

PBR VISVODAYA INSTITUTE OF TECHNOLOGY AND SCIENCE



DIABETES PREDICTION SYSTEM

A MACHINE LEARNING PROJECT

—

UNDER THE GUIDANCE OF:

Mr. Ch. Srinivasulu Reddy
(Assistant Professor)

SUBMITTED BY: BATCH 09 CSE - AI

Shaik Afsana	Annam Pavani
Edicharla Nani	Udaygiri Uday

ABSTRACT

Diabetes, a chronic and life-threatening disease, presents challenges in timely identification. Leveraging machine learning, our project develops a predictive model for early diabetes detection. Five ML models were evaluated on the 'Early stage diabetes risk prediction dataset,' resulting in Random Forest achieving 97.2% accuracy. Deployed via MySQL and Streamlite, our system offers a comprehensive web platform for user authentication, data storage, and interactive access to diabetes-related information. Our solution aims to enhance early intervention, reducing diabetes-related complications.

>> INTRODUCTION

>> PROBLEM STATEMENT

>> PROPOSED METHODOLOGY

>> REQUIREMENTS

>> MODULES INVOLVED

>> RESULTS

>> LIMITATIONS

>> CONCLUSION

>> FUTURE ENHANCEMENTS

CONTENTS



INTRODUCTION

- Diabetes is a chronic metabolic disease characterized by elevated blood glucose levels, leading to severe damage to various organs over time.
- Type 2 diabetes, the most common form, occurs when the body becomes resistant to insulin or doesn't produce enough insulin. Type 1 diabetes, on the other hand, results from the pancreas producing little or no insulin.
- The global prevalence of diabetes, especially type 2, has surged in the past few decades, posing a significant public health challenge.
- Symptoms of diabetes include frequent urination, thirst, hunger, weight loss, vision changes, and fatigue. Early diagnosis and awareness of risk factors are crucial.



SYMPTOMS



Always Hungry



Always Tired



Increased Thirst



Frequent Urination



Nausea



Blurry Vision



Sudden Weight Loss



Sexual Problems



Slow Healing of Wounds

COMPLICATIONS



Permanent Kidney Damage



Cerebrovascular Diseases



Eye Damage



Diabetic Foot



Peripheral Neuropathy



Coronary Heart Disease

TREATMENT



Diabetes Medications



Blood Sugar Monitoring



Insulin Pills



Insulin Injections

“ PREVENTION IS BETTER THAN CURE ”



**Healthy
Diet**



**Weight
Control**



**Break Bad
Habits**



Excercise

PROBLEM STATEMENT

- Diabetes poses significant dangers if left undetected and unmanaged.
- Traditional diagnostic methods present challenges such as time-consuming procedures, unnecessary tests, and financial burdens on patients and healthcare systems.
- However, the increasing use of machine learning algorithms helps address these challenges by analyzing healthcare data to identify hidden patterns for health diagnosis.
- Following the prevalence of prediction systems based on numerical values like BMI and glucose levels, barriers emerged for users lacking medical expertise or knowledge of diagnostic parameters.

- The PIMA INDIANS DIABETES DATASET serves as a prominent example of such systems.

- We thought of the growing need for user-friendly prediction systems that simplify the diagnostic process and empower individuals to monitor their health proactively.

SOLUTION: Symptom based Prediction System

ATTRIBUTE	DESCRIPTION	VALUE
Preg	Number of pregnancies	[0 – 17]
Plas	Plasma glucose concentration in an oral glucose tolerance test	[0-199]
Pres	Diastolic blood pressure	[0-122]
Skin	Triceps skin fold thickness	[0-99]
Insu	2-Hour serum insulin	[0-846]
Mass	Body mass index	[0-67]
Pedi	Diabetes pedigree function	[0-2.45]
Age	Age of an individual	[21-81]
class	Tested positive / negative	(0,1)

PROPOSED METHODOLOGY

- The proposed methodology involves conducting an Exploratory Data Analysis (EDA) of the Sylhet Diabetes Dataset to understand its underlying patterns and relationships.
- Subsequently, five machine learning algorithms, including Logistic Regression, Decision Tree, Support Vector Machine, Gradient Boost, and Random Forest, will be trained and evaluated to detect possible cases of diabetes.
- The model with best accuracy will then be deployed using a web framework, leveraging Streamlit, to make it accessible through a web browser.

REQUIREMENTS

SOFTWARE

OPERATING SYSTEM: Windows 10 or 11

LANGUAGE: Python 3.10

FRONT - END: Streamlit 3.7

BACK - END: Google Collab, MySQL

HARDWARE

PROCESSOR: Minimum Intel i3

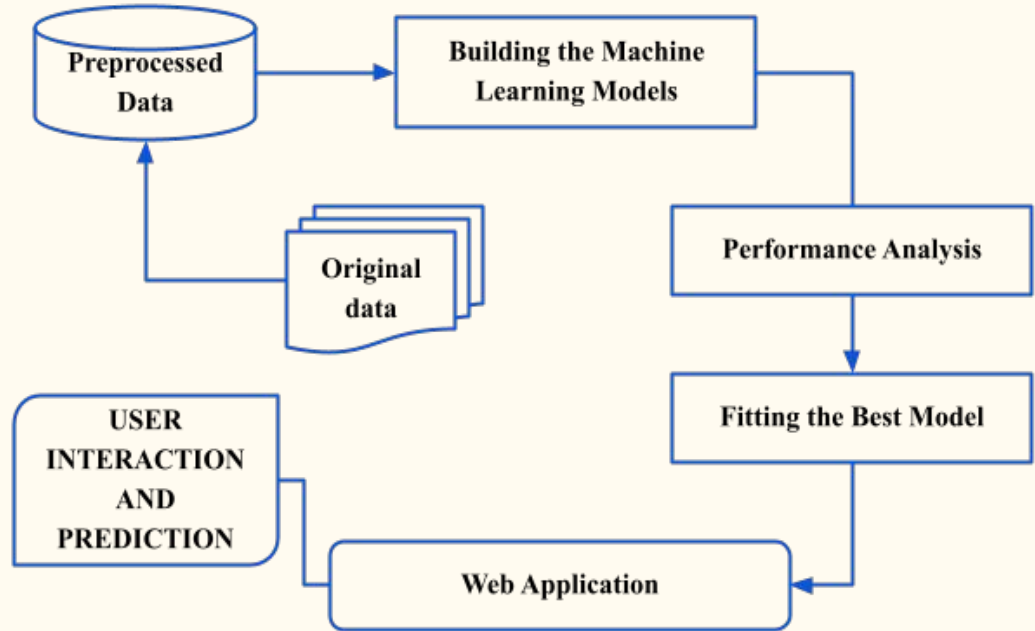
RAM: Minimum 4GB

HARD DISK: Minimum 250GB

KEY BOARD: Standard Windows Keyboard

MODULES INVOLVED

1. Data Collection
2. Data Preprocessing
3. Model Building
4. Performance Analysis
5. Loading best model and Deployment



BLOCK DIAGRAM

1. DATA COLLECTION

This module involves the acquisition of the "Early stage diabetes risk prediction dataset" from the UCI Machine Learning Repository. The dataset comprises responses gathered through direct questionnaires from patients at Sylhet Diabetes Hospital in Sylhet and has been approved by medical professionals. It includes 16 features and 1 target variable, with the target variable indicating the presence or absence of diabetes.

FEATURE NAME	RANGE
Age	20-65
Gender	Male/Female
Polyuria	Yes/No
Polydipsia	Yes/No
Sudden Weight Loss	Yes/No
Visual Blurring	Yes/No
Delayed Healing	Yes/No
Alopecia	Yes/No
Partial Paresis	Yes/No
Itching	Yes/No
Irritability	Yes/No
Obesity	Yes/No
Weakness	Yes/No
Muscle Stiffness	Yes/No
CLASS	Positive/Negative

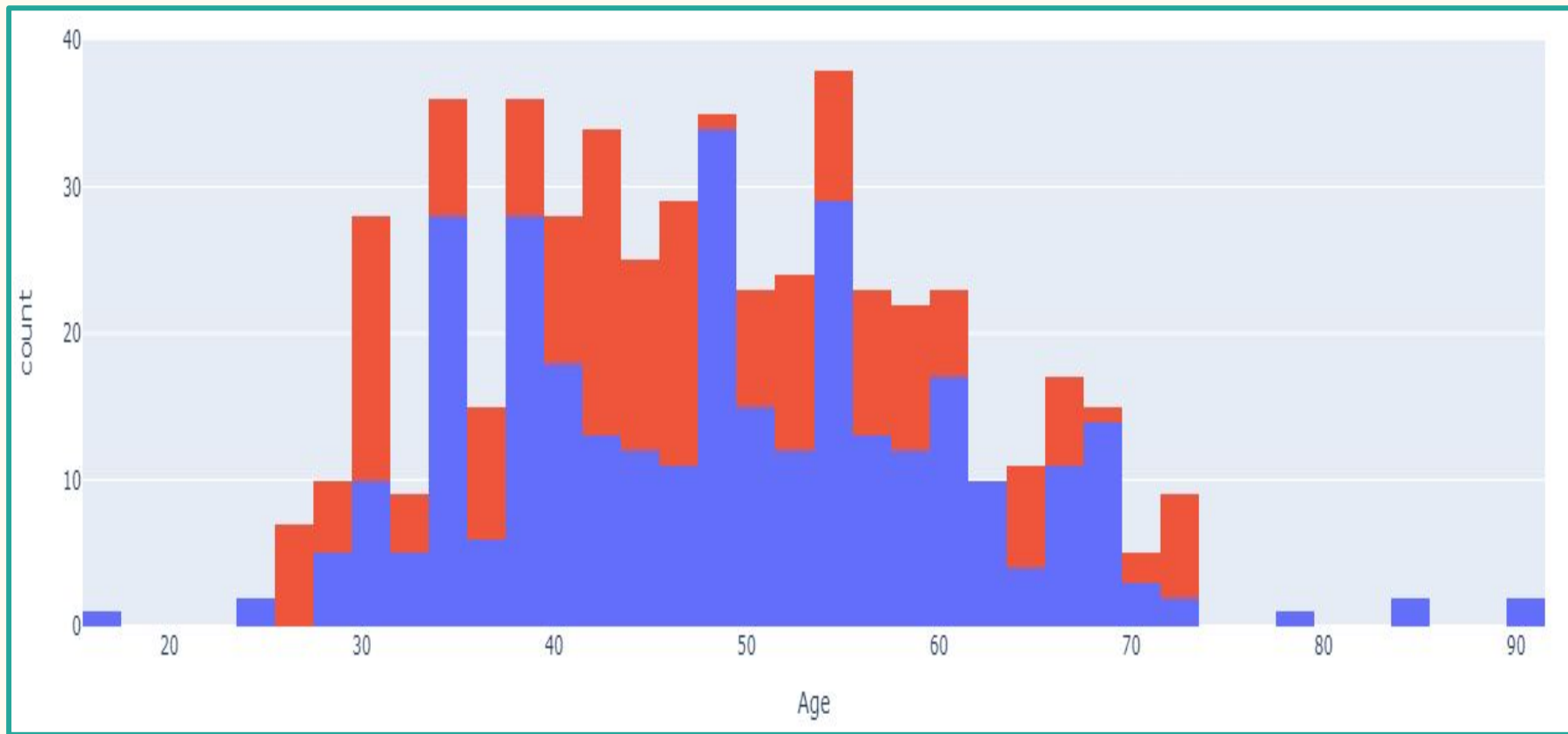
2. DATA PREPROCESSING

STEP 1: DATA QUALITY ASSESSMENT:

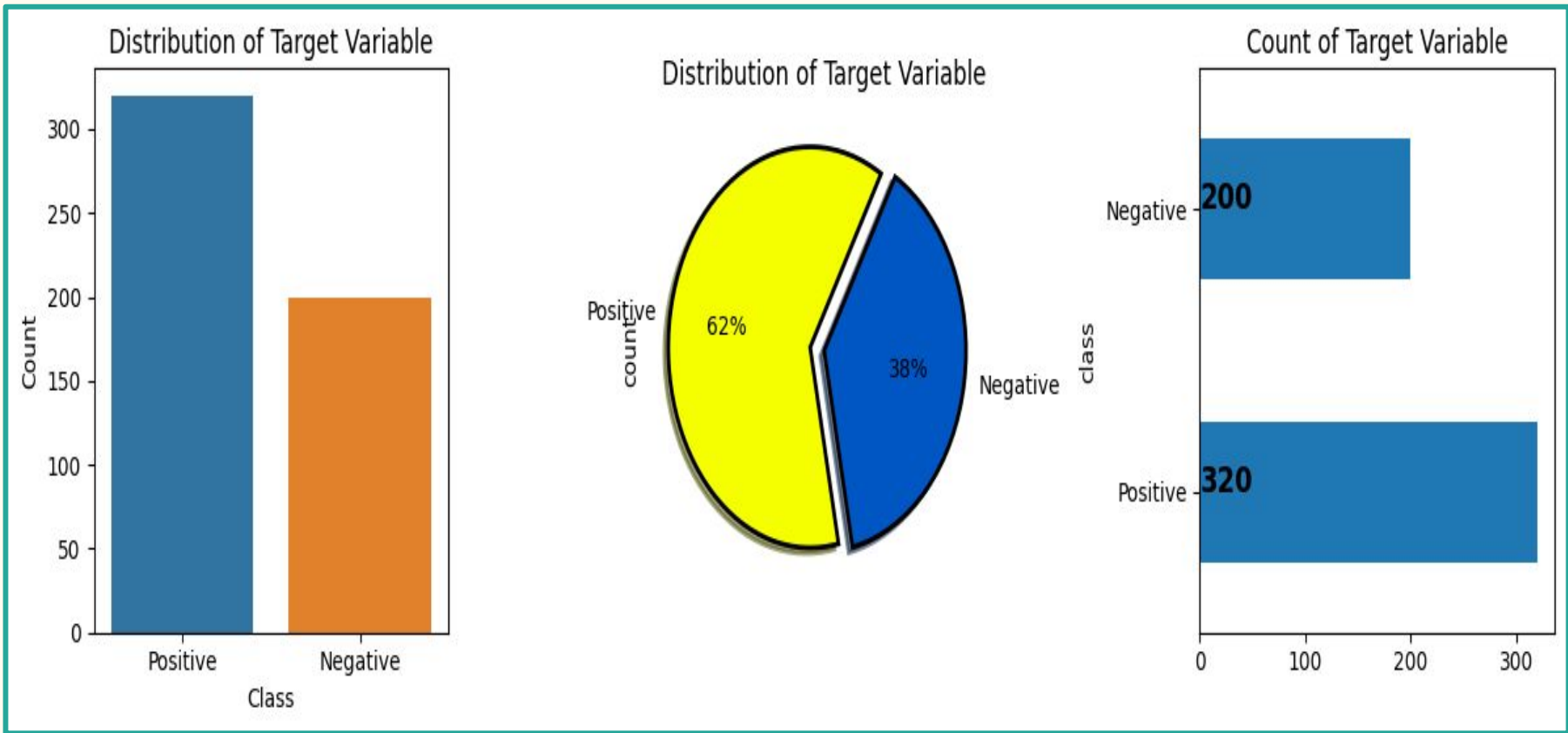
- It is conducted to address potential issues arising from imperfect data. Data is often sourced from multiple unreliable sources and may vary in format, it is crucial to assess and mitigate data quality issues before proceeding with analysis. Common challenges include missing values.
- To handle missing values effectively, two main strategies are employed:
 - Eliminating rows with missing data
 - Imputation with mean, median, or mode values.

STEP 2: EXPLORATORY DATA ANALYSIS

- This is a crucial step in understanding the characteristics and patterns present in a dataset before applying machine learning algorithms. The primary goal of EDA is to gain insights into the data, understand its underlying structure, detect patterns, anomalies, and relationships between variables.
- **Tools for EDA:** Popular libraries for EDA in Python include pandas, NumPy, Matplotlib, Seaborn, and Plotly for data manipulation, visualization, and statistical analysis.
- **Distribution of Different Features:** The analysis of the diabetes dataset reveals the distribution of positive and negative cases across various features. Using custom-colored count plots, we visualize the Age Distribution, class distribution etc..., enabling easy comparison of positive and negative instances.

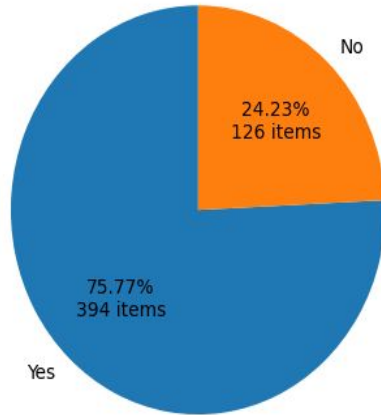


AGE DISTRIBUTION

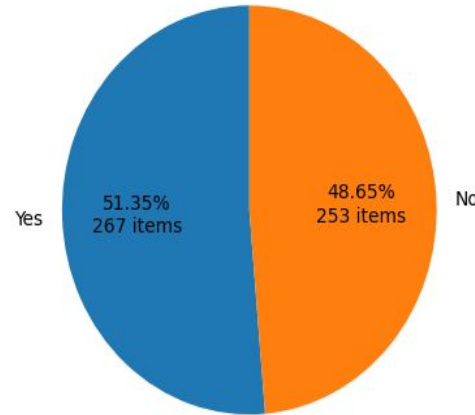


CLASS DISTRIBUTION

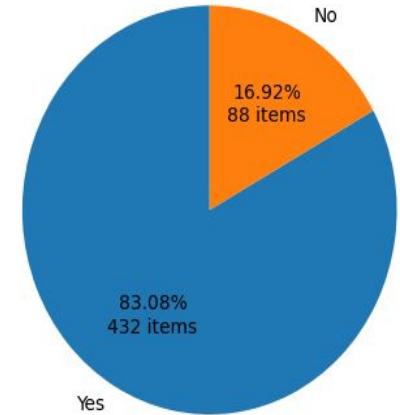
- **Occurrences of Symptoms in Patients:** Pie Chart is used to visualize the occurrence of symptoms among patients in the diabetes dataset.



Polydipsia



Polyphagia



Polyuria

STEP 3: ENCODING

- Categorical features are converted into numerical representations to enable machine learning models to process them effectively.
- Label Encoding is employed to convert categorical variables into numerical representations. This technique assigns a unique integer value to each category within a feature.
- Label Encoder is typically utilized for features where categories lack a natural order or hierarchy.
- In our dataset, for example the categorical features represent symptoms, where the presence or absence of a symptom is denoted by "Yes" or "No" respectively. Label Encoder can be applied to encode these categories as 0 and 1 respectively.

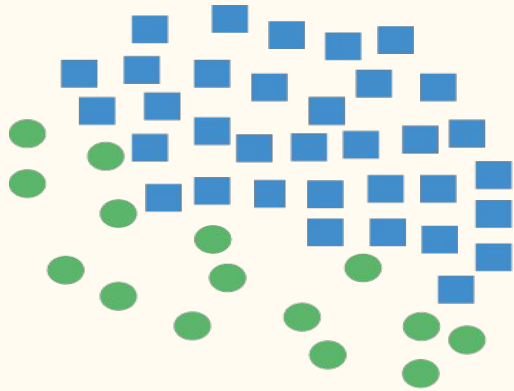
STEP 4: FEATURE SCALING

- Dropped the target variable from dataset, separated feature and target variables.
- The dataset is then divided into two subsets - training set and testing set.
- Numerical features in the dataset are standardized to have a mean of 0 and a standard deviation of 1. This process ensures that all features are on a similar scale, which is essential for many machine learning algorithms to perform effectively.

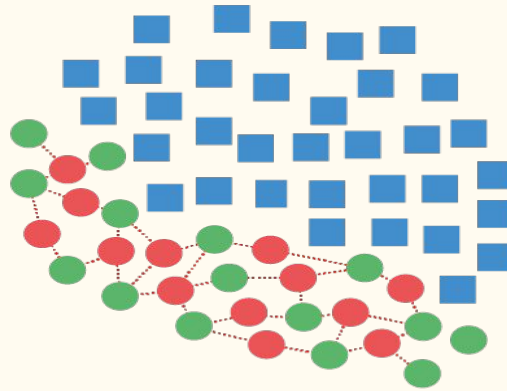
STEP 5: DEALING WITH IMBALANCED DATA

- Dealing with imbalanced classes is a crucial step in classification problems to ensure the model's robustness and performance otherwise model may Overfit.
- Imbalanced classes occur when one class significantly outnumbers the other(s) in the dataset, leading to biased model performance. Techniques such as under sampling, oversampling are employed to address class imbalance.

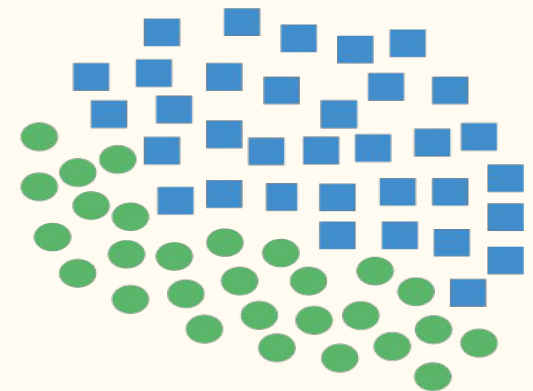
SMOTE (Synthetic Minority Over-sampling Technique) is a method used to address class imbalance in machine learning datasets by generating synthetic samples for the minority class. It creates new instances by selecting similar instances from the minority class and creating new instances along the line segments joining them in the feature space. By creating synthetic samples, SMOTE helps balance the class distribution without introducing exact duplicates, thus reducing the risk of overfitting.



Original Dataset



Generating Samples



Resampled Dataset

3. MODEL BUILDING

Now the the data is preprocessed for Model Building.

HYPER PARAMETER TUNING is the process of optimizing the hyper parameters of a machine learning model to improve its performance.

- Hyperparameters are parameters that are not learned during the training process but are set prior to training.
- Examples of hyperparameters include the learning rate, regularization strength, number of hidden layers in a neural network, and the choice of kernel in support vector machines.

GRID SEARCH: Grid search is a technique used to find the optimal hyperparameters for a machine learning model by exhaustively searching through a specified subset of the hyperparameter space.

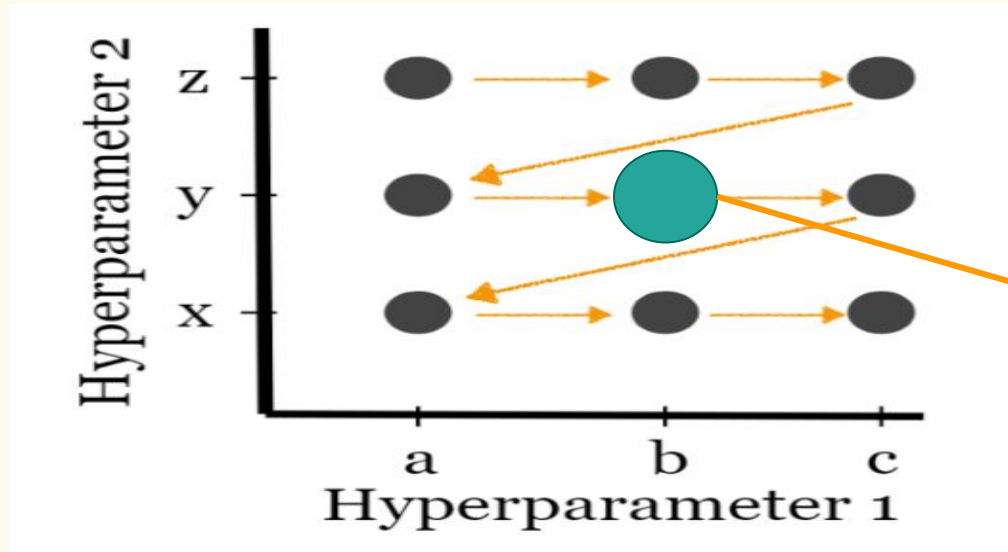
Working: The grid search process typically involves the following steps:

1. Define a grid of hyperparameter values for the model.
2. Train and evaluate the model for each combination of hyperparameter values on the grid using cross-validation.
3. Select the combination of hyperparameters that maximizes the performance metric of interest, such as accuracy, precision, recall, F1 score, or area under the ROC curve (AUC).

EXAMPLE: Hyperparameter one: $\{a,b,c\}$

Hyperparameter two: $\{x,y,z\}$

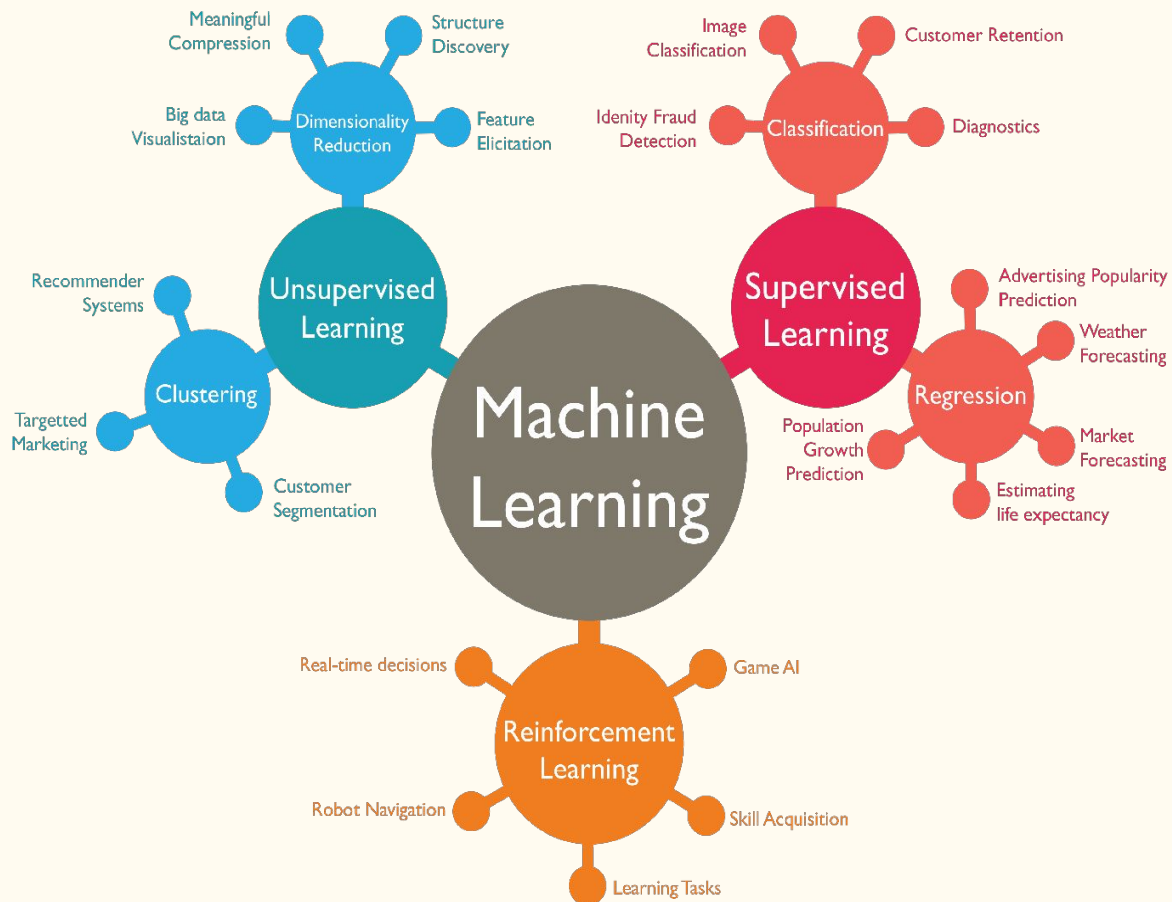
The Grid Search visualization is shown below



**Best hyperparameters
are $\{\{b\},\{y\}\}$**

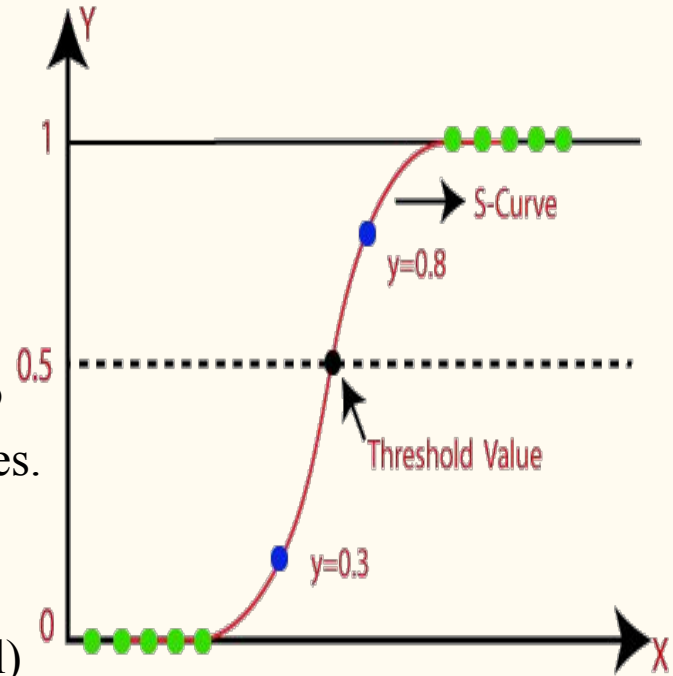
ALGORITHMS INVOLVED

1. Logistic Regression
2. Decision Tree Classifier
3. Random Forest
4. Support Vector Machine
5. Gradient Boost



LOGISTIC REGRESSION:

- It is a supervised classification algorithm used for estimating the probability of binary responses based on one or more predictors.
- It is employed when the goal is to classify data items into categories, such as determining whether a patient is positive or negative for diabetes.
- Logistic regression utilizes the sigmoid function to predict probabilities of positive and negative classes. Value lies between 0 to 1.
- **Sigmoid Function:** $P = 1 / (1 + e^{(-a - bx)})$
(P = probability, a and b = parameters of the model)



DECISION TREE

- Decision trees are fundamental classification models that are supervised and suitable for categorical response variables. They utilize a tree-like structure based on input features to describe the classification process.
- Each node in the tree represents an input feature, and the tree construction involves selecting the feature that provides the highest information gain to predict the output.

CONSTRUCTION:

1. A tree can be “learned” by splitting the source set into subsets based on an attribute value test.
2. This process is repeated on each derived subset in a recursive manner called recursive partitioning.
3. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions.

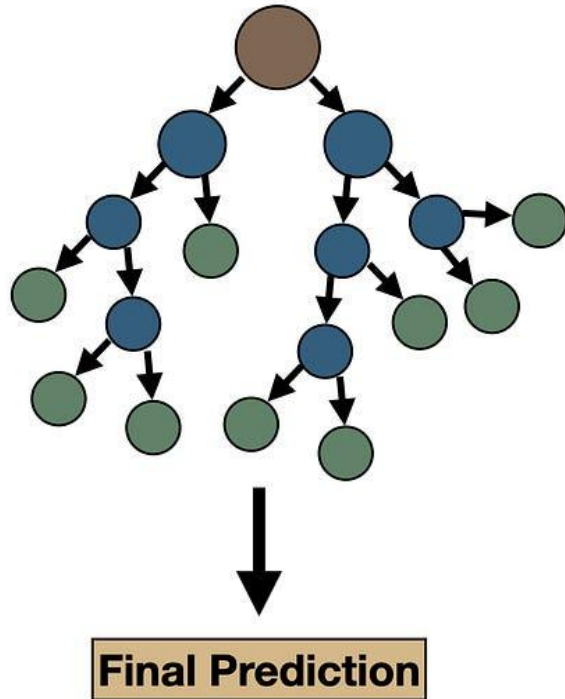
RANDOM FOREST

- Random Forest is an ensemble learning method suitable for both classification and regression tasks. It is known for its accuracy, particularly with large datasets.
- Random Forest constructs multiple decision trees during training and outputs the mode of the classes or mean prediction of individual trees.
- It mitigates variance and improves the performance of decision trees.

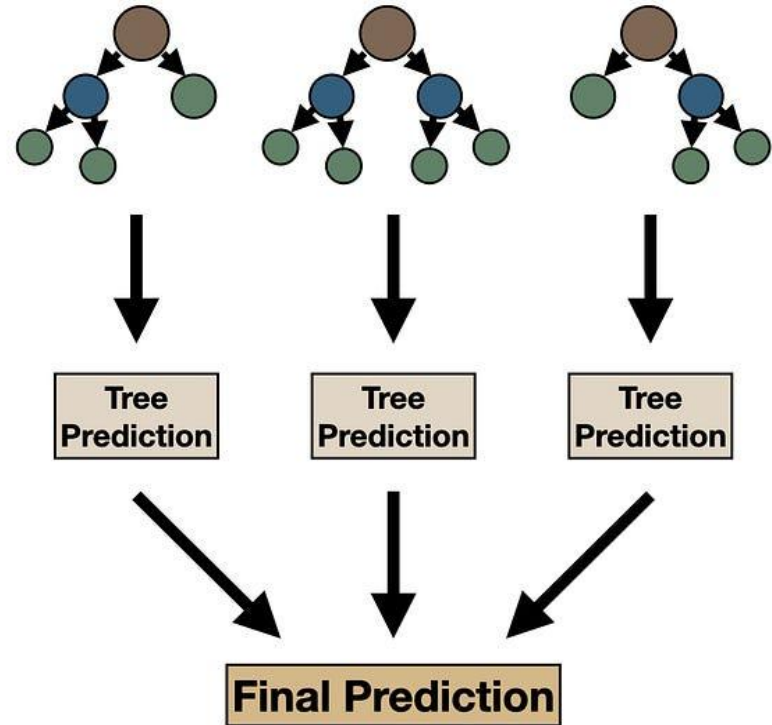
WORKING:

1. Create Bootstrap Dataset from Original data by Randomly choosing data(repetition is allowed).
2. Create Randomized Decision Tree from Bootstrap Dataset.
3. Finally output of the random forest is the class selected by most trees.

Single Decision Tree



Decision Tree Ensemble

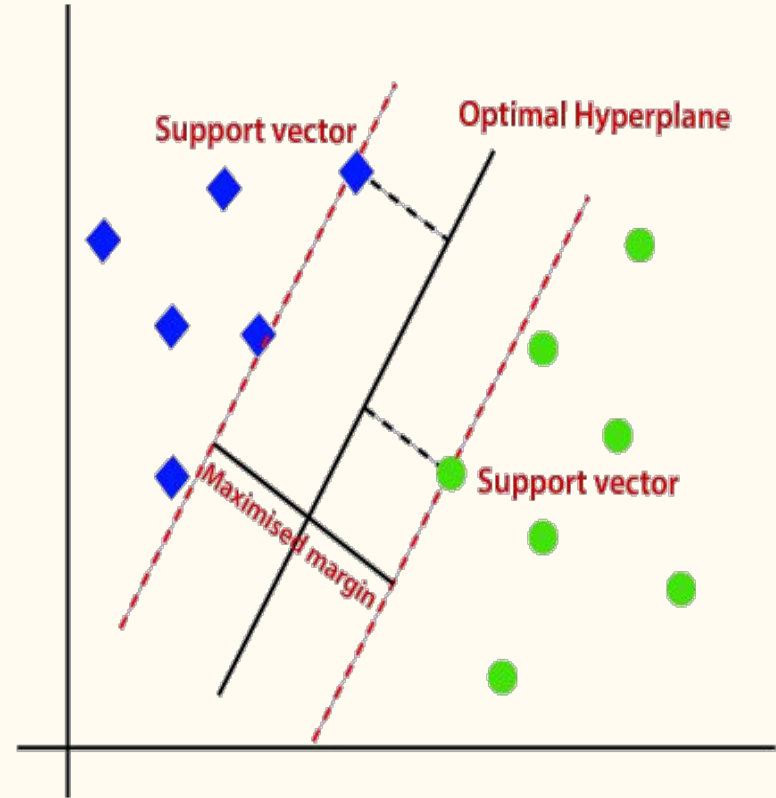


SUPPORT VECTOR MACHINE

- SVM is a renowned supervised machine learning algorithm used extensively in classification tasks.
- It is known for creating hyperplanes that effectively separate two classes.
- SVM operates efficiently in high-dimensional spaces and can classify instances not explicitly represented in the data.
- The separation is achieved by identifying a hyperplane that maximizes the margin between classes, minimizing the chance of misclassification.

WORKING

1. Selection of the hyperplane for optimal class separation.
2. Calculation of margin, representing the distance between classes.
3. Selection of the class with the highest margin, where
$$\text{margin} = \text{distance to positive point} + \text{distance to negative point}.$$

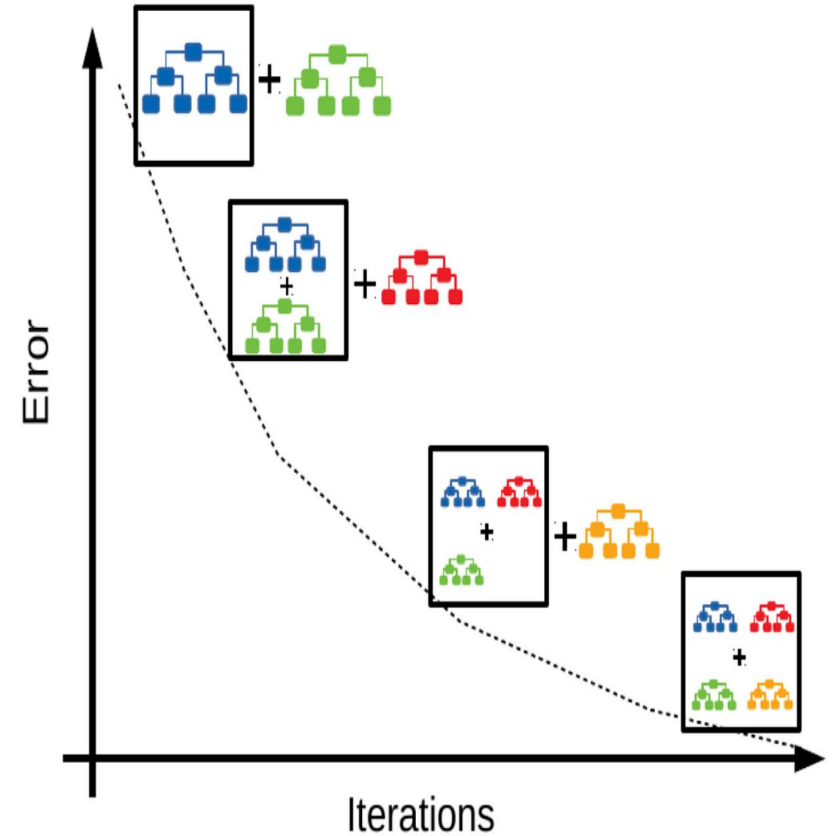


GRADIENT BOOST

- Gradient Boosting is a powerful ensemble learning technique widely used for both regression and classification tasks.
- It operates by building multiple decision trees sequentially, with each tree attempting to correct the errors of its predecessor.
- Gradient Boosting is particularly effective in capturing complex patterns in data and handling both numerical and categorical features. However, it may be prone to overfitting if not properly tuned.

WORKING:

1. The model starts with a simple decision tree, often called the base learner.
2. Subsequent trees are trained to correct the errors of the previous ones. Each new tree focuses on the residual errors of the ensemble, gradually improving the overall prediction.
3. The predictions of all trees are combined to make the final prediction. This combination is typically done by summing up the predictions of all trees.

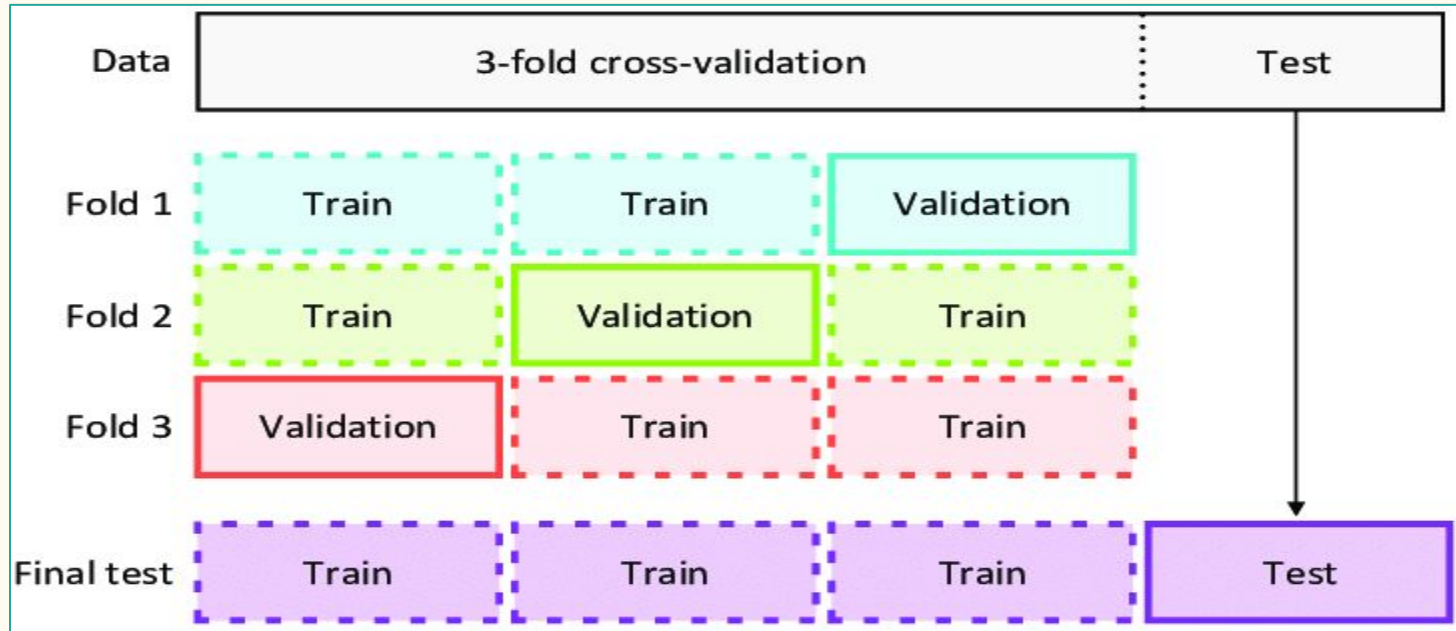


CROSS VALIDATION

After hyperparameter tuning, the model's performance is evaluated using cross-validation, a technique that splits the training data into multiple subsets for assessment of generalization.

1. Cross-validation, typically with k-fold ($k=10$) division, is applied to the resampled training data.
2. Each fold of the training data is used to evaluate the model, calculating performance metrics like accuracy, F1 score, precision, and recall.
3. The average performance metrics across all folds provide an overall assessment of the model's performance.
4. Standard deviation of performance metrics measures performance variability across folds.

- Performing cross-validation ensures a more robust estimate of the model's performance and its generalization ability, ensuring reliability and consistency across different training data subsets.



4. PERFORMANCE ANALYSIS

Evaluating models helps to determine the accuracy and effectiveness of the model in making predictions. Here are the evaluation metrics we have used.

CONFUSION MATRIX: A confusion matrix is a performance evaluation tool used in machine learning and statistics to assess the accuracy of a classification model. It is particularly useful when dealing with supervised learning problems where the data has predefined labels or classes.

- **True Positives (TP):** These are the cases where the model predicted a positive class, and the actual label was also positive.
- **True Negatives (TN):** These are the cases where the model predicted a negative class, and the actual label was also negative.

CONFUSION MATRIX

		Actual Values	
		Positive	Negative
Predicted Values	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

- **False Positives (FP):** These are the cases where the model predicted a positive class, but the actual label was negative (a type I error).

- **False Negatives (FN):** These are the cases where the model predicted a negative class, but the actual label was positive (a type II error).

ACCURACY: The proportion of correctly classified instances over the total number of instances

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

PRECISION: It is implied as the measure of correctly identified positive case (Tp) from all the predictive positive cases (TP + FP).

$$Precision = TP / (TP + FP)$$

RECALL: It is implied as the measure of correctly identified positive case(Tp) from all the actual positive cases.

$$\textit{Recall} = TP/(TP + FN)$$

F1 SCORE: It is the harmonic mean of precision and recall. This takes the contribution of both, so higher the F1 score, the better.

- So, a model does well in F1 score if the positive predicted are actually positives (precision) and doesn't miss out on positives and predicts them negative (recall).

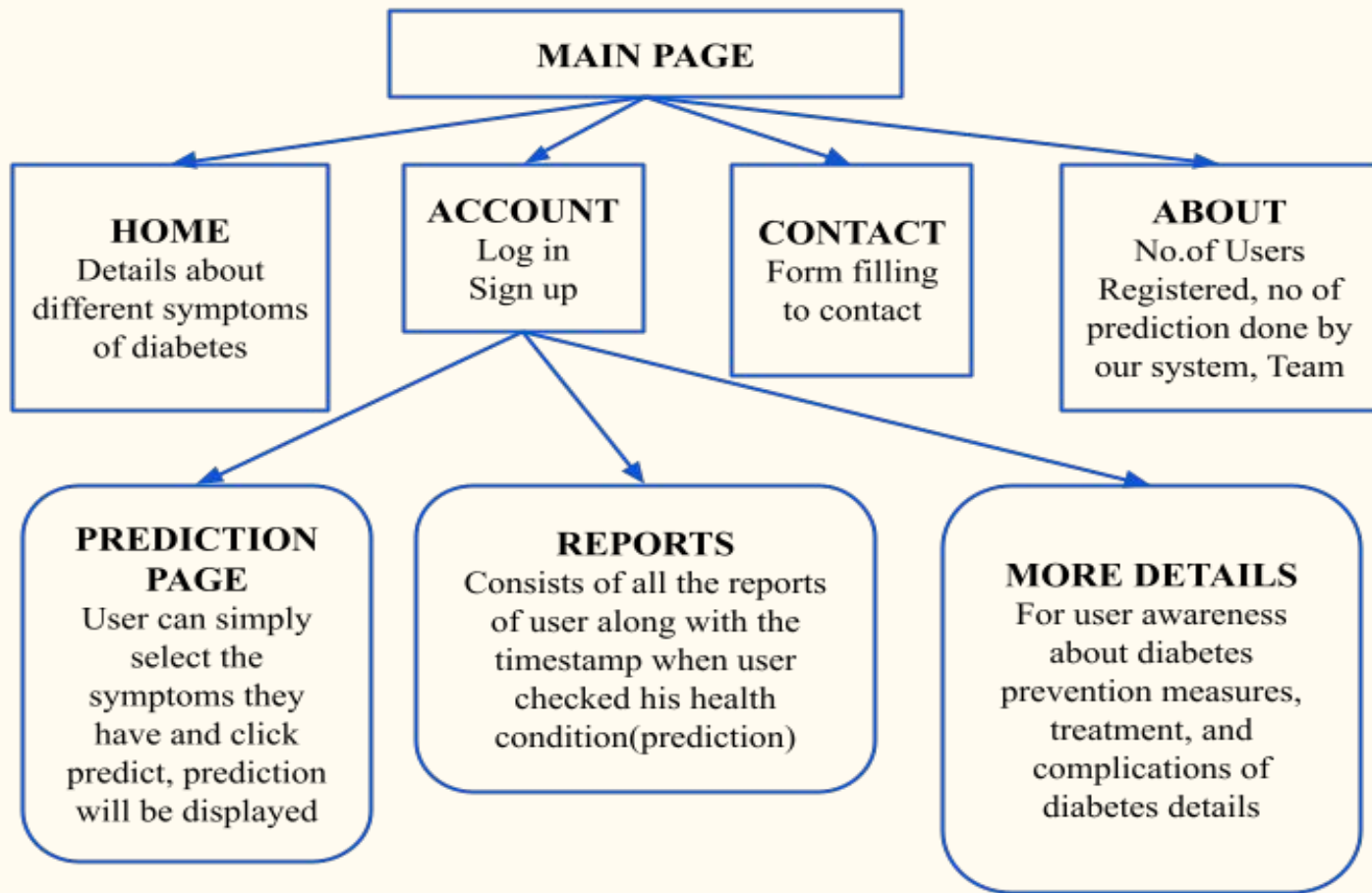
$$F1 = 2 \times (\textit{precision} + \textit{recall}) / (\textit{precision} \times \textit{recall})$$

	CV ACCURACY	PRECISION	RECALL	F1 SCORE
DECISION TREE CLASSIFIER	0.972569	0.979798	0.950980	0.965174
LOGISTIC REGRESSION	0.915328	0.960000	0.941176	0.950495
GRADIENT BOOST	0.970296	0.990099	0.980392	0.985222
RANDOM FOREST	0.972569	1.000000	0.990196	0.995074
SUPPORT VECTOR MACHINE	0.970402	0.990196	0.990196	0.990196

PERFORMANCE ANALYSIS

5. LOADING BEST MODEL AND DEPLOYMENT

- The Random Forest algorithm, identified as the top-performing model, is selected for deployment due to its superior accuracy in predicting diabetes risk.
- To ensure optimal performance in real-world scenarios, the model is trained on preprocessed data and serialized using a pickle file for efficient loading.
- Integrated the pickle file into the Streamlite web application enables users to input their medical parameters and receive predictions on early-stage diabetes risk.
- Utilized MySQL for database management ensures efficient user authentication and storage of health report information through which users can consult for treatment and doctor can also know the past trends of the symptoms in the user and thus helpful for treatment.



DIABETES PREDICTION WEBSITE FLOW

RESULTS

PREDICTION PAGE

CHECK YOUR HEALTH

What is your Age?

30

- +

What is your Gender?

☐ Male

☒ Female

SELECT SYMPTOMS YOU HAVE

☐ Do you urinate frequently?

☒ Do you feel unusually thirsty?

☐ Have you experienced sudden weight loss?

☒ Do you feel weak or fatigued?

☐ Do you have an increased appetite?

☐ Have you had recurring genital infections?

☐ Do you experience blurry vision?

☐ Do you suffer from itching, especially in the genital area?

☐ Are you irritable or experiencing mood changes?

☐ Do you notice that your wounds take longer to heal?

☒ Do you have numbness, tingling, or pain in your extremities?

☐ Do you experience muscle stiffness?

☐ Have you noticed hair loss?

☒ Are you overweight or obese?

Predict

DIABETES NEGATIVE CASE

Deploy

30 - +

What is your Gender?

☐ Male

☒ Female

SELECT SYMPTOMS YOU HAVE

☐ Do you urinate frequently?

☐ Do you feel unusually thirsty?

☐ Have you experienced sudden weight loss?

☐ Do you feel weak or fatigued?

☐ Do you have an increased appetite?

☐ Have you had recurring genital infections?

☐ Do you experience blurry vision?

☐ Do you suffer from itching, especially in the genital area?

☐ Are you irritable or experiencing mood changes?

☐ Do you notice that your wounds take longer to heal?

☐ Do you have numbness or pain in your extremities?

☐ Do you experience frequent stress?

☐ Have you noticed changes in your vision?

☒ Are you overweight or obese?

Predict

Hurray! You don't have Diabetes. Please consult with a Doctor for confirmation.

DIABETES POSITIVE CASE

What is your Age?

30

- +

What is your Gender?

☐ Male

☒ Female

SELECT SYMPTOMS YOU HAVE

☐ Do you urinate frequently?

☒ Do you feel unusually thirsty?

☐ Have you experienced sudden weight loss?

☒ Do you feel weak or fatigued?

☐ Do you have an increased appetite?

☐ Have you had recurring genital infections?

☐ Do you experience blurry vision?

☐ Do you suffer from itching, especially in the genital area?

☐ Are you irritable or experiencing mood changes?

☐ Do you notice that your wounds take longer to heal?

☒ Do you have numbness, tingling, or pain in your extremities?

☐ Do you experience muscle stiffness?


☐ Have you noticed hair loss?

☒ Are you overweight or obese?

Predict

You might have Diabetes. Please consult with a Doctor.

USER REPORTS - If user wants to Download only Particular Reports he/she can select those report ID's. They can just click on Generate to download all reports.

 MY HEALTH

MY REPORTS

MORE DETAILS

Generate Your Reports - Click On Generate

	ID	created_at	patient_id	age	gender	Polyuria	Polydipsia	Sudden_Weight_Loss	Weakness	Polyphagia	Genital_Thrush	Visual_Blurring	Itching	Irritability	Delayed_Healing	
6	7	2024-04-21 14:55:16	1	26	Male	No	No	No	No	No	No	Yes	No	No	No	
7	8	2024-04-23 13:21:24	1	30	Female	No	Yes	No	Yes	No	No	No	No	No	No	
8	9	2024-04-23 13:23:37	1	30	Female	No	No	No	No	No	No	No	No	No	No	

Enter Your Name

SMRITI MANDANA

Enter Your Age

30

Select Your Gender

☐ Male

☒ Female

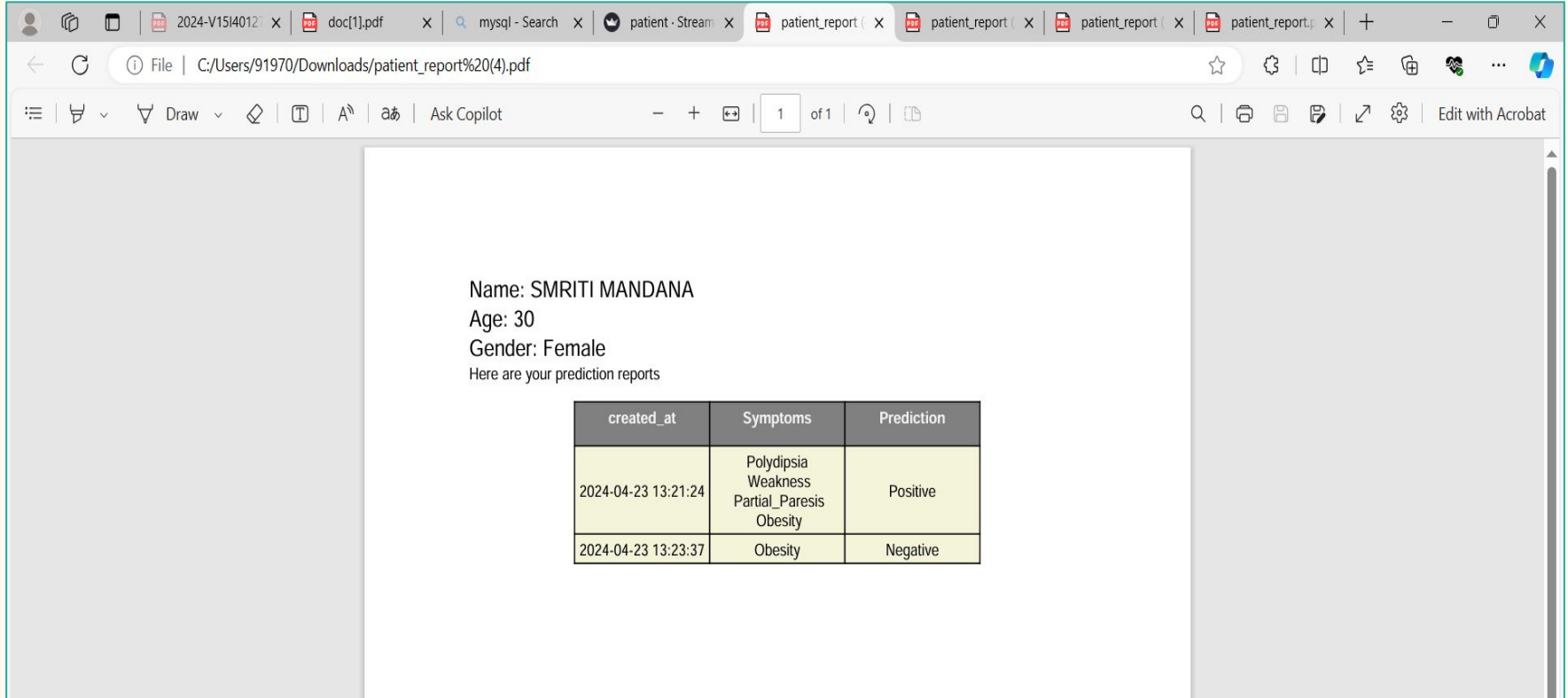
Select IDs for Report Generation (Leave empty to generate reports for all IDs)

8 x 9 x

GENERATE

Download PDF

GENERATED PDF OF SELECTED REPORTS



LIMITATIONS

Though there are many advantages with the proposed system there are also some limitations like.

- **Lack of access to technology:** Individuals who do not have access to technology or who are not comfortable using it may not be able to benefit from the system.
- **Limited availability of data:** The accuracy of the prediction model is highly dependent on the quality and quantity of data available. As there is limited data available the accuracy of the model may suffer.

Addressing these limitations is crucial for ensuring the system's accessibility and predictive accuracy.

CONCLUSION

In conclusion, the Early-stage Diabetic Prediction System leverages machine learning to forecast diabetes risk based on symptoms, aiming to enhance early detection and intervention. Accompanied by a user-friendly platform offering information on diabetes, prevention, and treatment, our system empowers individuals to make informed health decisions. By bridging the gap between technology and healthcare, we strive to mitigate the risk of diabetes complications and promote proactive health management.

FUTURE ENHANCEMENTS

- Explore advanced machine learning techniques like deep learning to enhance model performance.
- Expand the dataset through data augmentation or external sources for more meaningful insights.
- Consider ethical implications and collaborate with stakeholders for responsible advancements.
- Integrate domain-specific knowledge for improved model interpretability and relevance.
- Investigate novel data preprocessing techniques to improve model generalization and adoption in healthcare settings.

LINKS

PROJECT GITHUB LINK

<https://github.com/SHAIK-AFSANA/diabetespredictor>

CONTAINS DEPLOYABLE CODE

WEB APP LINK

<https://mldiabetespredictor.streamlit.app/>

THANK YOU