ABSTRACT

Data mining is an integrative field of study in the area of healthcare. Machine learning algorithms and data visualizations plays an important role in the field of healthcare. The project “Multiple Disease Prediction” mainly focuses on predicting different types of diseases by using a machine learning system. It predicts diseases like Diabetes, Heart disease, Parkinsons. The aim of the project is to predict multiple diseases based on the analysis of symptoms, medical history, and the patient’s lifestyle. The methodologies we are using for this project are: -Data collection, Data preprocessing, Model selection, Training and Testing, Model deployment. For metrics requirement algorithms are: - SVC: A Support Vector Classifier (SVC) is a supervised learning algorithm used for classification and regression tasks. It works by finding the optimal boundary (hyperplane) that best separates data into classes. SVMs are effective in high-dimensional spaces and are known for their ability to handle complex, non-linear data using kernel functions. KNN: The K-Nearest Neighbors (KNN) algorithm is a simple, supervised machine learning method used for classification and regression. It predicts the class or value of a data point by analyzing the 'k' closest labeled points in the feature space, making decisions based on the majority or average of these neighbors. Logistic Regression: We used logistic regression to predict the probability that a given input instance belongs to a particular class. It does this by modeling the relationship. between the input features and the probability of a binary outcome. Random Forest: We used for classification and regression tasks. It works by building multiple decision trees during training and combining their predictions to make more accurate and robust predictions than individual decision trees. We are using these algorithms over other algorithm because these algorithms provide a more comprehensive and clear evaluation of a model’s performance with great accuracy as compared to other metrices. By applying these algorithms, we will build a model. And for showing that model values we will create a dashboard where a person will enter their details and symptoms they are suffering from, to predict that they are suffering from any kind of chronic disease or not, and also to suggest for further treatment if needed. This project “Multiple Disease Prediction”, will act as a great blessing for the people who will usethis model for predicting their disease and keeping their health on a track and lead a healthy life.

TABLE OF CONTENTS

LIST OF FIGURES

ACRONYMS (if any)

KNN: K-Nearest neighbor
ML: Machine Learning
SVM: Support Vector Machine
ROC: Receiver Operating Characteristic
LR: Logistic Regression
RF: Random Forest

EQUATIONS

Chapter-1

INTRODUCTION

In today's world healthcare is the most important area to worry about. As in today's world people are much acquainted with Internet but they are not much aware about their physical health. People ignore their small problem and don't visit for regular checkup which at last causes a chronic disease. By taking the advantages of the growing usage of technology we can create a system which can spread self-awareness among people by predicting their diseases beforehand, and making them aware. The project multiple disease prediction can predict diseases like heart diseases, diabetes, Parkinson's disease. The main aim of the project is to use the machine learning algorithms to predict multiple diseases and reduce the risk of being affected by diseases. By identifying important patterns and detecting correlations and relationships among variables, and using various data mining tools and machine learning approaches has changed healthcare organizations which can help the society a lot. It actsas an important instrument in the medical sector, providing and comparing existing data for the future course of action. This technology combines multiple analytic methodologies with modern and complex algorithms, allowing for the exploration of huge amounts of data. It is used in healthcare sector to analyses, record, gather data in a systematic order. Use of machine learning inthe field of health sector not only provide us an easy way to care about ourself but also it makes usunderstand about the importance of our physical health.

In the diagnosis process of a person's disease one or more process is performed. And also, many assumptions are done to get the right disease to be predicted. Firstly, our machine learning model will see various factors of each patient like age, gender, past medical history, by this thing it will do personalized prediction / reading/ diagnosis of disease patterns and will say the unique customized treatment/cure/right treatment to it for ultimate good outcome, it's like saving doctor and patient time and resources/expenses as compare to the traditional way that seems to be far better.

In area of ethical and privacy ensuring that patient data is kept confidential and secure, and only used for the intended purpose of disease prediction. Creating awareness of the benefits of disease prediction technology and describing the accessibility to all individuals, regardless of their economic status, geographic location, or other factors that may affect access to healthcare. Ethical concerns such as patient privacy, data security, and algorithms of that machine learning is used because it is responsibly and good ethically in healthcare cases.

Chapter-2

LITERATURE REVIEW

2.1. Diabetes Disease:

The Diabetes Disease dataset, often used for machine learning and data analysis projects, is commonly available through repositories like Kaggle or the UCI Machine Learning Repository. It typically contains information on patients' medical records and attributes that may influence diabetes outcomes. These datasets aim to help build predictive models to diagnose diabetes or identify risk factors associated with the disease.

2.1.1. Key features in a diabetes dataset often include:

- **Pregnancies**: Number of times the patient has been pregnant.
- **Glucose**: Plasma glucose concentration after a glucose tolerance test.
- **Blood Pressure**: Diastolic blood pressure level.
- **Skin Thickness**: Triceps skinfold thickness, which can indicate body fat.
- **Insulin**: 2-hour serum insulin level.
- **BMI (Body Mass Index)**: A measure of body fat calculated from height and weight.
- **Diabetes Pedigree Function**: A function that assesses the likelihood of diabetes based on family history.
- **Age**: Age of the patient

The target variable in these datasets is usually a binary indicator of whether a patient has diabetes (1) or not (0).

2.1.2. Applications and Analyses

The Diabetes Disease dataset is valuable for:

- **Predictive Modelling**: Building models like Logistic Regression, KNN, Decision Trees, and Random Forests to predict diabetes.
- **Feature Analysis**: Identifying key risk factors or significant variables that influence the likelihood of diabetes.
- **Visualizations**: Using box plots, scatter plots, and correlation matrices to understand relationships between features.

2.2. Herat Disease:

The Heart Disease dataset is widely used in machine learning and medical research to predict the presence or absence of heart disease in individuals. This dataset, often sourced from databases like the UCI Machine Learning Repository, includes patient information and medical measurements that can help identify patterns associated with heart disease. By analysing this data, researchers and practitioners aim to build models for early diagnosis and to gain insights into the risk factors for heart disease.

2.2.1 Key Features

The dataset typically includes the following attributes:

- **Age**: Age of the patient.
- **Sex**: Gender of the patient, often coded as 1 for male and 0 for female.
- **Chest Pain Type (cp)**: Type of chest pain experienced, often with categories indicating severity and type (e.g., angina).
- **Resting Blood Pressure (trestbps)**: Blood pressure at rest, an important indicator of cardiovascular health.
- **Cholesterol (chol)**: Serum cholesterol in mg/dl, a common risk factor for heart disease.
- **Fasting Blood Sugar (fbs)**: Fasting blood sugar level, often recorded as >120 mg/dl or below.
- **Resting ECG (restecg)**: Electrocardiographic results, which can indicate abnormalities in heart function.
- **Max Heart Rate (thalach)**: Maximum heart rate achieved during physical activity.
- **Exercise-Induced Angina (exang)**: Indicates whether angina is induced by exercise (1) or not (0).
- **ST Depression (oldpeak)**: Exercise-induced ST segment depression, relevant for assessing heart disease severity.
- **Slope**: Slope of the peak exercise ST segment.
- **Number of Major Vessels (ca)**: Number of major vessels coloured by fluoroscopy, showing blood flow.
- **Thalassemia (thal)**: Blood disorder related to haemoglobin, sometimes used as a categorical attribute.

2.2.2 Applications and Analyses

The Heart Disease dataset is useful for:

- **Predictive Modelling**: Building classification models like Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines to predict heart disease.
- **Feature Importance Analysis**: Understanding which factors, like cholesterol or chest pain type, are most influential in heart disease prediction.
- **Data Visualization**: Scatter plots, correlation matrices, and histograms are used to observe

relationships between features and visualize risk trends.

2.3.Parkinson’sDisease:

Parkinson’s Disease (PD) datasets are essential for researching the progression, diagnosis, and management of Parkinson's disease—a neurodegenerative disorder that primarily affects movement. These datasets often include voice measurements, movement tracking, and clinical assessments of motor functions. The primary goal of analysing Parkinson’s datasets is to predict the presence of the disease or to assess its progression.

2.3.1. Key Features

Parkinson’s datasets often contain data from audio recordings, motor assessments, and other biomarkers, with features including:

MDVP (Maximum Vocal Fold Deviation) Measurements:

MDVP

- **(Hz)**: Average vocal fundamental frequency.
- **MDVP**
- **(Hz)** and **MDVP**
- **(Hz)**: Maximum and minimum vocal frequencies.

Jitter and Shimmer Measurements: Indicators of vocal stability, often analysed for tremors in voice:

1. **Jitter (Abs, %)**: Measures variations in frequency.
2. **Shimmer (dB)**: Measures variations in amplitude.

HNR (Harmonics-to-Noise Ratio): Reflects the ratio of harmonic sound to noise, used to assess voice quality.

RPDE (Recurrence Period Density Entropy) and **DFA (Detrended Fluctuation Analysis)**: Nonlinear features indicating signal complexity, which can be altered in Parkinson’s patients.

PPE (Pitch Period Entropy): Measures irregularities in voice pitch.

UPDRS (Unified Parkinson’s Disease Rating Scale): A clinical measure of Parkinson’s severity, including motor and non-motor symptoms.

The target variable often represents whether the individual has Parkinson’s Disease (1) or not (0), though some datasets include progression levels or specific symptom severity ratings.

2.3.2. Applications and Analyses

Parkinson’s Disease datasets support various research and diagnostic objectives, such as:

Predictive Modelling: Classifiers like Support Vector Machines (SVM), Random Forests, and

Neural Networks are commonly used to predict disease presence.

Feature Analysis: Identifying which features—such as vocal irregularities—strongly correlate with Parkinson’s, aiding in early detection.

Progression Tracking: Analysing how certain features change over time to predict disease progression.

et al., [1] Arumugam, K., in simple terms, we're using machine learning to create a system that can detect multiple diseases at once, like heart problems, diabetes, and cancer. Mostly the current systems that are being used focus on one disease at a time and aren't always accurate whereas this proposed system looks at various factors for each disease, like symptoms, test results and all such disease-specific parameters that are coupled together to give reliable predictions. This means people can keep track of their health better and take action early if needed, which could help them live longer and healthier lives. This not only helps in early diagnosis but accurate and efficient online predictions making it easy for medical sciences.

et al., [2] Yaganteeswarudu, Chronic diseases, which are long-lasting and life-threatening health conditions, are on the rise worldwide. Detecting these diseases early is crucial for improving patients' chances of survival. Researchers have been exploring different ways to predict these diseases before they become severe. The proposed system uses advanced artificial intelligence called artificial neural networks (ANN) combined with a technique called particle swarm optimization (PSO) to predict five common chronic diseases: breast cancer, diabetes, heartdiseases, lung cancer and Parkinson's disease. It was with seven other classification algorithms tosee how well it worked. The ANN model, integrated with PSO, performed better than the others and achieved an impressive accuracy rate of 99.67%. However, the accuracy of this model dependson the specific characteristics of the data we used. We also compared our results with those of other studies and found that our approach outperformed them. Additionally, our optimized ANN processing was faster than other methods like random forest, deep learning, and support vector machines. Our research has the potential to improve the early diagnosis of chronic diseases in hospitals and could even lead to the development of online diagnostic systems.

et al., [3] Uddin, S., Predicting chronic diseases for individuals is important, but past studies haven't fully looked at how different diseases relate to each other. To understand the risks better, we created a model called Multitask Learning Cox (MTL-Cox). It works by looking at a bunch of factors that might affect your health and figuring out how they relate to each disease. It uses a special method to learn from multiple diseases at once, considering how they affect each person differently. We tested our model on two datasets and found it performed better than other methods,improving prediction accuracy by up to 12%. This means we can better identify people at risk for nine different chronic diseases. Our work could help with early detection and personalized treatment plans for these conditions.

et al., [4] Uddin, S., In this computer world, huge data are generated in several fields. Statistics in the healthcare engineering provides data about many diseases and corresponding patient's information. These data help to evaluate a huge amount of data for identifying the unknown patterns in the diseases and are also utilized for predicting the disease. Hence, this work is to plan and implement a new computer-aided technique named modified Ensemble Learning with Weighted RBM Features (EL-WRBM). Data collection is an initial process, in which the data of various diseases are gathered from UCI repository and Kaggle. Then, the gathered data are pre-processed by missing data filling technique. Then, the pre-processed data are performed by deep belief network (DBN), in which the weighted features are extracted from the RBM regions. Then, the prediction is made by ensemble learning with classifiers, namely, support vector machine (SVM), recurrent neural network (RNN), and deep neural network (DNN), in which hyper-parameters are optimized by the adaptive spreading rate-based coronavirus herd immunity optimizer (ASR-CHIO). At the end, the simulation analysis reveals that the suggested model has implications to support doctor diagnoses.

et al., [5] Khadir, M. A., The primary focus of recent studies in healthcare has been on developing specific prediction models for individual illnesses, which is takes a long time and could be dangerous for people with more than one health problem. In the context of heart disease prediction, they conducted a study employing K-nearest neighbor give us good accuracy which is not an efficient score and the accuracy could have been increased if other classifiers would have been used as Support Vector machine (SVM). Diabetes is another disease that has been studied by researcher using machine learning algorithm. They employed logistic regression and K-nearest neighbor algorithm. The result show that logistic regression performs better than the K-nearest neighbor algorithm in term of accuracy.

et al., [6] Talasila, B., The study built a framework to predict multiple disease using patient data. They analyzed data from 4920 patients looking at 41 different diseases. To avoid overfitting, out of total 132 potential symptoms or side effects the researchers only selected 95 for inclusion. In this work Decision Tree, LightGBM gives the best accuracy which is 97.3% and 97.3%. this classifier gives us good accuracy but in order to increase the accuracy they employed Random

Forest which accuracy is 98.3%.

et al., [7] Singh, A., This research paper explores the application of machine learning in the detection of various diseases like diabetes and heart disease. For diabetes dataset they used mainly 4 main algorithms Decision Tree, Naïve Bayes and SVM algorithms and compared their accuracy to get a better result which is 85%, 77%, 77.3% respectively. They also employed an ANN to evaluate the classification result whether the disease is classified properly or not. Here they compared the precision, recall, F1 score, support and overall accuracy.

The main goal of the paper is to talk about how important the heart is for living things. Diagnosing ang predicting heart disease must be perfect and correct because it is very crucial which can lead to death. In this paper they calculate the accuracy of machine learning for predicting heart disease using K-nearest neighbor, decision tree, linear regression and SVM for training and testing the dataset. They compared the algorithm and their accuracy SVM 83%, Decision tree 79%, Linear regression 78%, K-nearest neighbor 87%.

et al., [8] Pulicherla, P., The study claims that diabetes is one of the most dangerous diseases in the world. Although it's easy and flexible to tell if someone is sick or not, the researcher used machine learning method to identify the diabetic condition. Here they compare the three key algorithm Decision tree, Naïve Bayes and SVM for their accuracy. They achieving the accuracy of SVM 77.3%, Naïve Bayes 77% and Decision Tree 85%. Additionally, they employed KNN method during the training phase to assess the network's response and disease categorization.

As the heart plays a important role for living things, the main aim of the paper is to accurately diagnose and predict heart related disease. To prevent heart disease, they used machine learning and artificial intelligence, which are helpful for forecasting natural events. They calculate the accuracy of machine learning for predicting heart disease using K-nearest neighbor, Decision Tree, Linear Regression and SVM. The result showed that SVM had 83% accuracy, Decision Tree had 79%, K-nearest neighbor had 87% and Linear Regression had 78%.

et al., [9] Vasavi, D., This article discusses the application of machine learning techniques and algorithm in predicting multiple diseases which include breast cancer, diabetes, heart diseases, and Parkinson's diseases. The authors highlight the importance of predictive analysis in health care for timely diagnosis and treatment, especially in the areas with limited medical infrastructure [9]. This article outlines the methodology of using machine learning algorithms like SVM and Logistic regression for disease prediction. And this article discusses the use of Streamlit, a Python based framework, for building and sharing machine learning and data science web applications instantly

[9]. The results and analysis section of the article presents the accuracy scores of the disease prediction models. In conclusion this paper it presents us the importance of utilizing machine learning techniques fordisease prediction and prevention.

et al., [10] Bayati, M., This paper discusses the importance of early identification for chronic disease to improve the lifestyle of people and reduce healthcare costs. It also highlights the limitations of current risk assessment methods used by employer's wellness programs and propose an approach to select a low-cost set of biomarkers for more accurate prediction of Multiple Disease. The proposed solution involves multi-task learning and reduction techniques from machinelearning and statistics. Empirical validation of the method is provided using data from electronic medical records systems and performing its performance against a statistical benchmark.

et al., [11] Bayati, M., The study proposes two approaches that is MTL and OLR-M models, for developing a clinical model to predict multiple diseases base on Health Risk Assessment (HRA) using some statistical methods and patient's data. While the results are very promising but the study acknowledges limitations such as the inability to compare prediction accuracy against common scoring. The investigation concludes that both MTL and OLR-M models offer a solution for designing a lower-cost HRA with comparable accuracy. This indicates that models maintain accuracy across different time horizon for positive diagnosis and under various performance metrics.

et al., [12] Dubey, A. K., This paper presents the prediction of multiple diseases using deep learning techniques. Most of the medical datasets including those from Kaggle and UCI repository were collected for diseases such as heart disease, diabetes, breast cancer, lung cancer, Parkinson's disease. In this the authors utilized a hybrid algorithm called L-BOA for optimal feature selection from the attribute sets of the collected datasets. First order statistical features were computed and merged with the optimally extracted features. These features were then applied to both neural network and deep brief learning.

et al., [13] Beg, A. A., The Multiple Disease Prediction System is an end-to-end machine learning solution designed to operate various medical conditions and predict patient's probabilities of suffering from illness. Traditional medical analysis system focuses on individual diseases, which results in fragmented approaches and cost increased. This system aims to make a revolutionize disease prediction by enabling concurrent prediction of multiple disease on a single platform. Utilizing machine learning algorithms and stream lit framework, it provides an efficient and accurate predictions based on the parameters. The project implements various algorithms such as

Logistic Regression, Support Vector Machine, and K-Nearest Neighbours for different disease prediction. Deployment via the Stream lit Cloud Server offers a user-friendly web interface.

et al., [14] Unnithan, D. R., In this paper they provide a brief description of telemedicine, its definition, benefits, and the integration of datamining techniques within telemedicine systems. Telemedicine refers to the provision of healthcare services remotely, and using of telecommunication technology. It provides healthcare professionals to diagnose, treat, and manage patients' health conditions without the need for one-to-one visits. Telecommunication provides various methods of communication, including video calls, phone calls, secure messaging, and remote monitoring devices.

et al., [15] Olajide, A. O., In this paper they have discuss about disease prediction using machine learning algorithms, mainly focusing on the challenges of diagnosing diseases with variety of symptoms and the prevalence of multimorbidity. Particularly in the context of multimorbidity, and the role of machine learning algorithms in enhancing diagnostic capabilities. It underscores the need for rigorous comparative analysis to identify the most effective classifier for disease prediction task.

et al., [16] Prakaash, A. S., In this research paper they mainly discuss the importance of disease prediction in healthcare, particularly focusing on multi-disease prediction using machine learning algorithms. Here, they have discussed mainly about the difference between single disease prediction system and multiple disease prediction system, uses of machine learning in disease prediction, what kind disease we can predict, etc.

et al., [17] Ramani, R., The dataset is from the Pima Indian community, consisting of 768 instances for training and 8 features like pregnancy count, blood pressure, and glucose levels. The class variable indicates diabetes presence (1) or absence (0). To improve accuracy, nontraditional data is normalized using min-max normalization. A MapReduce-based framework with a modified Artificial Neural Network (ANN) classifier is employed for predicting diabetic disease. This approach achieves high precision, accuracy, and recall compared to existing methods. Numerous studies explore big data analytics in healthcare for disease prediction and management. Backpropagation is utilized for error correction and weight updates in the neural network. The focus lies on early disease detection and patient care through advanced data analysis techniques.

et al., [18] Zhang, S., A survival analysis model is designed with a hazard function expressed as \(h(t; x; \beta; \lambda) = \lambda t^{\lambda-1} \). This model is tailored for personalized prediction of nine chronic diseases. It leverages a multitask learning framework to improve its performance. Evaluation is conducted using metrics such as the concordance index, AUC, specificity, sensitivity, and Youden index. Feature selection is enhanced using L1 and L2,1 norm regularization. Right-censoring is tackled through the Cox proportional hazards model. Furthermore, correlations among multiple diseases are considered for devising prevention strategies.

et al., [19] Rashid, J., This method shares parameters to improve stability and avoid strict task limitations. It uses soft parameter-sharing in a multitask model to predict nine chronic diseases personally. By employing a multitask approach, it trains models for multiple diseases simultaneously. This approach outperforms other methods in evaluating predictive performance. Evaluation metrics include concordance index, AUC, specificity, sensitivity, and Youden index. It deals with right-censoring by using the Cox proportional hazards model.

et al., [20] Buragadda, S., This discusses chronic diseases like Breast Cancer, Diabetes, Heart, Hepatitis, and kidney disease. It suggests an artificial neural network with Particle Swarm Optimization for feature selection. Compared to other methods like Decision Tree, Random Forest, Deep Learning, Naive Bayes, SVM, K-NN, and Logistic Regression, this approach performs better with 91.61% accuracy. The results indicate superior accuracy with artificial neural networks compared to other techniques. The proposed model utilizing artificial neural networks and Particle Swarm Optimization excels in predicting chronic diseases. Comparative analysis indicates that K-NN achieves higher accuracy. The study focuses on feature selection, model architecture, and performance evaluation for disease prediction.

Chapter-3

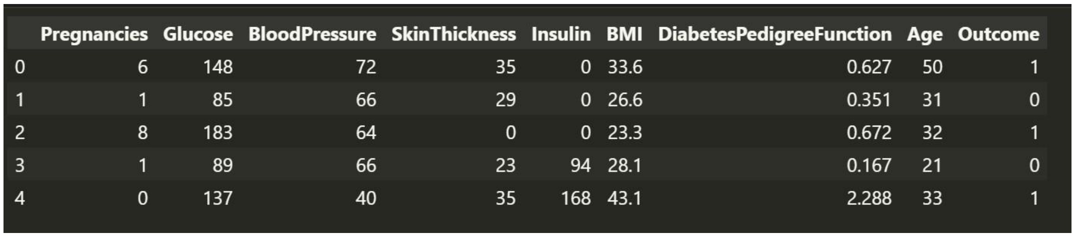
DATA PREPROCESSING

Data Pre-processing is a process of preparing the raw data and making a suitable machine learning model. It is a most important step while creating a machine learning model. Here we take the multiple disease dataset, in this dataset we have 3 different datasets i.e. Heart disease, Diabetes, and Parkinson’s.

ROC Curve: Plots TPR vs. FPR to evaluate classifier performance. TPR and FPR: Measures of correctly/incorrectly predicted positives. AUC: Indicates overall model performance; 1 is perfect,0.5 is random guessing.

3.1 Diabetes Disease: -

3.1.1 Dataset



	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

(Fig 3.01: Dataset of Diabetes)

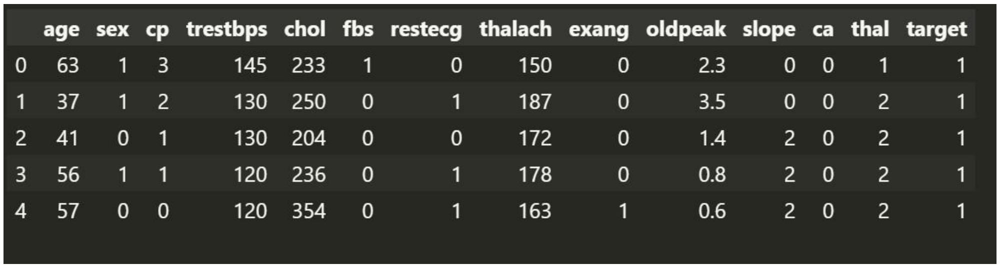
Diabetes, is a group of metabolic disorders in which there are high blood sugar levels over a prolonged period. Symptoms of high blood sugar include frequent urination, increased thirst, and increased hunger. If left untreated, diabetes can cause many complications. Acute complications can include diabetic ketoacidosis, hyperosmolar hyperglycemic state, or death. Serious long-term complications include cardiovascular disease, stroke, chronic kidney disease, foot ulcers, and damage to the eyes.

I have collected this dataset from Kaggle. The datasets consist of several medical predictor variables and one target variable, Outcome. Predictor variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

The dataset which we are using is consist of 769 rows and 9 columns. The 9 columns are representing all kinds of required things for finding a person is having diabetes or not. Those columns are Pregnancies, Glucose, Blood Pressure, Skin Thickness, BMI, Diabetes Pedigree Function, Age and Outcome.

3.2. Heart Disease

3.2.1. Dataset



	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

(Fig 3.02: Dataset of Heart Disease)

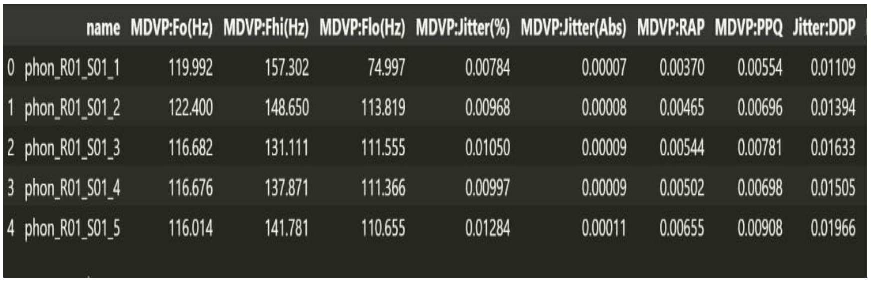
Heart disease encompasses various conditions affecting the heart, such as coronary artery disease, arrhythmias, and heart defects. Coronary artery disease, the most common type, involves the narrowing of arteries due to plaque buildup, restricting blood flow to the heart. Major risk factors include high blood pressure, high cholesterol, diabetes, smoking, obesity, and inactivity, along with genetic predispositions. Symptoms like chest pain, shortness of breath, and fatigue often indicate underlying heart issues.

Early detection and treatment are essential for managing heart disease. Lifestyle changes, including a heart-healthy diet, regular exercise, quitting smoking, and stress management, can significantly reduce the risk. In severe cases, medications or surgical interventions, like stents or bypass surgery, may be required. Regular check-ups and screenings play a crucial role in early diagnosis, helping prevent complications and improving long-term outcomes.

The dataset which we are using for this is consist of 304 rows and 14 columns. These 14 columns represent all the required data which are used for finding a person is having heart disease or not. Those columns are Age, Sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal and target.

3.3. Parkinson's Disease:

3.3.1. Dataset



	name	MDVP:Fo(Hz)	MDVP:Fhi(Hz)	MDVP:Flo(Hz)	MDVP:Jitter(%)	MDVP:Jitter(Abs)	MDVP:RAP	MDVP:PPQ	Jitter:DDP
0	phon_R01_S01_1	119.992	157.302	74.997	0.00784	0.00007	0.00370	0.00554	0.01109
1	phon_R01_S01_2	122.400	148.650	113.819	0.00968	0.00008	0.00465	0.00696	0.01394
2	phon_R01_S01_3	116.682	131.111	111.555	0.01050	0.00009	0.00544	0.00781	0.01633
3	phon_R01_S01_4	116.676	137.871	111.366	0.00997	0.00009	0.00502	0.00698	0.01505
4	phon_R01_S01_5	116.014	141.781	110.655	0.01284	0.00011	0.00655	0.00908	0.01966

(Fig-3.03: Parkinson's Dataset)

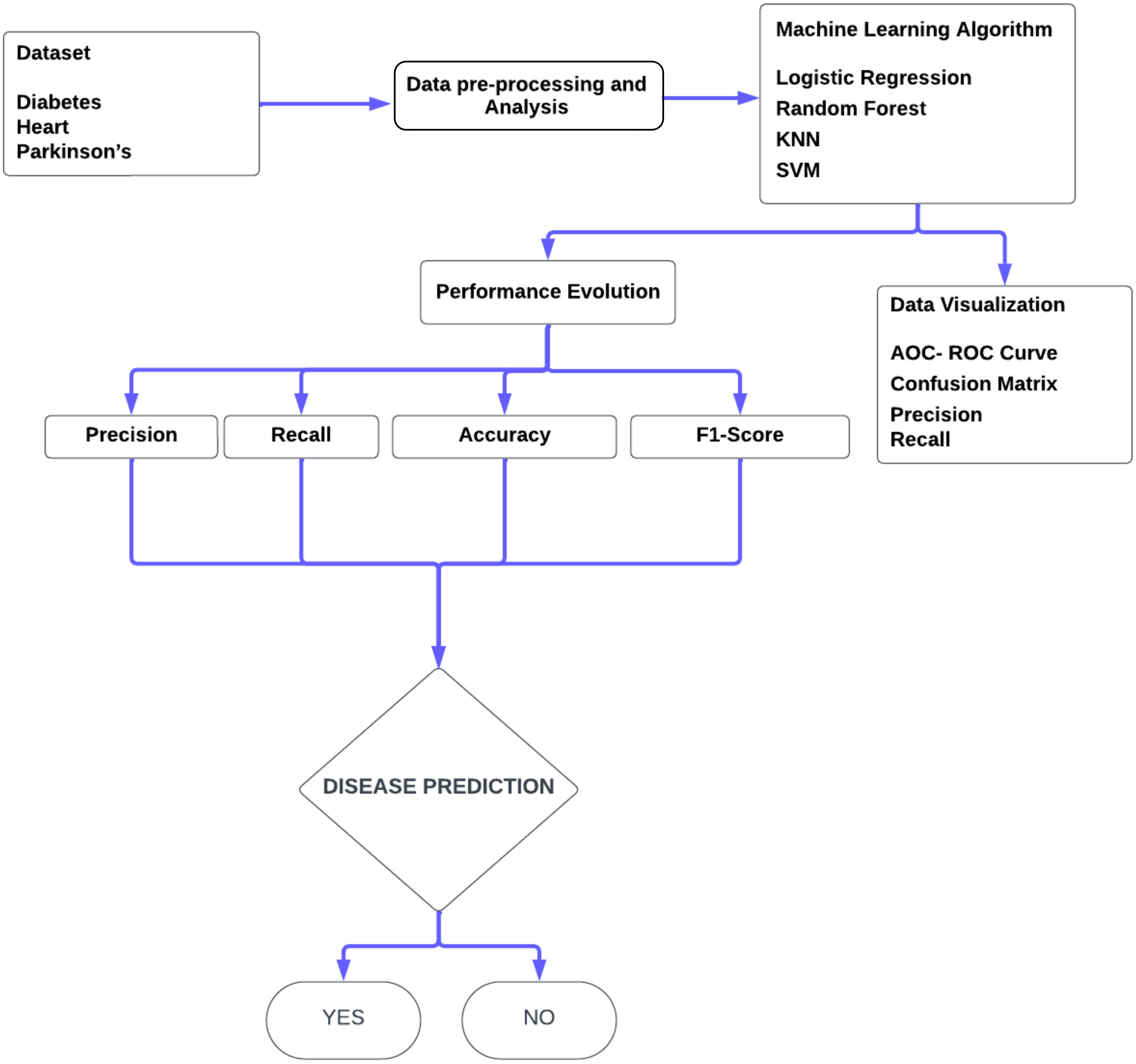
Parkinson's disease is a progressive neurological disorder that primarily affects movement control. It occurs when nerve cells in the brain, particularly those producing dopamine, gradually deteriorate. As dopamine levels decrease, symptoms such as tremors, stiffness, slowed movements, and balance difficulties emerge. The exact cause of Parkinson's remains unknown, but a combination of genetic and environmental factors is believed to contribute to its onset, often appearing in people over the age of 60.

Though there is no cure for Parkinson’s disease, early diagnosis can help manage symptoms and improve quality of life. Treatments may include medications to boost or mimic dopamine, as well as physical therapy to maintain mobility and muscle strength. In advanced cases, surgical options like deep brain stimulation can offer relief. A balanced diet, regular exercise, and support from healthcare professionals and caregivers are vital in managing the condition and slowing its progression.

The dataset which we are using for this is consist of 196 rows and 24 columns. These 14 columns represent all the required data which are used for finding a person is having heart disease or not. Those columns are name, 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, status, RPDE, DFA, spread1, spread2, D2 and PPE.

Chapter-4

PROPOSED MODEL



(Fig 4.01: Flow Diagram)

The following methodology outlines the steps for building and evaluating the proposed model:

4.1 . Data Preprocessing:

- **Cleaning**: Handle any missing values, outliers, or data inconsistencies identified during data visualization.
- **Feature Scaling**: Standardize or normalize features to ensure uniformity, especially when working with distance-based models.
- **Splitting Data**: Split the dataset into training and testing sets, typically in an 80-20 or 70-30 ratio, to assess model generalization.

4.2. Model Selection and Training:

- Train multiple algorithms such as Logistic Regression, Random Forest, K-Nearest Neighbours (KNN), and Support Vector Machine (SVM) to evaluate performance on the selected metrics.
- Use cross-validation to assess the model's consistency across different subsets of the data, minimizing overfitting.

4.3.Model Evaluation:

- After training, use the testing dataset to evaluate the model's performance based on the confusion matrix, precision, recall, accuracy, and F1-score.
- Plot the AUC-ROC curve and Precision-Recall curve for each model to visually compare performance across different algorithms.

4.4 Model Selection:

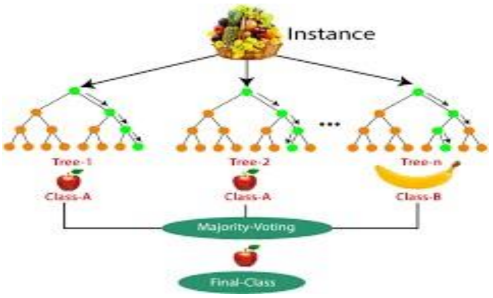
- Choose the model with the best performance on key metrics, prioritizing precision, recall, and F1-score, especially if data imbalance is present.
- If necessary, use ensemble methods to improve predictive power, combining multiple models to achieve more robust results.

Random Forest: -

Random Forest is a supervised learning technique. It can be used for both Classification and Regression problem in ML. It is based on the concept of ensemble learning which is a process of combining multiple classifiers to one classifier to solve a complex problem to improve the performance of the model.

Random Forest is a classifier that contain many decision trees on various subsets of the given dataset and takes the average to improve the prediction accuracy of the dataset. It is not relying on one decision tree, it take the prediction from each tree and based on the majority vote of prediction and it predict the final output.

Example: -

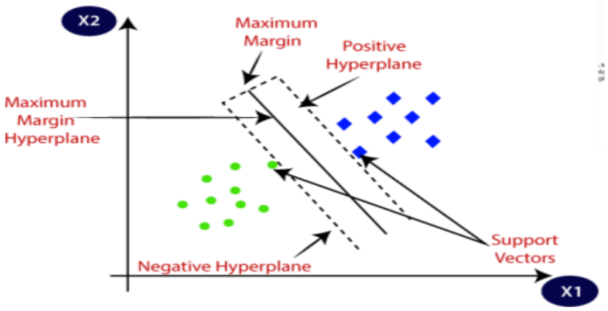


(Fig-4.02: Example for Random Forest)

Support Vector Machine: -

Support vector machine (SVM) is a powerful machine learning algorithm used for linear or nonlinear classification, regression, and even outlier detection. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors.

Example: -



(Fig-4.03: Example for SVM)

Logistic Regression: -

Logistic Regression is a supervised machine learning. It is used for predict the probability that an instance belongs to a given class or not. Logistic Regression is a statistical algorithm which analyse the relationship between 2 data factors. Logistic Regression is used for binary classification, that takes input as independent variable and produces a probability value between 0 and 1.

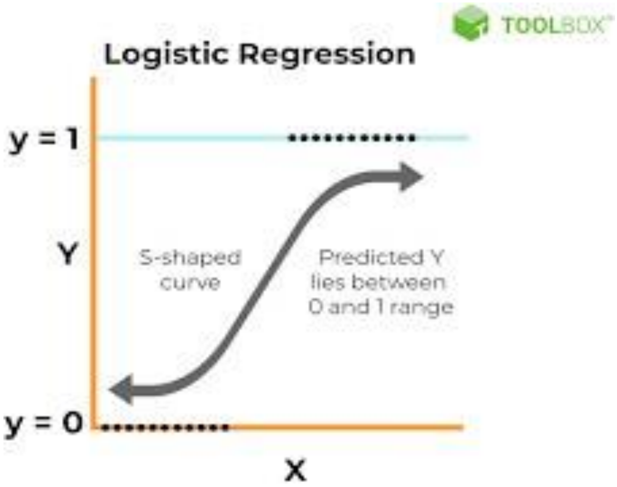
Logistic Regression predicts the output of a categorical dependent variable. Therefore, the output must be categorical or discreate value. It can be either Yes or No, 0 or 1, True or False.

Formula: -

$$f(x) = \frac{1}{1 + e^{-x}}$$

(Equation: 4.01)

Where, e = Base of natural logarithms
Example: -



(Fig-4.0: Example for Logistic Regression)

K-Nearest Neighbor: -

K-Nearest Neighbors (KNN) is a simple, non-parametric, and lazy learning algorithm used in classification and regression problems. The main idea behind KNN is to predict the category or value of a data point based on the classes or values of its closest neighbors in the feature space. This algorithm does not make any assumptions about the underlying data distribution, making it very versatile for various types of data.

For two points $p = (p_1, p_2, \ldots, p_n)$ and $q = (q_1, q_2, \ldots, q_n)$, the Euclidean distance $d(p, q)$ is calculated as:

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

(Equation: 4.02)

Graphical Analysis:

4.5. AUC-ROC Curve:

AUC-ROC (Area Under the Receiver Operating Characteristic curve) algorithm is used to:

Measure the ability of the model to distinguish between patients who have disease and those who don’t, plot the true positive rate against the false positive rate at various levels. It provides a single value AUC score that will represents the performance of the model, where higher scores indicate better performance rate researchers to assess the effectiveness of the model in identifying disease cases.

4.6. Confusion Matrix:

The Confusion Matrix is a table that summarizes the performance of a classification algorithm by counting the correct and incorrect predictions. It typically has four components:

True Positive (TP): Correctly predicted positive cases.

True Negative (TN): Correctly predicted negative cases.

False Positive (FP): Incorrectly predicted positive cases (Type I error).

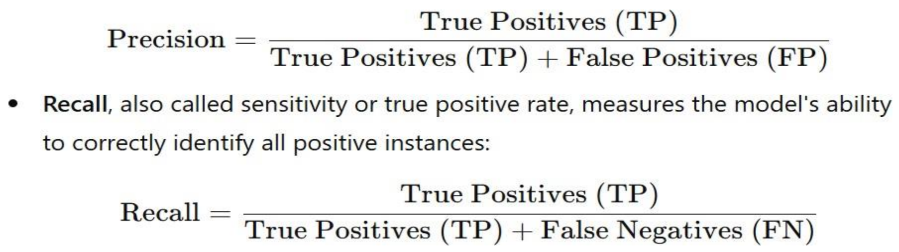
False Negative (FN): Incorrectly predicted negative cases (Type II error).

The Confusion Matrix enables calculation of key metrics, such as accuracy, precision, recall, and F1-score, helping to diagnose a model’s performance comprehensively.

4.7. Precision and Recall:

Precision measures the accuracy of positive predictions, defined as the proportion of true positive predictions out of all positive predictions made by the model

Recall, also called sensitivity or true positive rate, measures the model's ability to correctly identify all positive instances



$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

- **Recall**, also called sensitivity or true positive rate, measures the model's ability to correctly identify all positive instances:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

(Equation: 4.03)

Chapter-5

RESULTS AND DISCUSSION/ANALYSIS

5.1. Diabetes-

5.1.2. Accuracy:

Accuracy is a metric that measures how often a machine learning model correctly predicts the outcome. We got the accuracy for different models applied in our machine learning article is followed by: Logistic Regression = 0.7662, Random Forest = 0.7207, SVM = 0.7833, KNN = 0.7272.

5.1.3. Precision:

Precision is a metric that measures how often a machine learning model correctly predicts the positive class. We got the precision for different models applied in our machine learning article is followed by: Logistic Regression = 0.81, Random Forest = 0.79, SVM = 0.7727, KNN = 0.81.

5.1.4. Recall:

Recall is a metric that measures how often a machine learning model correctly identifies positive instances (true positives) from all the actual positive samples in the dataset. We got the precision for different models applied in our machine learning article is followed by: Logistic Regression = 0.84, Random Forest = 0.78, SVM = 0.6545, KNN = 0.81.

5.1.5. F1 Score:

The F1 score or F-measure is described as the harmonic mean of the precision and recall of a classification model. The two metrics contribute equally to the score, ensuring that the F1 metric correctly indicates the reliability of a model. We got the F1 Score for different models applied in our machine learning article is followed by: Logistic Regression = 0.84, Random Forest = 0.78, SVM = 0.6605, KNN = 0.81.

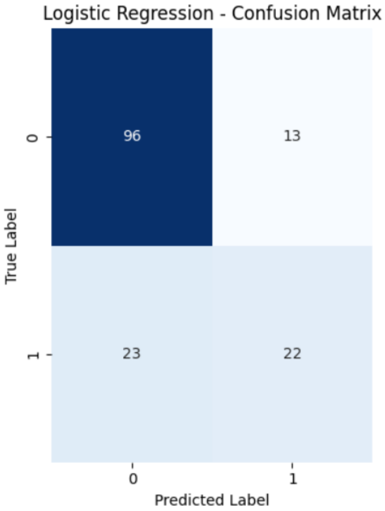
5.1.6. AUC-ROC Score:

An ROC curve, or receiver operating characteristic curve, is like a graph that shows how well a classification model performs. It helps us see how the model makes decisions at different levels of certainty. 24 We got the AUC-ROC Score for different models applied in our machine learning article is followed by: Logistic Regression = 0.7966, Random Forest = 0.6979, SVM = 0.8152, KNN = 0.8423.

Tables:5.1

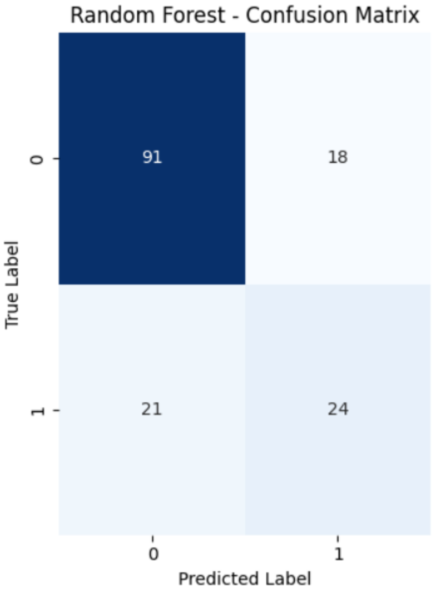
SL No	Algorithm	Accuracy	Precision	Recall	F1 Score	AUC-ROC
1	**Logistic Regression**	0.7662	0.81	0.84	0.84	0.7966
2	**Random Forest**	0.7207	0.79	0.78	0.78	0.6979
3	**SVM**	0.7833	0.7727	0.6545	0.6605	0.8152
4	**KNN**	0.7272	0.81	0.81	0.81	0.8423

Figure:



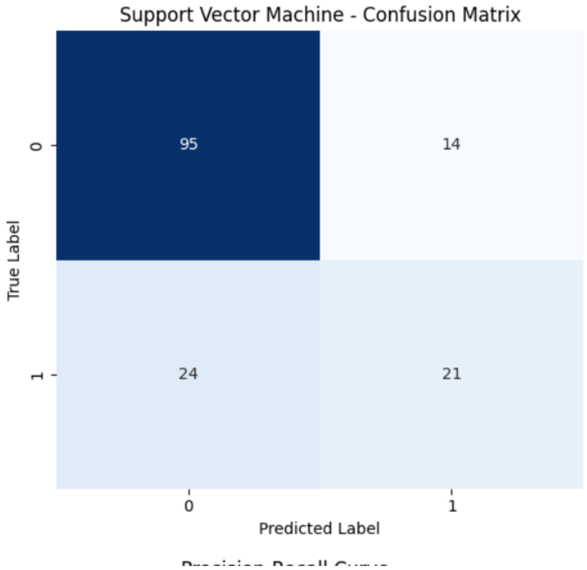
(Fig: 5.01 Confusion Matrix – Logistic Regression)

The above figure is the confusion matrix of Logistic Regression which has a very high relation between their false positive and false negative values which can give a very stable result as compared to others.



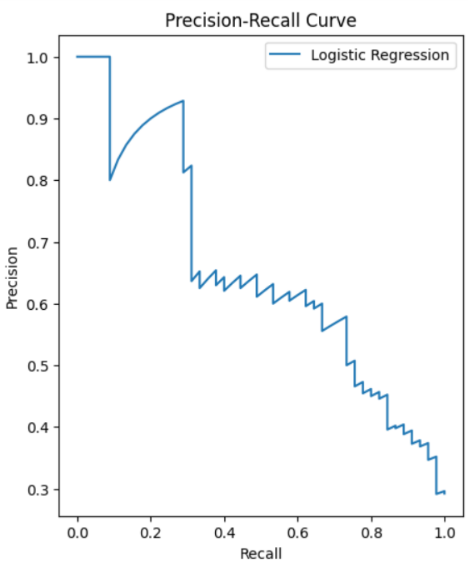
(Fig:5.02 Confusion Matrix – Random Forest)

The above figure is the Confusion Matrix of Random Forest which is used in our project and this has a high fluctuation between false positive and false negative values which may affect our model’s accuracy. So, we are avoiding the use of it in out Machine Learning Model.



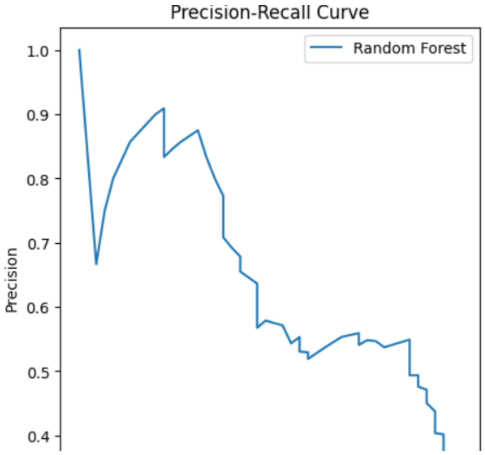
(Fig:5.03 Confusion Matrix – Support Vector Classifier)

The above figure is the confusion matrix of Support Vector Classifier which has a very high relation between their false positive and false negative values which can give a very stable result as compared to others. So, we got the recommendation of using it strongly for its better accuracy.



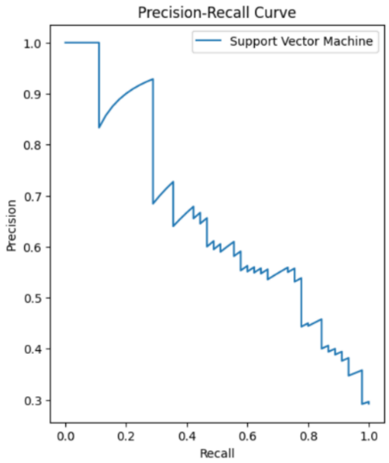
(Fig: 5.04 Precision-Recall Curve-Logistic)

If the classes are imbalanced, a precision-recall curve might give a more informative picture than the ROC curve. The precision-recall curve is a plot of precision (TP/P) versus recall (TPR) at different thresholds. This is to measure how well a model performs when the positive class is not only as small as 5% of the entire target set**.**



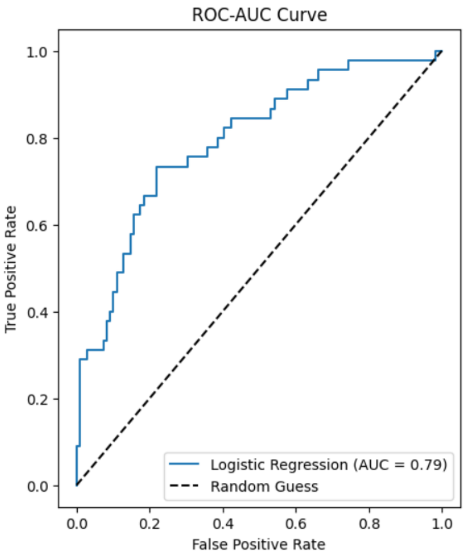
(Fig: 5.05 Precision-Recall Curve-Random Forest)

If the classes are imbalanced, a precision-recall curve might give a more informative picture than the ROC curve. The precision-recall curve is a plot of precision (TP/P) versus recall (TPR) at different thresholds. This is to measure how well a model performs when the positive class is not only as small as 5% of the entire target set.



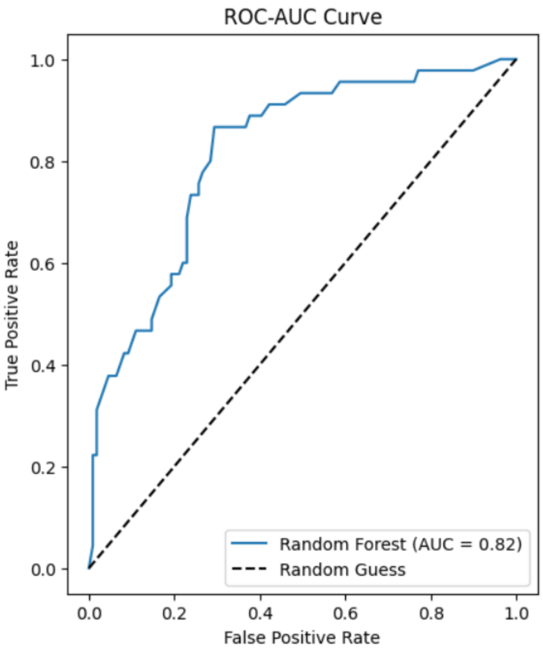
(Fig: 5.06 Precision-Recall Curve-SVC)

If the classes are imbalanced, a precision-recall curve might give a more informative picture than the ROC curve. The precision-recall curve is a plot of precision (TP/P) versus recall (TPR) at different thresholds. This is to measure how well a model performs when the positive class is not only as small as 5% of the entire target set.



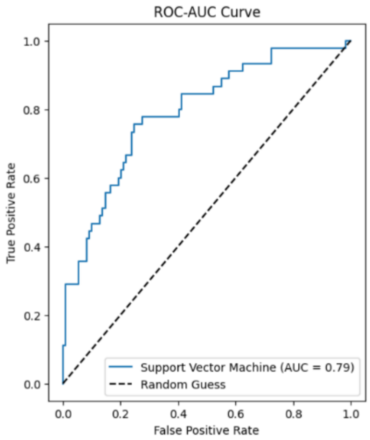
(Fig: 5.07 ROC Curve- Logistic)

The ROC curve in the image shown here means that the model you evaluate is a bad model in comparison to even random chance, this can be because: The model could be a random classifier, there may be an error in calculation, or There is a too imbalanced class. More investigation is definitely needed here to actually find out what is causing this and handle it so that really it can be used for model evaluation.



(Fig: 5.08 ROC Curve Random Forest)

The ROC curve in the image shown here means that the model you evaluate is a bad model in comparison to even random chance, this can be because: The model could be a random classifier, there may be an error in calculation, or There is a too imbalanced class. More investigation is definitely needed here to actually find out what is causing this and handle it so that really it can be used for model evaluation.



(Fig: 5.09 ROC Curve-SVC)

The ROC curve in the image shown here means that the model you evaluate is a bad model in comparison to even random chance, this can be because: The model could be a random classifier, there may be an error in calculation, or There is a too imbalanced class. More investigation is definitely needed here to actually find out what is causing this and handle it so that really it can be used for model evaluation.

5.2. Heart Disease

5.2.1. Accuracy:

Accuracy is a metric that measures how often a machine learning model correctly predicts the outcome. We got the accuracy for different models applied in our machine learning article is followed by: Logistic Regression = 0.85, Random Forest = 0.8023, KNN = 0.6224.

5.2.2. Precision:

Precision is a metric that measures how often a machine learning model correctly predicts the positive class. We got the precision for different models applied in our machine learning article is followed by: Logistic Regression = 0.7962, Random Forest = 0.77, KNN = 0.65.

5.2.3. Recall:

Recall is a metric that measures how often a machine learning model correctly identifies positive instances (true positives) from all the actual positive samples in the dataset. We got the precision for different models applied in our machine learning article is followed by: Logistic Regression = 0.8921, Random Forest = 0.82, KNN = 0.67.

5.2.4. F1 Score:

The F1 score or F-measure is described as the harmonic mean of the precision and recall of a classification model. The two metrics contribute equally to the score, ensuring that the F1 metric correctly indicates the reliability of a model. We got the F1 Score for different models applied in our machine learning article is followed by: Logistic Regression = 0.8625, Random Forest = 0.79, KNN = 0.66.

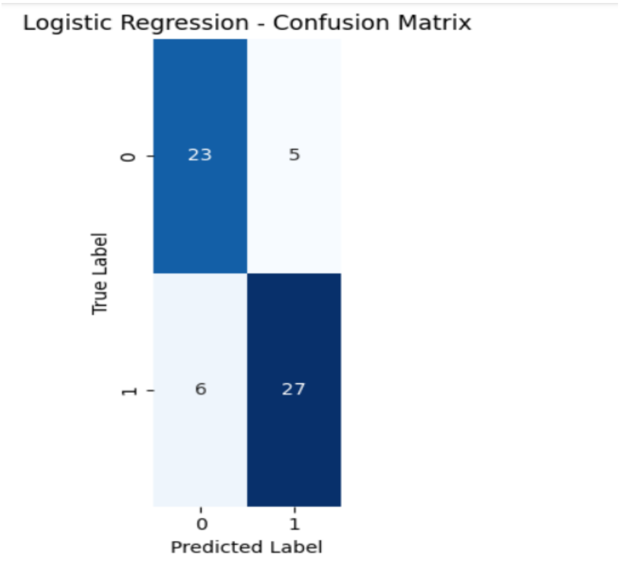
5.2.5. AUC-ROC Score:

An ROC curve, or receiver operating characteristic curve, is like a graph that shows how well a classification model performs. It helps us see how the model makes decisions at different levels of certainty. 24 We got the AUC-ROC Score for different models applied in our machine learning article is followed by: Logistic Regression = 0.8586, Random Forest = 0.84, KNN = 0.8294.

Tables:5.2

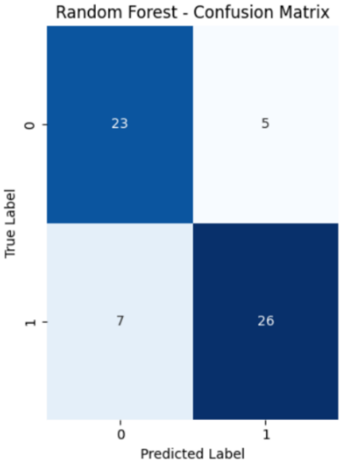
SL No	Algorithm	Accuracy	Precision	Recall	F1 Score	AUC-ROC
1	**Logistic Regression**	0.85	0.7962	0.8921	0.8625	0.8586
2	**Random Forest**	0.8023	0.77	0.82	0.79	0.84
3	**KNN**	0.6224	0.65	0.57	0.66	0.8294

Figure:



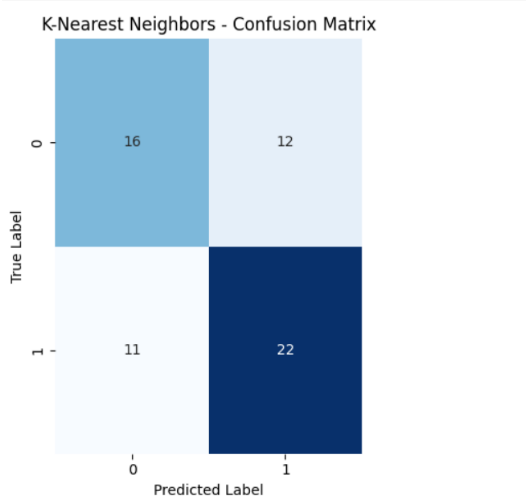
(Fig: 5.10 Confusion Matrix – Logistic Regression)

The above figure is the confusion matrix of Logistic Regression which has a very high relation between their false positive and false negative values which can give a very stable result as compared to others.



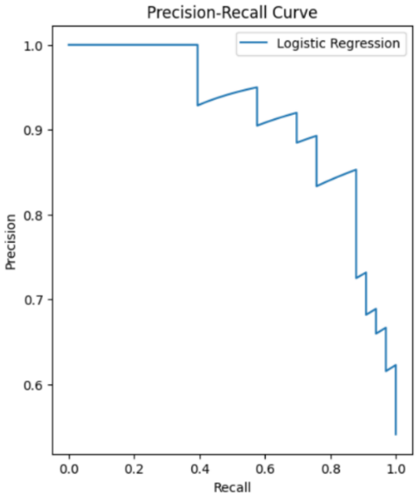
(Fig:5.11 Confusion Matrix – Random Forest)

The above figure is the Confusion Matrix of Random Forest which is used in our project and this has a high fluctuation between false positive and false negative values which may affect our model's accuracy. So, we are avoiding the use of it in out Machine Learning Model.



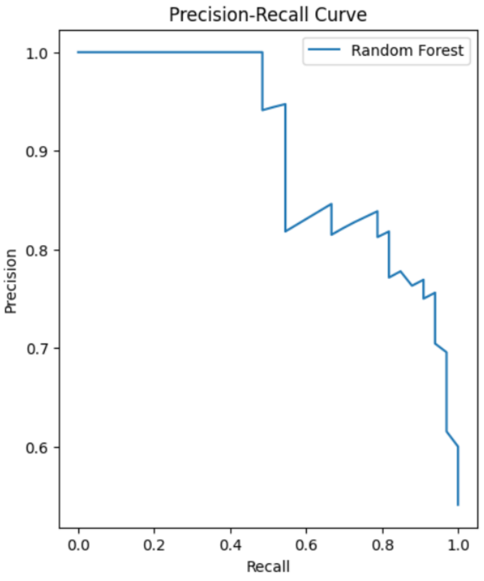
(Fig: 5.12 Confusion Matrix – K Nearest neighbor)

The above figure is the confusion matrix of K Nearest Neighbor which has a stable relation between their false positive and false negative values which can give a very stable result as compared to others.



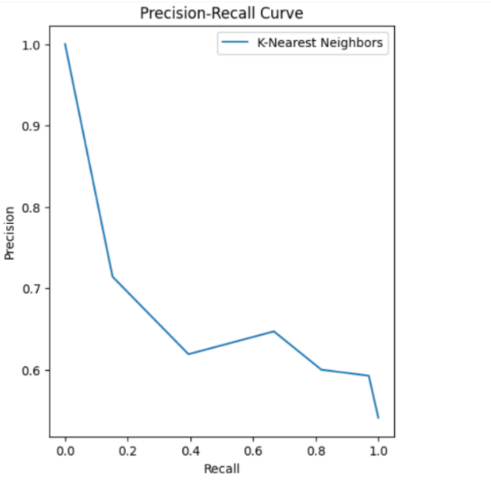
(Fig: 5.13Precision-Recall Curve-Logistic)

If the classes are imbalanced, a precision-recall curve might give a more informative picture than the ROC curve. The precision-recall curve is a plot of precision (TP/P) versus recall (TPR) at different thresholds. This is to measure how well a model performs when the positive class is not only as small as 5% of the entire target set.



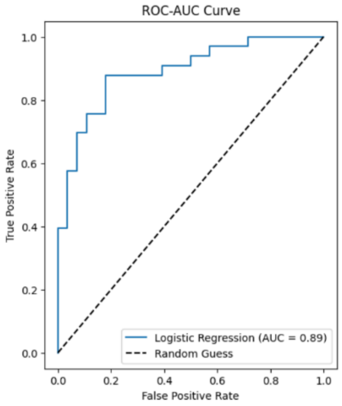
(Fig: 5.14 Precision-Recall Curve-Random Forest)

If the classes are imbalanced, a precision-recall curve might give a more informative picture than the ROC curve. The precision-recall curve is a plot of precision (TP/P) versus recall (TPR) at different thresholds. This is to measure how well a model performs when the positive class is not only as small as 5% of the entire target set.



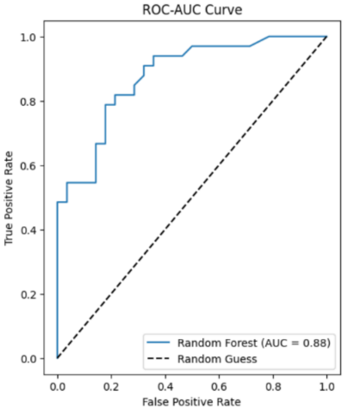
(Fig: 5.15 Precision-Recall Curve-KNN)

If the classes are imbalanced, a precision-recall curve might give a more informative picture than the ROC curve. The precision-recall curve is a plot of precision (TP/P) versus recall (TPR) at different thresholds. This is to measure how well a model performs when the positive class is not only as small as 5% of the entire target set.



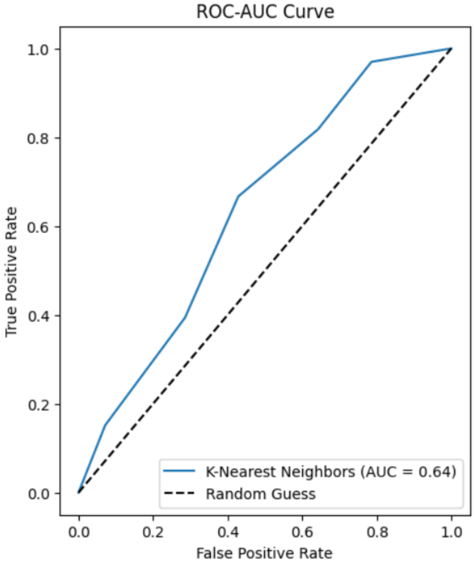
(Fig: 5.16 ROC Curve-Logistic)

The ROC curve in the image shown here means that the model you evaluate is a bad model in comparison to even random chance, this can be because: The model could be a random classifier, there may be an error in calculation, or There is a too imbalanced class. More investigation is definitely needed here to actually find out what is causing this and handle it so that really it can be used for model evaluation.



(Fig: 5.17 ROC Curve-Random Forest)

The ROC curve in the image shown here means that the model you evaluate is a bad model in comparison to even random chance, this can be because: The model could be a random classifier, there may be an error in calculation, or There is a too imbalanced class. More investigation is definitely needed here to actually find out what is causing this and handle it so that really it can be used for model evaluation.



(Fig: 5.18 ROC Curve-KNN)

The ROC curve in the image shown here means that the model you evaluate is a bad model in comparison to even random chance, this can be because: The model could be a random classifier, there may be an error in calculation, or There is a too imbalanced class. More investigation is definitely needed here to actually find out what is causing this and handle it so that really it can be used for model evaluation.

5.3. Parkinson's Disease

5.3.1. Accuracy:

Accuracy is a metric that measures how often a machine learning model correctly predicts the outcome. We got the accuracy for different models applied in our machine learning article is followed by: Logistic Regression = 0.87, Random Forest = 0.7948, SVM = 0.87, KNN = 0.74.

5.3.2. Precision:

Precision is a metric that measures how often a machine learning model correctly predicts the positive class. We got the precision for different models applied in our machine learning article is followed by: Logistic Regression = 0.88, Random Forest = 0.90, SVM = 0.8590, KNN = 0.89.

5.3.3. Recall:

Recall is a metric that measures how often a machine learning model correctly identifies positive instances (true positives) from all the actual positive samples in the dataset. We got the precision for different models applied in our machine learning article is followed by: Logistic Regression = 0.97, Random Forest = 0.84, SVM = 0.8752, KNN = 0.77.

5.3.4. F1 Score:

The F1 score or F-measure is described as the harmonic mean of the precision and recall of a classification model. The two metrics contribute equally to the score, ensuring that the F1 metric correctly indicates the reliability of a model. We got the F1 Score for different models applied in our machine learning article is followed by: Logistic Regression = 0.92, Random Forest = 0.87, SVM = 0.8962, KNN = 0.83.

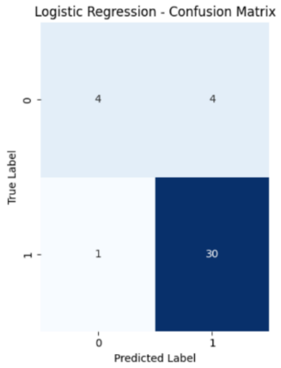
5.3.5. AUC-ROC Score:

An ROC curve, or receiver operating characteristic curve, is like a graph that shows how well a classification model performs. It helps us see how the model makes decisions at different levels of certainty. 24 We got the AUC-ROC Score for different models applied in our machine learning article is followed by: Logistic Regression = 0.8798, Random Forest = 0.9754, SVM = 0.9043, KNN = 0.8238.

Tables:2

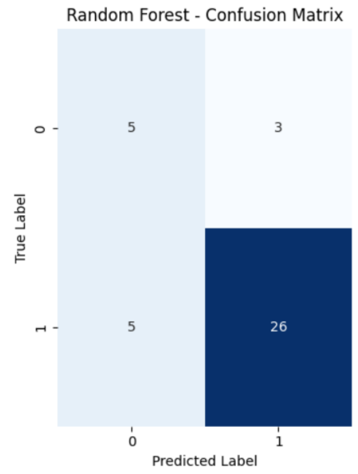
SL No	Algorithm	Accuracy	Precision	Recall	F1 Score	AUC-ROC
1	**Random Forest**	0.7948	0.90	0.84	0.87	0.9754
2	**SVM**	0.87	0.8590	0.8752	0.8962	0.9043
3	**KNN**	0.74	0.89	0.77	0.83	0.8238
4	**Logistic Regression**	0.87	0.88	0.97	0.92	0.8798

Figure:



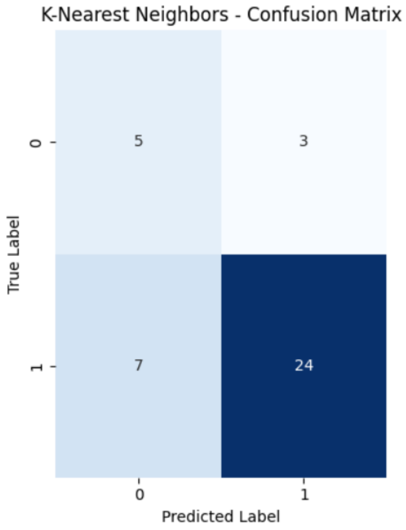
(Fig: 5.19 Confusion Matrix – Logistic Regression)

The above figure is the confusion matrix of Logistic Regression which has a very high relation between their false positive and false negative values which can give a very stable result as compared to others.



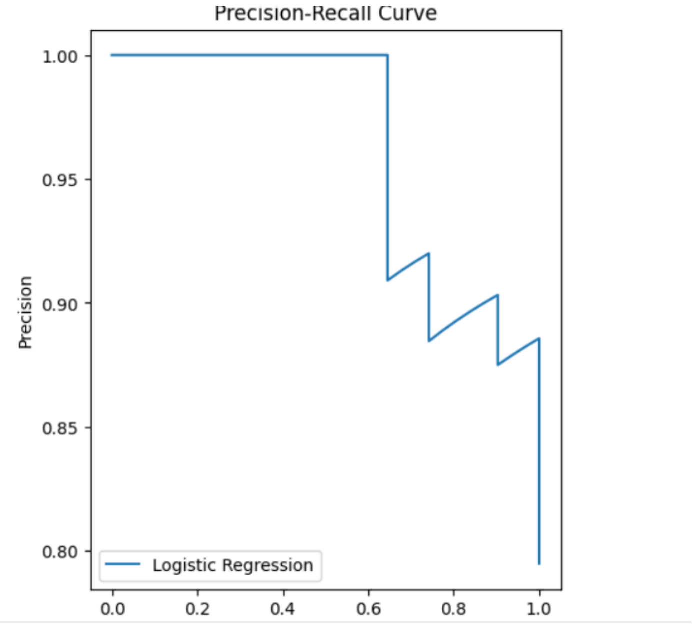
(Fig:5.20 Confusion Matrix – Random Forest)

The above figure is the Confusion Matrix of Random Forest which is used in our project and this has a high fluctuation between false positive and false negative values which may affect our model's accuracy. So, we are avoiding the use of it in out Machine Learning Model.



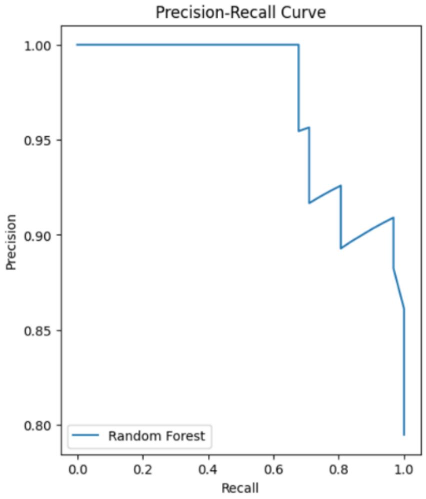
(Fig: 5.21 Confusion Matrix – K Nearest neighbor)

The above figure is the confusion matrix of K Nearest Neighbor which has a stable relation between their false positive and false negative values which can give a very stable result as compared to others.



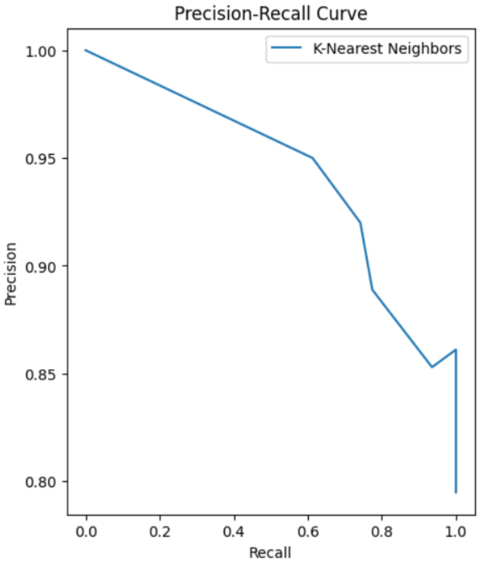
(Fig: 5.22 Precision-Recall Curve- Logistic)

If the classes are imbalanced, a precision-recall curve might give a more informative picture than the ROC curve. The precision-recall curve is a plot of precision (TP/P) versus recall (TPR) at different thresholds. This is to measure how well a model performs when the positive class is not only as small as 5% of the entire target set.



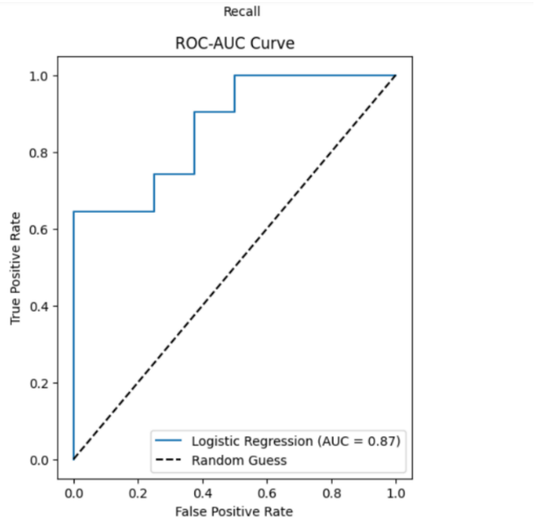
(Fig: 5.23 Precision-Recall Curve- Random Forest)

If the classes are imbalanced, a precision-recall curve might give a more informative picture than the ROC curve. The precision-recall curve is a plot of precision (TP/P) versus recall (TPR) at different thresholds. This is to measure how well a model performs when the positive class is not only as small as 5% of the entire target set.



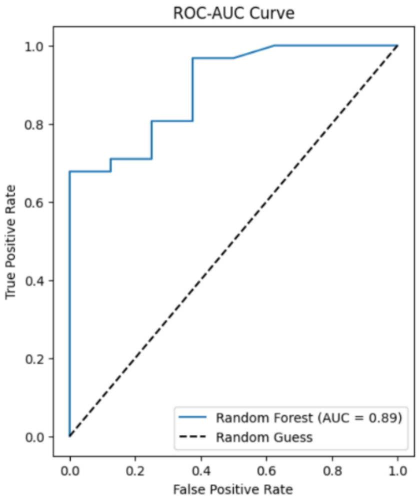
(Fig: 5.24 Precision-Recall Curve-KNN)

If the classes are imbalanced, a precision-recall curve might give a more informative picture than the ROC curve. The precision-recall curve is a plot of precision (TP/P) versus recall (TPR) at different thresholds. This is to measure how well a model performs when the positive class is not only as small as 5% of the entire target set.



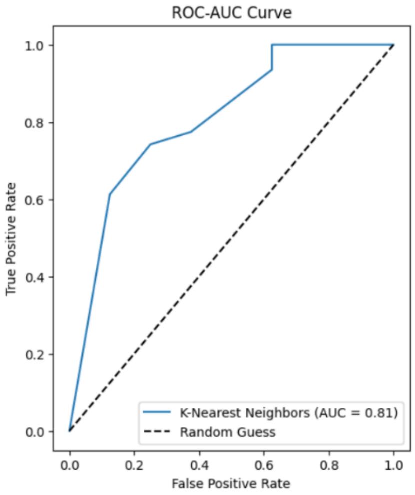
(Fig: 5.25 ROC Curve - Logistic)

The ROC curve in the image shown here means that the model you evaluate is a bad model in comparison to even random chance, this can be because: The model could be a random classifier, there may be an error in calculation, or There is a too imbalanced class. More investigation is definitely needed here to actually find out what is causing this and handle it so that really it can be used for model evaluation**.**



(Fig: 5.26 ROC Curve – Random Forest)

The ROC curve in the image shown here means that the model you evaluate is a bad model in comparison to even random chance, this can be because: The model could be a random classifier, there may be an error in calculation, or There is a too imbalanced class. More investigation is definitely needed here to actually find out what is causing this and handle it so that really it can be used for model evaluation**.**



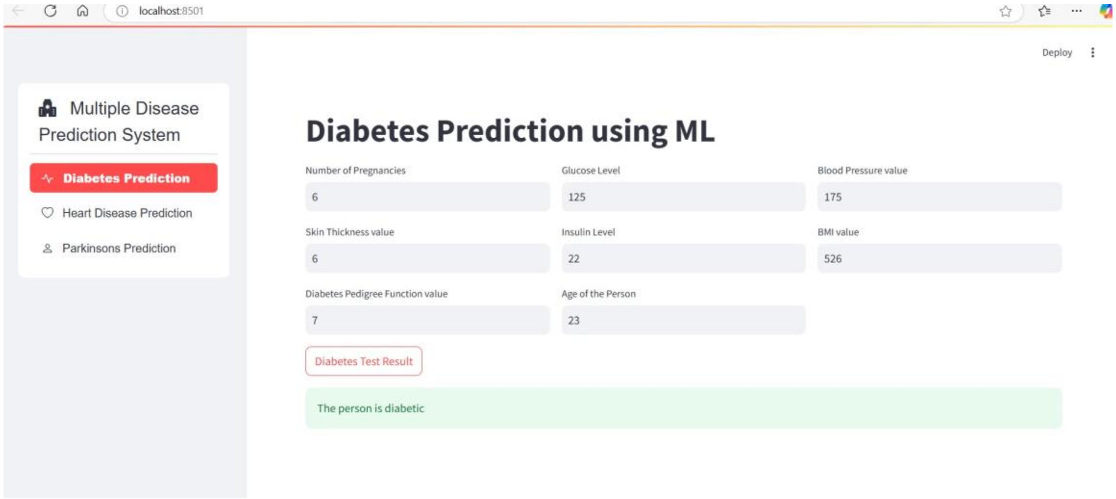
(Fig: 5.27 ROC Curve- KNN)

The ROC curve in the image shown here means that the model you evaluate is a bad model in comparison to even random chance, this can be because: The model could be a random classifier, there may be an error in calculation, or There is a too imbalanced class. More investigation is definitely needed here to actually find out what is causing this and handle it so that really it can be used for model evaluation.

User Interface:

We create the user Interface through the Stream Lead where the multiple features to be input that showing the result.

For Diabetes Prediction:



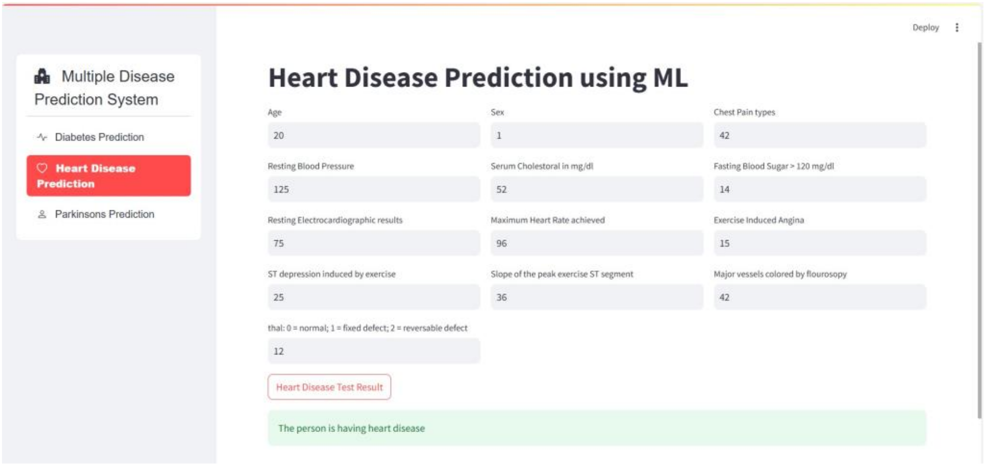
Diabetes Prediction using ML

Number of Pregnancies: 6
Glucose Level: 125
Blood Pressure value: 175
Skin Thickness value: 6
Insulin Level: 22
BMI value: 526
Diabetes Pedigree Function value: 7
Age of the Person: 23

The person is diabetic

(Fig: 5.28 UI of Diabetics Disease)

For Heart Prediction:



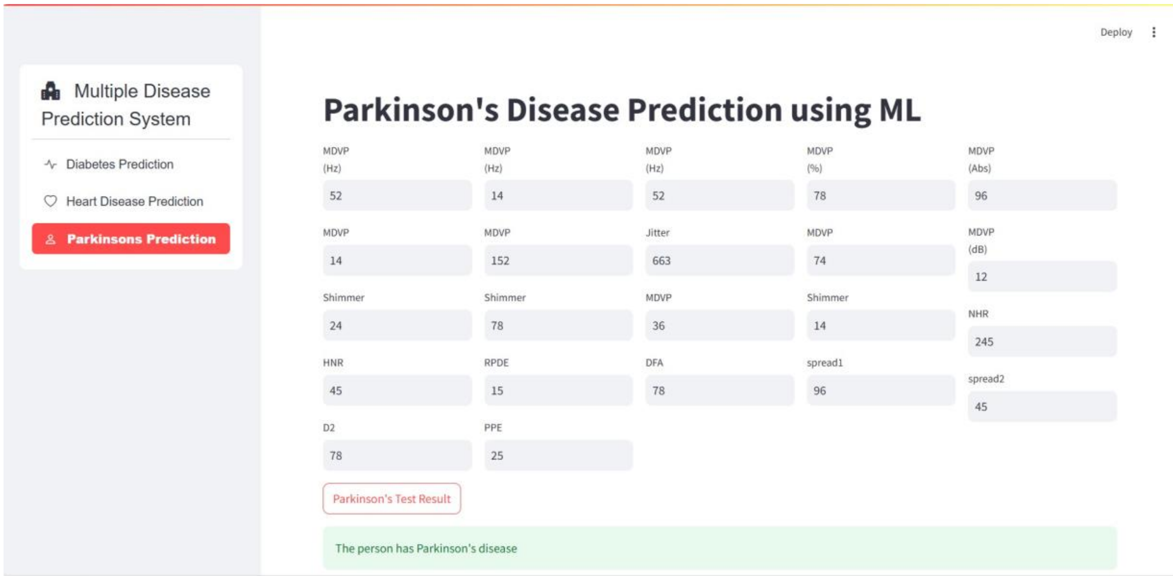
Heart Disease Prediction using ML

Age: 20
Sex: 1
Chest Pain types: 42
Resting Blood Pressure: 125
Serum Cholestoral in mg/dl: 52
Fasting Blood Sugar > 120 mg/dl: 14
Resting Electrocardiographic results: 75
Maximum Heart Rate achieved: 96
Exercise Induced Angina: 15
ST depression induced by exercise: 25
Slope of the peak exercise ST segment: 36
Major vessels colored by flourosopy: 42
thal: 0 = normal; 1 = fixed defect; 2 = reversable defect: 12

The person is having heart disease

(Fig: 5.29 UI of Heart Disease)

For Parkinson's Disease:



Parkinson's Disease Prediction using ML

MDVP (Hz): 52

MDVP (Hz): 14

MDVP (Hz): 52

MDVP (%): 78

MDVP (Abs): 96

MDVP: 14

MDVP: 152

Jitter: 663

MDVP: 74

MDVP (dB): 12

Shimmer: 24

Shimmer: 78

MDVP: 36

Shimmer: 14

NHR: 245

HNR: 45

RPDE: 15

DFA: 78

spread1: 96

spread2: 45

D2: 78

PPE: 25

The person has Parkinson's disease

(Fig: 5.30 UI of Parkinson's Disease)

Chapter-6

CONCLUSIONS AND SCOPE FOR FUTURE

Our project focused on the development of a Multiple Disease Prediction using machine learning algorithms. The aim was to create a robust system capable of predicting various diseases simultaneously, namely Heart disease, Lung’s disease, Diabetes, Breast and Parkinson's disease. Because of this project the user doesn’t need to traverse different websites which saves time as well. Diseases if predicted early can increase your life expectancy as well as save you from financial troubles. For this purpose, we have used various machine learning algorithms like Random Forest, Logistic Regression, SVM, andK nearest neighbor (KNN) to achieve maximum accuracy. It eliminates the need for long-distance travel by bridging the gap between patients and healthcare professionals, especially in rural or underserved areas.

REFRENCE

1. Arumugam, K., Naved, M., Shinde, P. P., Leiva-Chauca, O., Huaman-Osorio, A., & Gonzales-Yanac, T. (2023). Multiple disease prediction using Machine learning algorithms. *Materials Today: Proceedings*, *80*, 3682-3685.
2. Yaganteeswarudu, A. (2020, June). Multi disease prediction model by using machine learning and Flask API. In *2020 5th International Conference on Communication and Electronics Systems (ICCES)* (pp. 1242-1246). IEEE.
3. Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learningalgorithms for disease prediction. *BMC medical informatics and decision making*, *19*(1), 1-16.
4. Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learningalgorithms for disease prediction. *BMC medical informatics and decision making*, *19*(1), 1-16.
5. Khadir, M. A., Mohd, A., Ali, M., & Khan, P. A. (2023). Multiple Disease Prediction System Using Machine Learning. *Mathematical Statistician and Engineering Applications*, *72*(1), 1435-1445.
6. Talasila, B., Kolli, S., Kumar, K. V. N., & Anudeep, P. (2021). Symptoms based multiple disease prediction model using machine learning approach. *International Journal of Innovative Technology and Exploring Engineering*.
7. Singh, A., Yadav, A., Shah, S., & Nagpure, R. (2022). Multiple Disease Prediction System. *International Research Journal of Engineering and Technology*.
8. Pulicherla, P., Akash, P., Raviteja, P., Krishna, V. S., & Vikas, R. (2023). Human Multiple Disease Prediction. *Journal of Engineering Sciences*, *14*(03).
9. Vasavi, D., Venkatesh, D., Kumar, S. S., Sahaja, S., & Kumar, V. S. MULTIPLE DISEASE PREDICTION USING MACHINE LEARNING.
10. Bayati, M., Bhaskar, S., & Montanari, A. (2018). Statistical analysis of a low cost method for multiple diseaseprediction. *Statistical methods in medical research*, *27*(8), 2312-2328.
11. Bayati, M., Bhaskar, S., & Montanari, A. (2015). A low-cost method for multiple disease prediction. In *AMIA Annual Symposium Proceedings* (Vol. 2015, p. 329). American Medical Informatics Association.
12. Dubey, A. K. Optimized hybrid learning for multi disease prediction enabled by lion with butterfly optimization algorithm. Sādhanā 46, 63 (2021).

13. Beg, A. A., Maqsood, F., & Siddiqi, S. (2023). Multiple disease prediction system using ML. *International Journal Of Engineering And Management Research*, *13*(3), 88-94.

14. Unnithan, D. R., & Jeba, J. R. (2024). A novel framework for multiple disease prediction in telemedicine systems using deep learning. *Automatika*, *65*(3), 763-777.

15. Olajide, A. O. (2022). Estimating The Accuracy of Classifiers in Analyzing Multiple Diseases. *International Journal of Research and Innovation in Applied Science*, *7*(9), 92-96.

16. Prakaash, A. S., Sivakumar, K., Surendiran, B., Jagatheswari, S., & Kalaiarasi, K. (2022). Design and development of modified ensemble learning with weighted RBM features for enhanced multi-disease prediction model. *New Generation Computing*, *40*(4), 1241-1279.

17. Ramani, R., Devi, K. V., & Soundar, K. R. (2020). MapReduce-based big data framework using modified artificial neural network classifier for diabetic chronic disease prediction. *Soft Comput*, *24*(21), 16335- 16345.

18. Zhang, S., Yang, F., Wang, L., Si, S., Zhang, J., & Xue, F. (2023). Personalized prediction for multiple chronic diseases by developing the multi-task Cox learning model. *PLoS Computational Biology*, *19*(9), e1011396.

19. Rashid, J., Batool, S., Kim, J., Wasif Nisar, M., Hussain, A., Juneja, S., & Kushwaha, R. (2022). An augmented artificial intelligence approach for chronic diseases prediction. *Frontiers in Public Health*, *10*, 860396.

20. Buragadda, S., VP, S. K. P., Kavya, D. K., & Khanam, S. S. (2023). Multi Disease Classification System Based on Symptoms using The Blended Approach. *International Research Journal on Advanced Science Hub*, *5*(03).