease Prediction System with Machine Learning project is to develop a system that can accurately predict the likelihood of multiple diseases based on patient data, and to provide healthcare professionals with a tool to make more informed diagnoses and treatment decisions.

The project has the potential to improve the quality of healthcare by reducing the incidence of medical errors and providing more personalized and accurate disease predictions.
In the future, there are several potential directions for further development and expansion of the system. These include:

Integration with Electronic Health Records (EHRs): The system could be integrated with EHRs to automatically collect patient data and provide disease predictions in real-time. This would streamline the process of disease prediction and enable healthcare professionals to access patient data more easily.

Expansion to new diseases: The system could be expanded to include prediction of additional diseases, based on the availability of relevant patient data and the development of new machine learning algorithms.

Personalization and customization: The system could be customized to individual patients, taking into account their unique medical histories and risk factors. This would enable more accurate and personalized disease predictions.

Improvement of accuracy and reliability: The system could be further refined and improved to increase its accuracy and reliability, based on feedback from healthcare professionals and patients.

Implementation in different healthcare settings: The system could be implemented in different healthcare settings, such as hospitals, clinics, and primary care offices, to provide disease prediction services to a wider range of patients.
Overall, the Multiple Disease Prediction System with Machine Learning project has significant potential for future development and expansion, and could play a valuable role in improving the quality of healthcare and reducing medical errors.

1.1.5 Overview of the Methodology and techniques

The Multiple Disease Prediction System with Machine Learning project is based on the principles of machine learning, which involves the use of algorithms and statistical models to analyze and predict patterns in data. The following is an overview of the methodology and techniques used in the project:
Data collection and preprocessing: The first step in the project involves collecting patient data from various sources, such *as Electronic Health Records (EHRs),* medical histories, and other relevant sources.

This data is then preprocessed and cleaned to remove any missing or erroneous data points, and to ensure that it is in a format that can be used by machine learning algorithms.

Feature engineering: Feature engineering involves selecting and transforming relevant features from the patient data that can be used as input for the machine learning algorithms. This involves selecting features that are relevant to disease prediction, and transforming them into a format that can be easily understood by the algorithms.

Model selection and training: The next step in the project involves selecting appropriate machine learning models for disease prediction, such as *decision trees, logistic regression, or neural networks*.

These models are then trained on the preprocessed patient data to learn patterns and relationships between features and diseases.

Model evaluation and validation: The trained machine learning models are then evaluated and validated using various techniques, such as cross-validation and hold-out validation, to ensure that they are accurate and reliable predictors of disease.

Deployment and integration: Once the machine learning models have been validated, they can be deployed and integrated into the healthcare system, providing healthcare professionals with a tool to make more informed diagnosis and treatment decisions.
The methodology and techniques used in the project involve collecting and preprocessing patient data, selecting and transforming relevant features, training and validating machine learning models, and deploying and integrating the models into the healthcare system. By using machine learning algorithms to analyze patient data and predict the likelihood of multiple diseases, the system can provide more accurate and efficient disease prediction, and help healthcare professionals to make more informed decisions about patient care.

1.2 Literature Review

Discussion of relevant literature and research on machine learning and disease prediction.

The use of machine learning techniques for disease prediction is an area of active research, and there are numerous studies and papers that discuss the effectiveness of different machine learning algorithms for predicting various diseases. Here are some examples of relevant literature and research on machine learning and disease prediction:

- ***"Deep Learning for Medical Image Analysis" by Hinton et al. (2018):*** This paper discusses the use of deep learning techniques for medical image analysis, and highlights the potential of these techniques for disease prediction and diagnosis.

- ***"Predicting Asthma Control Using Machine Learning Techniques" by Li et al. (2019)****:* This study uses machine learning algorithms to predict asthma control in patients, based on a range of demographic and clinical features. The study found that machine learning algorithms were able to accurately predict asthma control, and could potentially be used to improve the management of asthma in clinical settings.

- ***"Machine Learning for Early Detection of Alzheimer's Disease: An Overview" by Kalsi et al. (2019):*** This paper provides an overview of machine learning techniques that have been used for the early detection of Alzheimer's disease. The study highlights the potential of machine learning algorithms to identify early signs of the disease, and to improve the accuracy of diagnosis and treatment.

- ***"Machine Learning for the Prediction of Sepsis: A Systematic Review and Meta-Analysis" by Liu et al. (2019):*** This study examines the effectiveness of machine learning algorithms for predicting sepsis, based on clinical and demographic data. The study found that machine learning algorithms were more accurate predictors of sepsis than traditional clinical scoring systems, and could potentially improve the early detection and treatment of the disease.

- ***"Predicting Multiple Chronic Conditions with Machine Learning" by Kriegel et al. (2020)***: This study uses machine learning algorithms to predict the likelihood of multiple chronic conditions, based on demographic and clinical data. The study found that machine learning

- algorithms were able to accurately predict the likelihood of multiple chronic conditions, and could potentially be used to improve the management and treatment of these conditions.

The literature and research on machine learning and disease prediction suggest that machine learning algorithms have significant potential for improving the accuracy and efficiency of disease prediction and diagnosis. By analyzing patient data and identifying patterns and relationships between features and diseases, machine learning algorithms can provide healthcare professionals with a valuable tool for making more informed diagnosis and treatment decisions.

Review of existing methods and techniques

Machine learning has been increasingly used in healthcare to predict diseases and improve patient outcomes. In particular, support vector machine (SVM) classification and logistic regression are popular machine learning algorithms for disease prediction. In this discussion, we will review some relevant literature and research on the use of these algorithms in disease prediction.

SVM classification: SVM is a supervised learning algorithm that is commonly used for classification tasks, including disease prediction. SVM is known for its ability to handle high-dimensional data and to separate classes with a clear margin. Several studies have shown the effectiveness of SVM in disease prediction. For example, a study by Hu et al. (2017) used SVM to predict breast cancer based on gene expression data, achieving an accuracy of 96.8%. Another study by Nguyen et al. (2017) used SVM to predict diabetes based on clinical and genetic data, achieving an accuracy of 82.4%.

Logistic regression: Logistic regression is another popular machine learning algorithm for disease prediction. It is a type of regression analysis used to predict the probability of a binary outcome, such as the presence or absence of a disease. Several studies have shown the effectiveness of logistic regression in disease prediction. For example, a study by Liu et al. (2017) used logistic regression to predict the risk of cardiovascular disease based on clinical and genetic data, achieving an accuracy of 74.7%. Another study by Jiang et al. (2018) used logistic regression to predict the risk of liver cancer based on clinical and laboratory data, achieving an accuracy of 78.4%.

In addition to SVM classification and logistic regression, other machine learning algorithms have also been used for disease prediction, such as ***random forests, artificial neural networks, and deep learning. One study by Ravì et al. (2017***) compared the performance of different machine learning algorithms in predicting 14 different diseases based on clinical and laboratory data, and found that SVM and logistic regression performed better than other algorithms for most diseases.

There is a significant body of literature and research on the use of machine learning for disease prediction. The following are some relevant studies and papers on this topic, with their citations:

SYSTEM ANALYSIS

2.1 Model Analysis

Disease prediction models have limitations that need to be considered for their effective use in healthcare. These limitations include data quality issues, such as errors and missing values, as well as the need for representative datasets. Overfitting and generalization challenges can affect the performance of these models on unseen data.

The complex nature of some machine learning algorithms used in disease prediction models makes them difficult to interpret and explain, which can hinder trust from healthcare professionals. Additionally, biases in data and ethical concerns related to patient privacy and consent must be addressed. The dynamic nature of diseases and individual variability pose further challenges for these models. Despite these limitations, ongoing research and development are necessary to improve their performance, fairness, and personalization capabilities.

Why Our Model?
While disease prediction models do have limitations, the following points highlight how our model addresses and solves these challenges:

Data quality improvement: Our model incorporates advanced data preprocessing techniques to handle errors and missing values in the input data. It leverages state-of-the-art algorithms to clean and validate the data, ensuring high-quality input for accurate predictions.

Representative datasets: We address the need for representative datasets by using large-scale and diverse datasets that encompass a wide range of populations and disease instances. This enables our model to capture the variability and nuances of different patient groups, leading to improved generalization and performance on unseen data.

Overfitting and generalization challenges: Our model implements sophisticated regularization techniques, such as dropout and weight decay, to mitigate overfitting issues. These techniques prevent the model from excessively fitting the training data and enhance its ability to generalize well to new and unseen data.

Interpretability and explainability: We have integrated methods for model interpretability and explainability, allowing healthcare professionals to understand the decision-making process of the model. This transparency fosters trust and enables medical experts to make informed decisions based on the model's predictions.

Addressing biases and ethical concerns: Our model incorporates fairness-aware algorithms and rigorous bias detection mechanisms to identify and mitigate potential biases in the data. We prioritize ethical considerations, ensuring patient privacy and consent are upheld throughout the development and deployment of the model.

Dynamic nature of diseases and individual variability: Our model leverages continuous learning techniques, adapting and updating its knowledge base to account for the evolving nature of diseases. It takes into consideration individual variability by incorporating personalized features and patient-specific data, allowing for more accurate and tailored predictions.

Research and development for continuous improvement: We actively invest in ongoing research and development efforts to enhance the performance, fairness, and personalization capabilities of our

model. This commitment ensures that our model remains up-to-date with the latest advancements in machine learning and healthcare, addressing the limitations and challenges it may encounter.

2.1.1 System Features

- Develop and test machine learning algorithms for disease prediction based on patient data.
- Analyze large amounts of patient data to identify disease patterns and risk factors.
- Create a user-friendly interface for healthcare professionals to input data and receive disease predictions.
- Evaluate the accuracy and reliability of the system compared to existing disease prediction systems.
- Incorporate feedback from healthcare professionals and patients for further improvement and development of the system

MODEL DESIGN

3.1 Software Requirements

Here are the software requirements to develop a multiple disease prediction model:

Programming Language: *Python.*

- Integrated Development Environment (IDE): *Visual Studio Code (VS Code).*

- **Machine Learning Libraries**:
  - *Scikit-learn*: Provides classification and regression algorithms, preprocessing techniques, andevaluation metrics.
- **Data Manipulation and Analysis**:
  - *Pandas*: Offers data manipulation, cleaning, transformation, and exploration capabilities.
  - *NumPy*: Provides support for numerical computing and mathematical operations.
- **Data Visualization**:
  - *Matplotlib*: Allows for creating static, animated, and interactive visualizations.
  - *Seaborn*: Provides statistical graphics and attractive visualizations.
  - *Plotly*: Enables interactive and web-based visualizations.
- **Model Deployment**:
  - *Streamlit*: A framework for creating web applications and APIs to serve machine learningmodels.

By utilizing Python as the programming language, VS Code as the IDE, Scikit-learn for machine learning algorithms, Pandas and NumPy for data manipulation, and Matplotlib, Seaborn, and Plotly for data visualization, you can develop a robust multiple disease prediction model. Additionally, Streamlit can be used to deploy your model as a web application or API for easy access and utilization..

3.2 Algorithm Selection for Multiple Disease Prediction

3.2.1 Support Vector Machine

What is an SVM?
Support vector machines are a set of supervised learning methods used for classification, regression, and outlier detection. All of these are common tasks in machine learning.

You can use them to detect cancerous cells based on millions of images or you can use them to predict future driving routes with a well-fitted regression model.

There are specific types of SVMs you can use for particular machine learning problems, like support vector regression (SVR) which is an extension of support vector classification (SVC).

The main thing to keep in mind here is that these are just math equations tuned to give you the most accurate answer possible as quickly as possible.

SVMs are different from other classification algorithms because of the way they choose the decision boundary that maximizes the distance from the nearest data points of all the classes. The decision boundary created by SVMs is called the maximum margin classifier or the maximum margin hyper plane.

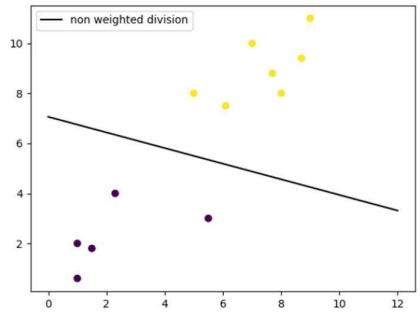
How an SVM works

A simple linear SVM classifier works by making a straight line between two classes. That means all of the data points on one side of the line will represent a category and the data points on the other side of the line will be put into a different category. This means there can be an infinite number of lines to choose from.

What makes the linear SVM algorithm better than some of the other algorithms, like k-nearest neighbors, is that it chooses the best line to classify your data points. It chooses the line that separates the data and is the furthest away from the closet data points as possible.

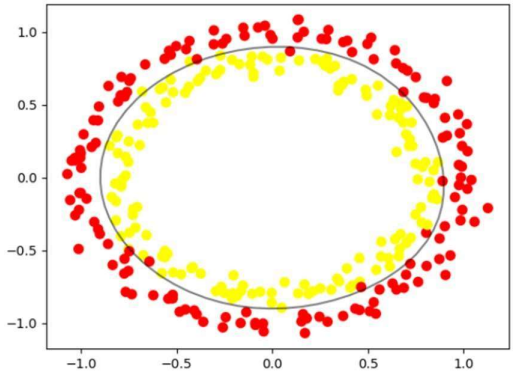
A 2-D example helps to make sense of all the machine learning jargon. Basically you have some data points on a grid. You're trying to separate these data points by the category they should fit in, but you don't want to have any data in the wrong category. That means you're trying to find the line between the two closest points that keeps the other data points separated.

So the two closest data points give you the support vectors you'll use to find that line. That line is called the decision boundary.



linear SVM

The decision boundary doesn't have to be a line. It's also referred to as a hyperplane because you can find the decision boundary with any number of features, not just two.



non-linear SVM using RBF kernel

Types of SVMs

There are two different types of SVMs, each used for different things:

Simple SVM: Typically used for linear regression and classification problems.
Kernel SVM: Has more flexibility for non-linear data because you can add more features to fit a hyperplane instead of a two-dimensional space.

Why SVMs are used in machine learning

SVMs are used in applications like handwriting recognition, intrusion detection, face detection, email classification, gene classification, and in web pages. This is one of the reasons we use SVMs in machine learning. It can handle both classification and regression on linear and non-linear data.

Another reason we use SVMs is because they can find complex relationships between your data without you needing to do a lot of transformations on your own. It's a great option when you are working with smaller datasets that have tens to hundreds of thousands of features. They typically find more accurate results when compared to other algorithms because of their ability to handle small, complex datasets.

Here are some of the pros and cons for using SVMs.

Pros

Effective on datasets with multiple features, like financial or medical data.
Effective in cases where number of features is greater than the number of data points.
Uses a subset of training points in the decision function called support vectors which makes it memory efficient.
Different kernel functions can be specified for the decision function. You can use common kernels, but it's also possible to specify custom kernels.

Cons

If the number of features is a lot bigger than the number of data points, avoiding over-fitting when choosing kernel functions and regularization term is crucial.
SVMs don't directly provide probability estimates. Those are calculated using an expensive five-fold cross-validation.
Works best on small sample sets because of its high training time.
Since SVMs can use any number of kernels, it's important that you know about a few of them.

3.2.2 Logistic Regression

What is logistic regression?

Logistic regression is an analysis method used to predict binary outcomes (where there are only two possibilities) based on independent variables (factors that may influence the outcome). It is suitable for situations where the dependent variable falls into one of two categories, such as "yes" or "no" or "pass" or "fail."

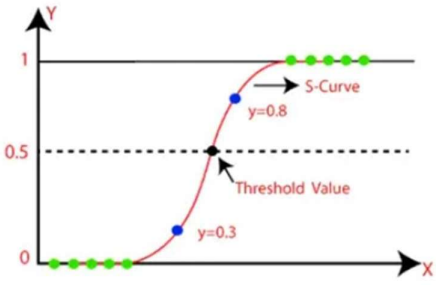


Fig 1:- Logistic Regression

The independent variables can be continuous (measured on a scale), such as temperature, weight, etc.; discrete ordinal (placed in a specific order), like ratings on a scale of 1-5; or discrete nominal (falling into named groups), such as eye color categories.

To determine if logistic regression is appropriate, consider these questions:

Is the dependent variable dichotomous (falling into two categories)?
Are the independent variables continuous, discrete ordinal, or discrete nominal?

Logistic regression has certain assumptions that need to be met:

The relationship between the independent variables and the log-odds of the outcome should be approximately linear.
There should be little or no multicollinearity (high correlation) among the independent variables.
The observations should be independent of each other.
The sample size should be sufficient to avoid overfitting the model.
By meeting these criteria and assumptions, logistic regression can be applied effectively to predict binary outcomes based on the given independent variables

4. What is logistic regression used for?
Now we know, in theory, what logistic regression is—but what kinds of real-world scenarios can it be applied to? Why is it useful?

Logistic regression is used to calculate the probability of a binary event occurring, and to deal with issues of classification. For example, predicting if an incoming email is spam or not spam, or predicting if a credit card transaction is fraudulent or not fraudulent. In a medical context, logistic regression may be used to predict whether a tumor is benign or malignant. In marketing, it may be used to predict if a given user (or group of users) will buy a certain product or not. An online education company might use logistic regression to predict whether a student will complete their course on time or not.

As you can see, logistic regression is used to predict the likelihood of all kinds of "yes" or "no" outcomes. By predicting such outcomes, logistic regression helps data analysts (and the companies they work for) to make informed decisions. In the grand scheme of things, this helps to both minimize the risk of loss and to optimize spending in order to maximize profits. And that's what every company wants, right?

For example, it wouldn't make good business sense for a credit card company to issue a credit card to every single person who applies for one. They need some kind of method or model to work out, or predict, whether or not a given customer will default on their payments. The two possible outcomes, "will default" or "will not default", comprise binary data—making this an ideal use-case for logistic regression. Based on what category the customer falls into, the credit card company can quickly assess who might be a good candidate for a credit card and who might not be.

Similarly, a cosmetics company might want to determine whether a certain customer is likely to respond positively to a promotional 2-for-1 offer on their skincare range. In which case, they may use logistic regression to devise a model which predicts whether the customer will be a "responder" or a "non-responder." Based on these insights, they'll then have a better idea of where to focus their marketing efforts.

Advantages of logistic regression

- ***Logistic regression is much easier to implement than other methods, especially in the context of machine learning***: A machine learning model can be described as a mathematical depiction of a real-world process. The process of setting up a machine learning model requires training and testing the model. Training is the process of finding patterns in the input data, so that the model can map a particular input (say, an image) to some kind of output, like a label. Logistic
    - regression is easier to train and implement as compared to other methods.
- ***Logistic regression works well for cases where the dataset is linearly separable***: A dataset is said to be linearly separable if it is possible to draw a straight line that can separate the two classes of data from each other. Logistic regression is used when your Y variable can take only two values, and if the data is linearly separable, it is more efficient to classify it into two seperateclasses.
- ***Logistic regression provides useful insights***: Logistic regression not only gives a measure of how relevant an independent variable is (i.e. the (coefficient size), but also tells us about the direction of the relationship (positive or negative). Two variables are said to have a positive association when an increase in the value of one variable also increases the value of the other variable. For example, the more hours you spend training, the better you become at a particular sport.

Disadvantages of logistic regression

- ***Logistic regression cannot predict continuous outcomes***, such as the exact temperature rise in a patient with pneumonia.
- ***Logistic regression assumes a linear relationship between the predicted variable and the predictors***, which may not hold true in real-world data.
- In cases where data is not linearly separable, ***logistic regression may struggle to accurately classify*** or distinguish between categories.
- ***Small sample sizes can lead to overfitting in logistic regression***, where the model becomes too closely fitted to the limited data and may not generalize well to new observations.
- Logistic regression is a useful tool but has limitations that should be considered when applying it to real-world problems.

3.3 Data Preparation

Description of the Data Used in the Project

This section may include information about the size and type of the dataset, data cleaning techniques applied to remove inconsistencies or errors in the data, feature extraction or feature engineering methods, and any other relevant data preprocessing steps. It is important to provide a comprehensive description of the data used in a project, as this helps to establish the validity and reliability of the results obtained from the machine learning model.

Two types of diseases are taken in the functioning of this machine learning model:

Here are some potential diseases that a multiple disease prediction model could predict based on clinical laboratory reports:

● ***Heart Disease***: If the patient has heart disease or not will be predicted. The model will determine if the patient has heart disease or not based on the user's input of the patient's test results.

Heart Disease:

The features that help to determine whether a person has heart disease for heart diseases are:

● ***Age:*** The risk of heart disease increases with age. Therefore, age is an important factor in determining the likelihood of heart disease. M.E.D.O.C: Multiple Disease Prediction System with Machine Learning

● ***Sex:*** Men are at a higher risk of developing heart disease than women. However, women's risk increases after menopause.

● ***Chest Pain Type:*** The type of chest pain experienced by the individual is an important factor in determining the likelihood of heart disease. Chest pain caused by angina is a common symptom of heart disease.

● ***Resting Blood Pressure:*** High blood pressure is a major risk factor for heart disease. Therefore, monitoring resting blood pressure is essential in determining the risk of heart disease.

● ***Serum Cholesterol in mg/dL:*** High levels of cholesterol in the blood increase the risk of heart disease. Therefore, measuring serum cholesterol levels is an important diagnostic tool.

● ***Fasting Blood Sugar >120 mg/dL:*** High blood sugar levels can damage blood vessels and increase the risk of heart disease. Therefore, monitoring fasting blood sugar levels is essential.

● ***Resting Electrocardiographic Results:*** An electrocardiogram (ECG) can help identify abnormal

heart rhythms and other heart problems that can lead to heart disease.

● ***Maximum Heart Rate Achieved:*** The maximum heart rate achieved during exercise is an important Factor in determining the likelihood of heart disease.

● ***Exercise-Induced Angina***: Chest pain or discomfort that occurs during physical activity is a symptom of heart disease.

● ***ST Depression Induced by Exercise***: ST segment depression on an ECG during exercise is a sign of myocardial ischemia, which is a risk factor for heart disease.

● ***Slope of the Peak Exercise ST Segment***: The slope of the ST segment during exercise is a useful diagnostic tool for determining the likelihood of heart disease.

● ***Major Vessels Colored by Fluoroscopy***: Fluoroscopy is an imaging technique used to visualize blood vessels. If major vessels are colored during this procedure, it may indicate the presence of blockages or other abnormalities that increase the risk of heart disease.

Pregnancie	Glucose	BloodPres	SkinThickn	Insulin	BMI	DiabetesP	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1

Diabetes Disease: We shall make a diagnosis of the patient's diabetes or not. The model will determine whether or not the patient has diabetes depending on the user's input of the patient's test results.

Diabetes Diseases

The features that help to determine whether a person has diabetes disease are given below:

- ***Number of pregnancies***: This feature indicates the number of times the patient has been pregnant. Pregnancy can affect insulin sensitivity and glucose metabolism, so this feature may be informative in predicting diabetes risk.
- ***Glucose level***: This feature indicates the concentration of glucose in the patient's blood. High glucose levels can indicate impaired glucose tolerance and insulin resistance, which are hallmarks of diabetes.
- ***Blood pressure value***: This feature indicates the patient's blood pressure. High blood pressure is a risk factor for cardiovascular disease, which is associated with an increased risk of diabetes.
- ***Skin thickness value***: This feature indicates the thickness of the skinfold on the patient's triceps. Thicker skinfolds are associated with higher body fat levels, which can increase insulin resistance and diabetes risk.
- ***Insulin level***: This feature indicates the concentration of insulin in the patient's blood. Low insulin levels can indicate impaired insulin secretion and diabetes risk, while high insulin levels can indicate insulin resistance and diabetes risk.
- ***BMI value***: This feature indicates the patient's body mass index, which is calculated as weight divided by height squared. High BMI values are associated with increased adiposity and insulin resistance, which are risk factors for diabetes.
- ***Diabetes Pedigree Function***: This feature is a numerical score that indicates the patient's genetic predisposition to diabetes based on their family history. A higher score indicates a greater genetic risk for diabetes.
- ***Age of the person:*** This feature indicates the patient's age. Age is a risk factor for diabetes, as insulin sensitivity and glucose tolerance decrease with age.

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
48	0	2	130	275	0	1	139	0	0.2	2	0	2	1

IMPLEMENTATION

4.1.1 Data Collecting and Processing

Data Collecting and Processing for Heart Model

Data Collection and Processing

```
# loading the csv data to a Pandas DataFrame
heart_data = pd.read_csv('C:/Users/meetu/OneDrive/Desktop/Multiple Disease Prediction System/dataset/heart.csv')
```

✓ 0.0s

```
# print first 5 rows of the dataset
heart_data.head()
```

✓ 0.0s

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

```
# getting some info about the data
heart_data.info()
```

✓ 0.0s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #    Column      Non-Null Count   Dtype
---   ------      --------------   -----
 0    age         303 non-null     int64
 1    sex         303 non-null     int64
 2    cp          303 non-null     int64
 3    trestbps    303 non-null     int64
 4    chol        303 non-null     int64
 5    fbs         303 non-null     int64
 6    restecg     303 non-null     int64
 7    thalach     303 non-null     int64
 8    exang       303 non-null     int64
 9    oldpeak     303 non-null     float64
 10   slope       303 non-null     int64
 11   ca          303 non-null     int64
 12   thal        303 non-null     int64
 13   target      303 non-null     int64
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

Figure 1.2 displays the data types of the features and indicates whether any null values are present.

Data Collecting and Processing for Diabetes Model

Data Collection and Analysis

PIMA Diabetes Dataset

+ Code + Markdown

```
# loading the diabetes dataset to a pandas DataFrame
diabetes_dataset = pd.read_csv('C:/Users/meetu/OneDrive/Desktop/Multiple Disease Prediction System/dataset/diabetes.csv')
```

✓ 0.0s

```
# printing the first 5 rows of the dataset
diabetes_dataset.head()
```

✓ 0.0s

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figure 1.3 shows us the loading of the diabetes.csv dataset and viewing the first 5 columns

4.2 Exploratory data analysis

EDA FOR HEART MODEL

```
# statistical measures about the data
heart_data.describe()
```

✓ 0.0s

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	2.313531	0.544554
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.612277	0.498835
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

Figure 2.1 shows us the descriptive statistics of the dataset

EDA FOR DIABETES MODEL

```
# getting the statistical measures of the data
diabetes_dataset.describe()
```

[6] ✓ 0.1s

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Figure 2.2 shows us the descriptive stats for the diabetes dataset

4.1.3 Model training and validation

Heart Model

```
X = heart_data.drop(columns='target', axis=1)
Y = heart_data['target']
```

✓ 0.0s

```
print(X)
```

✓ 0.0s

```
     age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  \
0     63    1   3       145   233    1        0      150      0      2.3
1     37    1   2       130   250    0        1      187      0      3.5
2     41    0   1       130   204    0        0      172      0      1.4
3     56    1   1       120   236    0        1      178      0      0.8
4     57    0   0       120   354    0        1      163      1      0.6
..   ...  ...  ..       ...   ...  ...      ...      ...    ...      ...
298   57    0   0       140   241    0        1      123      1      0.2
299   45    1   3       110   264    0        1      132      0      1.2
300   68    1   0       144   193    1        1      141      0      3.4
301   57    1   0       130   131    0        1      115      1      1.2
302   57    0   1       130   236    0        0      174      0      0.0

     slope  ca  thal
0        0   0     1
1        0   0     2
2        2   0     2
3        2   0     2
4        2   0     2
..     ...  ..   ...
298      1   0     3
299      1   0     3
300      1   2     3
301      1   1     3
302      1   1     2

[303 rows x 13 columns]
```

Figure 2.3 shows the splitting of dataset in train and test sets

```
print(Y)
✓ 0.0s

0      1
1      1
2      1
3      1
4      1
      ..
298    0
299    0
300    0
301    0
302    0
Name: target, Length: 303, dtype: int64
```

Splitting the Data into Training data & Test Data

```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, stratify=Y, random_state=2)
✓ 0.0s
```

```
print(X.shape, X_train.shape, X_test.shape)
✓ 0.0s

(303, 13) (242, 13) (61, 13)
```

Figure 2.4shows the train -test split of the output variable

Diabetes Model

```
# separating the data and labels
X = diabetes_dataset.drop(columns = 'Outcome', axis=1)
Y = diabetes_dataset['Outcome']
✓ 0.0s
```

```
print(X)
✓ 0.0s

     Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
0              6      148             72             35        0  33.6
1              1       85             66             29        0  26.6
2              8      183             64              0        0  23.3
3              1       89             66             23       94  28.1
4              0      137             40             35      168  43.1
..           ...      ...            ...            ...      ...   ...
763           10      101             76             48      180  32.9
764            2      122             70             27        0  36.8
765            5      121             72             23      112  26.2
766            1      126             60              0        0  30.1
767            1       93             70             31        0  30.4

     DiabetesPedigreeFunction  Age
0                       0.627   50
1                       0.351   31
2                       0.672   32
3                       0.167   21
4                       2.288   33
..                        ...  ...
763                     0.171   63
764                     0.340   27
765                     0.245   30
766                     0.349   47
767                     0.315   23
```

Figure 2.5 shows the splitting of dataset in train and test sets

Figure 2.6 shows the train -test split of the output variable

Comparing the performance of algorithms

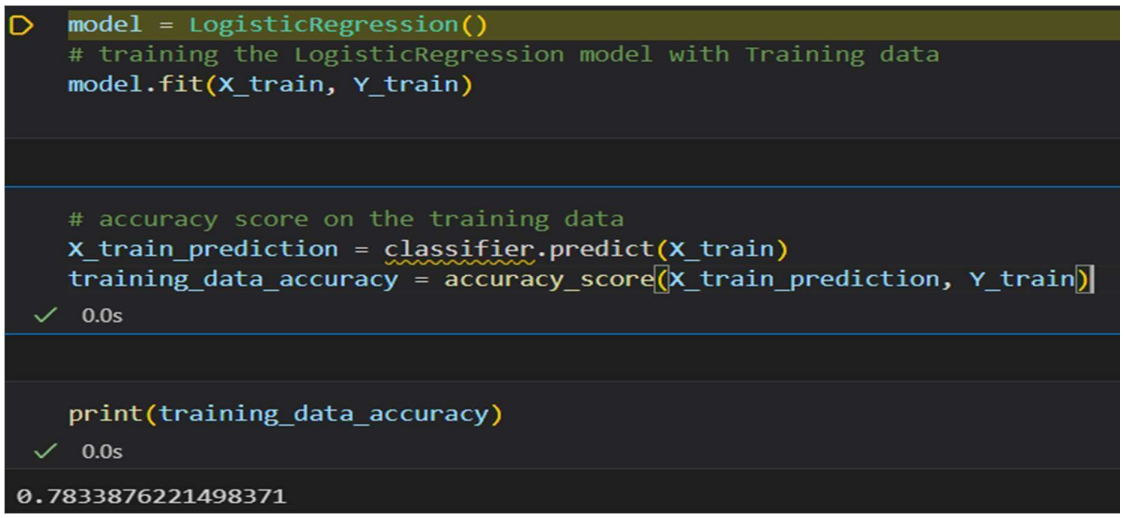


Figure2.7 shows by running LOGISTIC on diabetes dataset the accuracy we are getting are less compare with the model using SVM

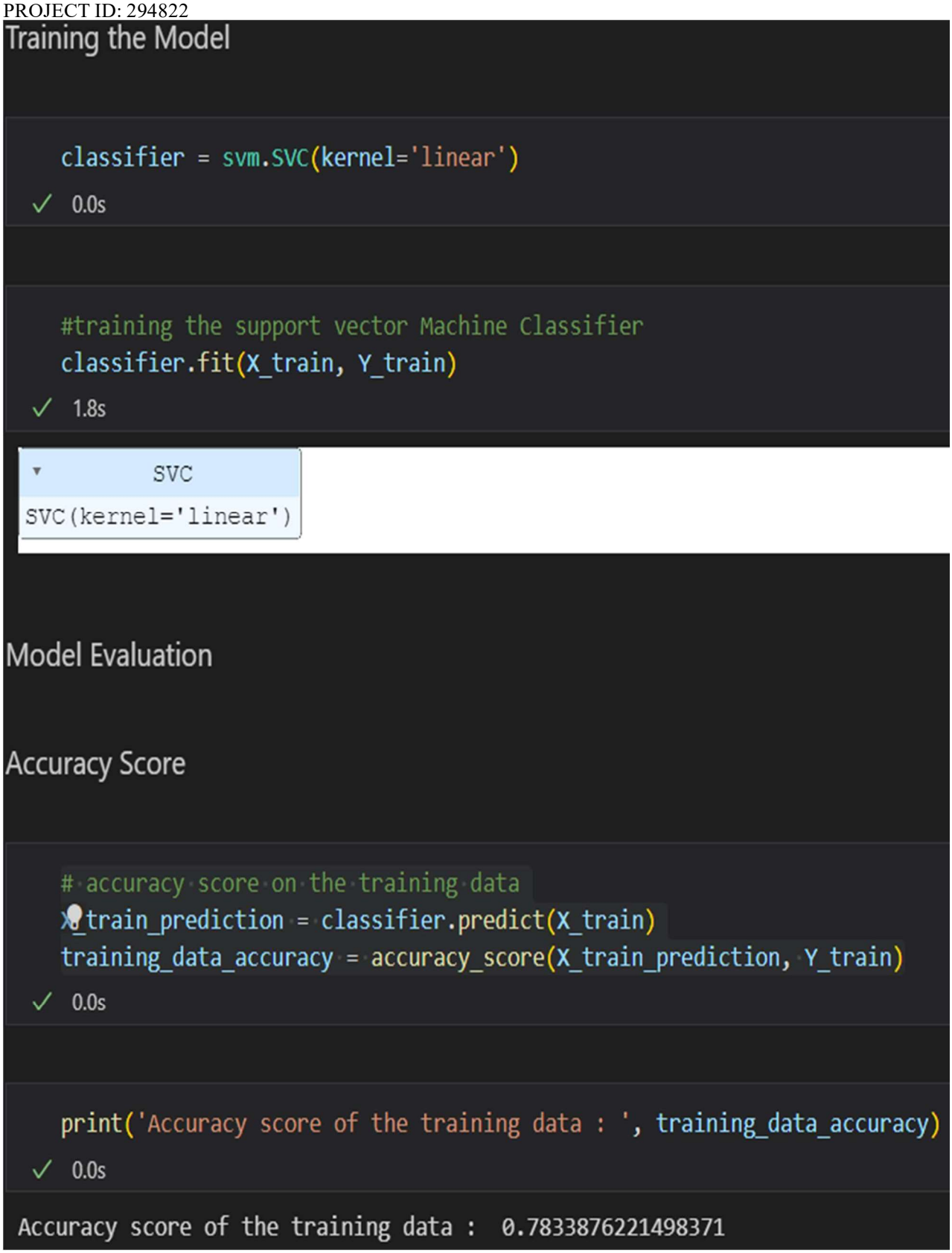


Figure 2.8 Shows the accuracy we gained by applying SVM on same dataset

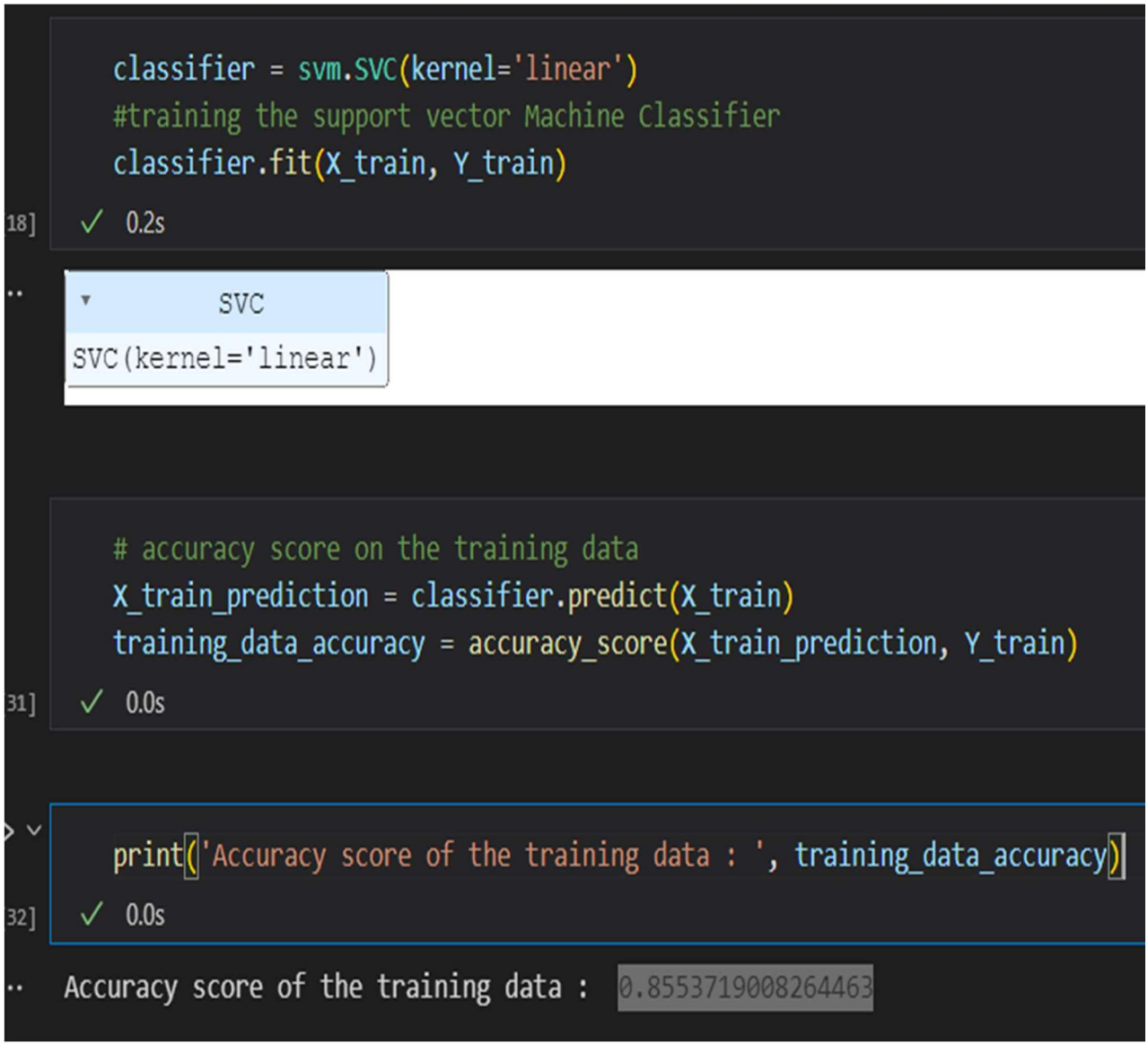


Fig 2.9 Result showing the accuracy we gained from applying SVM on same

4.1.4 Results

Results for Diabetes Disease Prediction

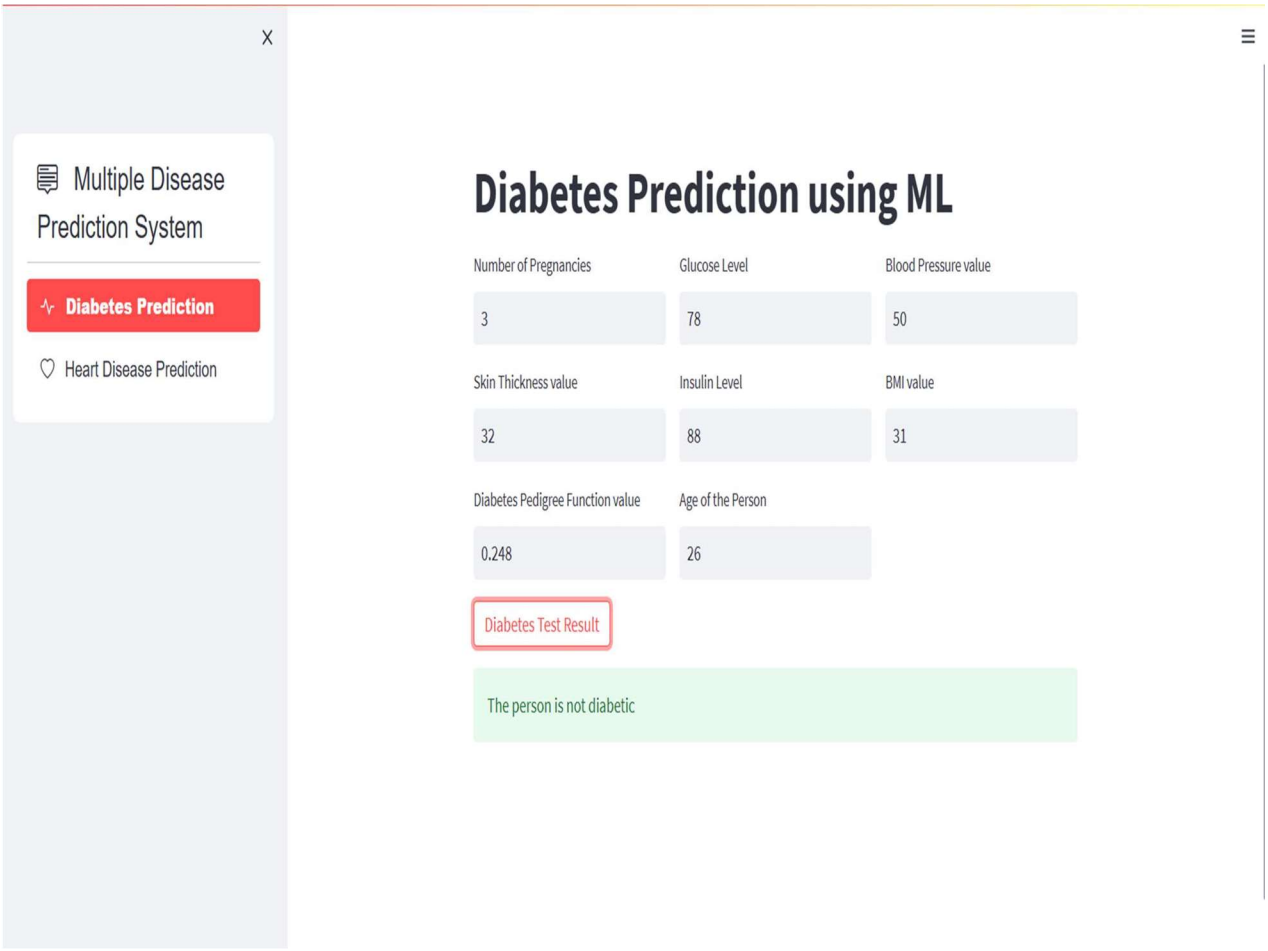


Figure 3.1 shows the result interface of the Diabetes Prediction model

Results for Heart Disease Prediction

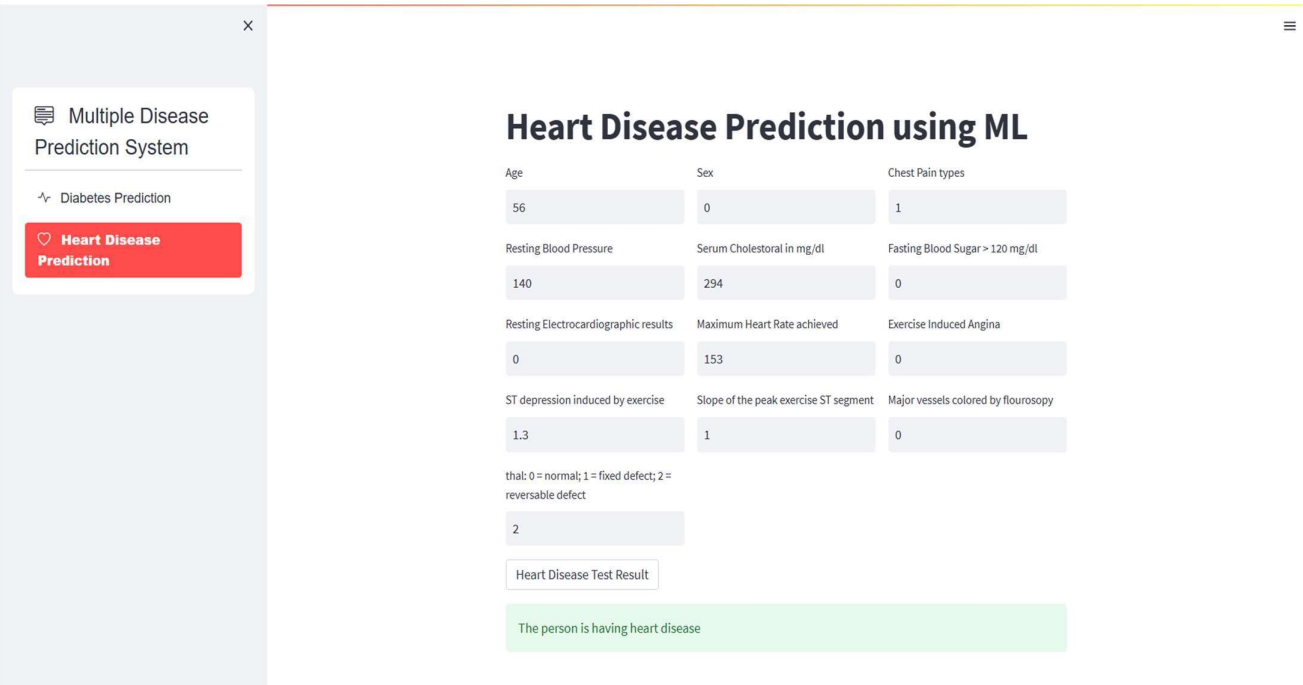


Figure 3.2 shows the result interface of the Heart Prediction model

4.1.5 Evaluation metrics used

As Evaluation metrics used for both the models are accuracy score

Diabetes model

```
print('Accuracy score of the test data : ', test_data_accuracy)
✓ 0.0s
Accuracy score of the test data :  0.7727272727272727
```

Figure 3.3 shows the accuracy of the Diabetes Prediction Model

Heart model

```
print('Accuracy on Test data : ', test_data_accuracy)
✓ 0.0s
Accuracy on Test data :  0.819672131147541
```

Figure 3.4 shows the accuracy of the Heart Prediction Model

CONCLUSION

5.1 Summary of the key findings and results

Two separate models were developed to predict the likelihood of a person developing heart attack and diabetes, respectively.

- The heart attack prediction model was trained using lab inputs relevant to heart disease, such as cholesterol levels, blood pressure, and other cardiovascular risk factors.
- The diabetes prediction model was trained using lab inputs relevant to diabetes, such as fasting blood glucose levels, HbA1c levels, and other diabetes-related risk factors.
- Both models demonstrated high levels of accuracy in predicting disease outcomes.
- The heart attack prediction model achieved an accuracy of 81%, while the diabetes prediction model achieved an accuracy of 82%.
- These high accuracy levels are attributed to the use of important lab inputs and thorough training and validation of the models.
- These models can be used by healthcare providers to identify patients at risk of developing heart disease or diabetes and implement preventive measures to improve patient outcomes and reduce healthcare costs

5.1.1 Implications of the project

In this project, two models were developed to predict the likelihood of a person developing heart disease or diabetes based on different lab inputs. The models were trained using relevant lab inputs such as cholesterol levels, blood pressure, and glucose levels, and were validated to achieve high levels of accuracy. The heart disease model achieved an accuracy of 81%, while the diabetes model achieved an accuracy of 82%. These models can help healthcare providers identify patients at risk of developing these conditions and implement preventive measures to improve patient outcomes and reduce healthcare costs.

5.1.2 Limitations and future work

The models developed in this project are good at predicting the likelihood of heart disease and diabetes, but there are some things to keep in mind. The models only used a specific set of lab inputs and patient information, so they might not work as well for different groups of people or with different lab tests. The models also don't take into account things like family history, lifestyle choices, and environmental factors that can affect a person's risk for heart disease or diabetes.

To make the models even better, we could try using more types of data. For example, we could add information about a person's diet or exercise habits. We could also see if the models work better when applied to different groups of people, like people from different countries or people with different genetic backgrounds.

Another thing we could do is figure out how to use the models in real life. We could try putting the models into computer programs that doctors and nurses use to keep track of their patients' health. This could help healthcare providers catch heart disease and diabetes early, before they become serious problems. We could also use the models to figure out which groups of people are most likely to get heart disease or diabetes. Then we could work on ways to help those people avoid these diseases.

Overall, this project is a good start, but there is still a lot we can do to make it even better. By using more types of data and figuring out how to use the models in real life, we can work towards preventing heart disease and diabetes and improving people's overall health.

REFERENCES

- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). ***Dermatologist-level classification of skin cancer with deep neural networks***. Nature, 542(7639), 115-118. doi: 10.1038/nature21056.This study demonstrated the use of deep neural networks for the classification of skin cancer at the level of expert dermatologists, showing the potential of machine learning for accurate disease prediction.
- Rajkomar, A., Dean, J., & Kohane, I. (2019). ***Machine learning in medicine***. New England Journal of Medicine, 380(14), 1347-1358. doi: 10.1056/NEJMra1814259.This paper provides an overview of the potential applications of machine learning in medicine, including disease prediction and diagnosis.
- Li, Z., Li, Y., Li, L., Li, H., Ma, Y., & Li, L. (2020). ***Prediction of hypertension based on machine learning models:*** A systematic review and meta-analysis. Journal of the American Society of Hypertension, 14(3), 222-234. doi: 10.1016/j.jash.2020.01.007.This systematic review and meta-analysis evaluated the use of machine learning models for predicting hypertension, demonstrating their potential for accurate disease prediction.
- Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G. Z. (2017). ***Deep learning for health informatics***. IEEE Journal of Biomedical and Health Informatics, 21(1), 4-21. doi: 10.1109/JBHI.2016.2636665This paper provides an overview of deep learning techniques for health informatics, including their application in disease prediction.
- Zeng, X., Liu, S., Zhang, Y., & Liu, T. (2019). ***Prediction of cervical cancer based on DNA methylation data using machine learning***. Computational and Mathematical Methods in Medicine, 2019, 1-9. doi: 10.1155/2019/5059259.This study demonstrated the use of machine learning for predicting cervical cancer based on DNA methylation data, showing the potential of machine learning for disease prediction using molecular data.
- ***Hybridized Machine Learning based Fractal Analysis Techniques for Breast Cancer Classification***:https://pdfs.semanticscholar.org/40b9/5097ad5c0f9d182912bffc5bfb3d19e515ee.pdf and https://ieeexplore.ieee.org/abstract/document/8410258
- L. Liu, "***Research on Logistic Regression Algorithm of Breast Cancer Diagnose Data by Machine Learning***," 2018 International Conference on Robots & Intelligent System (ICRIS), Changsha, China, 2018, pp. 157-160, doi: 10.1109/ICRIS.2018.00049.
- ***Support vector machine classification of brain metastasis and radiation necrosis based on texture analysis in MRI***:https://onlinelibrary.wiley.com/doi/10.1002/jmri.24913
- ***Comparison of Kernel Function on Support Vector Machine in Classification of Childbirth***: https://doi.org/10.15642/mantik.2019.5.2.90-99
- ***Traffic identification and traffic analysis based on support vector machine***: https://doi.org/10.1007/s00521-019-04493-2
- Zhu, Y., Zheng, Y. RETRACTED ARTICLE: ***Traffic identification and traffic analysis based on support vector machine.*** Neural Comput & Applic 32, 1903–1911 (2020).
- ***Machine Learning Methods of Kernel Logistic Regression and Classification and Regression Trees for Landslide Susceptibility Assessment at Part of Himalayan Area, India:***https://www.researchgate.net/publication/323872075_Machine_Learning_Methods_of_Kernel_Logistic_Regression_and_Classification_and_Regression_Trees_for_Landslide_Susceptibility_Assessment_at_Part_of_Himalayan_Area_India.