



# DIABETES PREDICTION USING MACHINE LEARNING MODELS

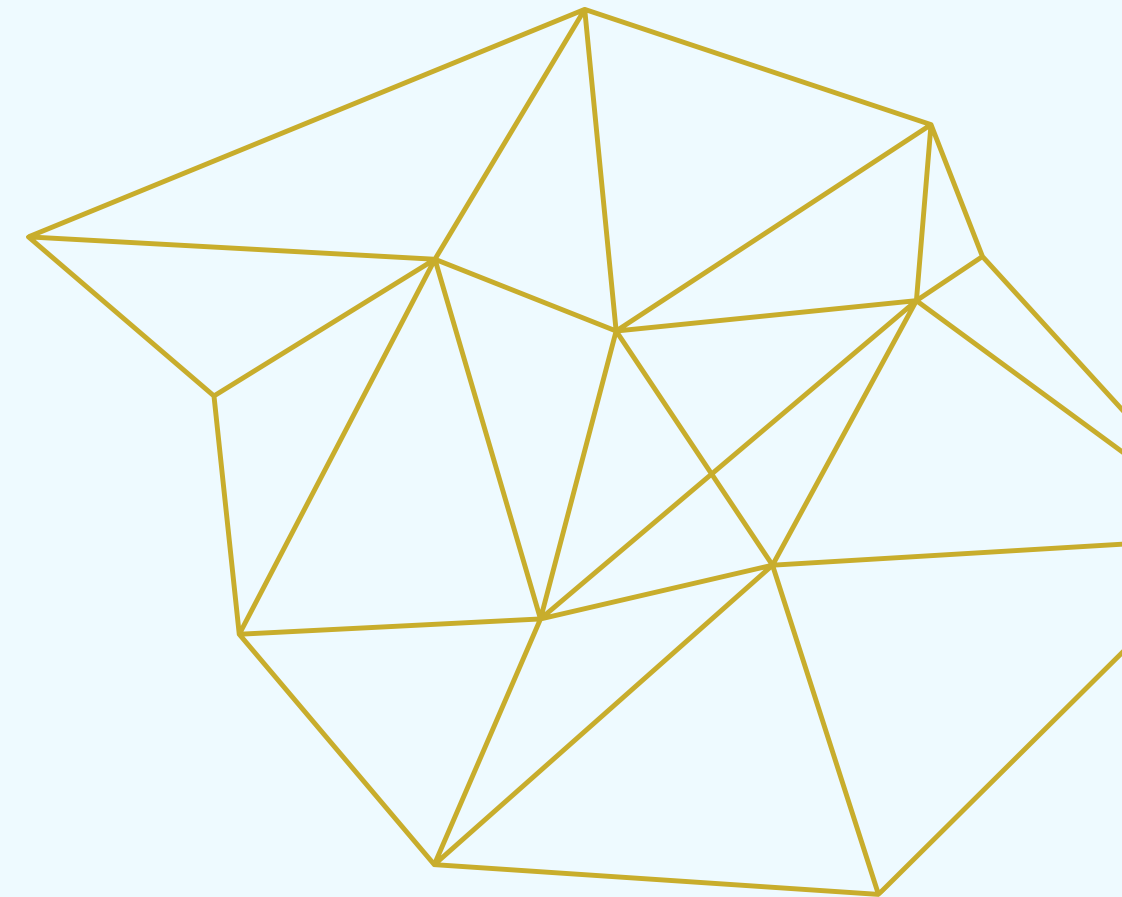
**Presented by Team Subhash and Co**

Varada Mahanth Naidu

Nenavath Preetham

Faiz Hashim

Doddy Subhash





# OVERVIEW

1

Project Abstract

2

Data Preprocessing

3

Exploratory Data Analysis

4

Models Training and Evaluation

5

Conclusions



# Abstract

---

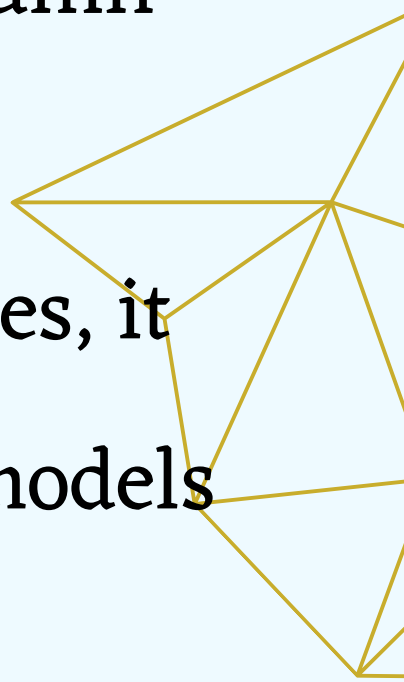
Diabetes Mellitus (**DM**) presents a growing global health concern, demanding effective preventive strategies. This project explores the application of machine learning algorithms/models to predict the onset of diabetes using the famous **PIMA INDIAN DATASET**.

Feature engineering techniques are employed to preprocess the given dataset and followed by a comparative analysis of various ML models, such as Smote-based LSTM, Logistic Regression, Random Forest, Support Vector Machines, Linear Discriminant Analysis, Quadratic Discriminant Analysis etc. Finally, we compare the metrics such as accuracy, precision etc for the models and try to find the best model which fits for diabetes prediction.



# PREPROCESSING



- Since there may be some missing values in the given dataset, it's always essential to preprocess the data ensuring that the dataset is more suitable for training.
  - In our dataset, certain data points contain attribute values represented as zeros, indicating **missing values**. Initially, these zero values were replaced with **NaN** (Not a Number). Subsequently, the medians were computed for columns with inconsistent or missing data and then these missing values were replaced with the respective column medians to make the data consistent.
  - By replacing the missing values with **median values** specific to the outcome classes, it tries to minimise the impact of missing data on subsequent analyses or predictive models built from this dataset.
- 

# Exploratory Data Analysis

---

- Exploratory Data Analysis (**EDA**) is an approach of analysing the dataset to summarise their main characteristics( here we use it for **Outlier Detection**) .
- **Outlier** is an observation that is numerically distant from the rest of the data (or) in simple words it is the value which is out of the range.
- Using EDA, we can see that some features like Insulin, BloodPressure, SkinThickness, BMI, DiabetesPedigreeFunction contain outliers in our dataset.
- We used **IQR(Inter Quartile Range)** method and **robust Z score** for outlier detection and removal of these outliers.

# Inter Quartile Range (IQR) Method


- For each feature in the dataset, calculate the first quartile (**Q1**), third quartile (**Q3**), and the Interquartile Range (**IQR**) using the formulas :  
$$\mathbf{Q1 = 25th\ percentile, Q3 = 75th\ percentile, IQR = Q3 - Q1}$$
- We define a range for identifying outliers: typically, values below  $\mathbf{Q1 - 1.5 * IQR}$  or above  $\mathbf{Q3 + 1.5 * IQR}$  are considered outliers.
- Generate a boolean mask (outliers) for each feature, marking rows that contain outliers based on the defined range and use the boolean mask to filter the dataset, removing rows identified as outliers for all features.





# Robust Z\_score Method

---

- For each data point in a feature column, compute the **Z-score** using the formula:  
$$\mathbf{Z} = (\mathbf{X} - \mu) / \sigma$$
 where  $X$  is the data point,  $\mu$  is the mean of the feature, and  $\sigma$  is the standard deviation of the feature.
  - Define a threshold value (often, an absolute Z-score of 3 or more) to determine outliers. Data points with Z-scores beyond this threshold are considered outliers.
  - Generate a boolean mask marking rows that contain outliers based on the defined threshold for each feature and use the boolean mask to filter the dataset, removing rows identified as outliers based on their Z-scores.
- 

## AdaBoost Classifier :-

- Base Model Accuracy for this model without EDA is around **84.41%** .
- Model Accuracies after data pre-processing and hyper-parameters tuning using **gridsearch** are **87.66%**(using IQR data), **86.36%**(using Z\_score data) .
- Since this model is a weak learner classifier there is an increase in the accuracies and also the increase is mainly due to data pre-processing and EDA.

## XGBoost Classifier :-

- Base Model Accuracy for this model without EDA is around **87.01%** .
- Model Accuracies after data pre-processing and hyper-parameters tuning using **gridsearch** are **83.11%** (using IQR data), **85.06%**(using Z\_score data).
- During hyper-parameter tuning the model complexity may have increased.  
So due to overfitting of the data the test accuracies got decreased.



# Linear Discriminant Analysis (LDA) :-

- **LDA** involves modeling the distribution of features for individuals with and without diabetes separately, using this information to predict if new individuals have diabetes or not. By analyzing how the features vary among the two classes, LDA determines boundaries that best separate the classes, enabling accurate classification of new instances based on their features.
- Base Model Accuracy for this model without EDA is around **78.57%** .
- Model Accuracies after data pre-processing and hyper-parameters tuning using **gridsearch** are **81.81%**(using IQR data), **77.27%**(using Z\_score data) .

# Quadratic Discriminant Analysis(QDA) :-

- **QDA** is similar to LDA but the difference is that this does not assume equal co-variance among the classes. This method models class-specific covariance matrices, allowing for more flexible decision boundaries compared to LDA, potentially leading to improved performance when classes have **different covariance** structures. Adjusting parameters or refining features can optimize QDA's performance for diabetes prediction.
- Base Model Accuracy for this model without EDA is around **85.06%**.
- Model Accuracies after data pre-processing and hyper-parameters tuning using **gridsearch** are **84.41%**(using IQR data), **85.06%**(using Z\_score data) .

## Logistic Classifier :-

- Base Model Accuracy for this model without EDA is around **77.27%** .
- Model Accuracies after data pre-processing and hyper-parameters tuning using **gridsearch** are **81.81%**(using IQR data), **77.92%**(using Z\_score data) .
- By removing outliers we can see the increase in IQR data and for Z\_Score feature scaling may not be critical as we have almost similar values

## SVM Classifier :-

- Base Model Accuracy for this model without EDA is around **83.76%** .
- Model Accuracies after data pre-processing and hyper-parameters tuning using **gridsearch** are **81.16%** (using IQR data), **82.46%**(using Z\_score data).
- We can observe decrease in both because removal of outliers might change the decision boundaries learned and original scales were not a major issue



# SMOTE Based LSTM - I

---

- **SMOTE**, which stands for Synthetic Minority Over-sampling Technique, is a method used to address class imbalance in datasets, particularly in machine learning tasks where one class is significantly underrepresented compared to others. When combined with LSTM networks, SMOTE can be used to enhance the performance of the LSTM model when dealing with imbalanced datasets.
- Assuming that the data set's sampling magnification is  $N$ , there are  $N$  samples from  $K$ -NN and  $K > N$ . The interpolation formula of SMOTE is

$$S_i = X + rand(0,1) \times (y_i - X)$$

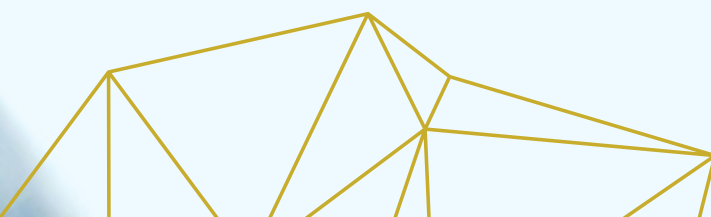
# SMOTE Based LSTM - II

---

- Initially, the data outliers were removed using EDA, then the normalization of the data set is performed by using **min-max transformation** method.
- Then we use the **SMOTE** method which handles the imbalanced data and turns this into balanced data. Then data is passed through the **LSTM** .
- Next the balanced data set will be reshaped into 3D data. **LSTM** accepts input data to be a 3D tensor such as batch size, timesteps and features.
- Base Model Accuracy for this model without EDA is around **82.47%** .
- Model Accuracies after data pre-processing and hyper-parameters tuning are **87.50%**(using IQR data), **87.96%**(using Z\_score data) .



# Random Forest Classifier :-

- This is an **Ensemble learning** method using multiple decision trees.
  - Base Model Accuracy for this model without EDA is around **88.31%** .
  - Tuning numerous parameters in this model can lead to over-fitting so carefully tuned only specific parameters using **Gridsearch** to balance model complexity .
  - Those specific parameters were n\_estimators, max\_depth, min\_samples\_leaf, min\_samples\_split and used **best\_params\_** from these for the test prediction.
  - Model Accuracies after data pre-processing and hyper-parameters tuning using **gridsearch** are **87.01%**(using IQR data), **87.66%**(using Z\_score data) .
- 



## K-Nearest Neighbours Classifier :-

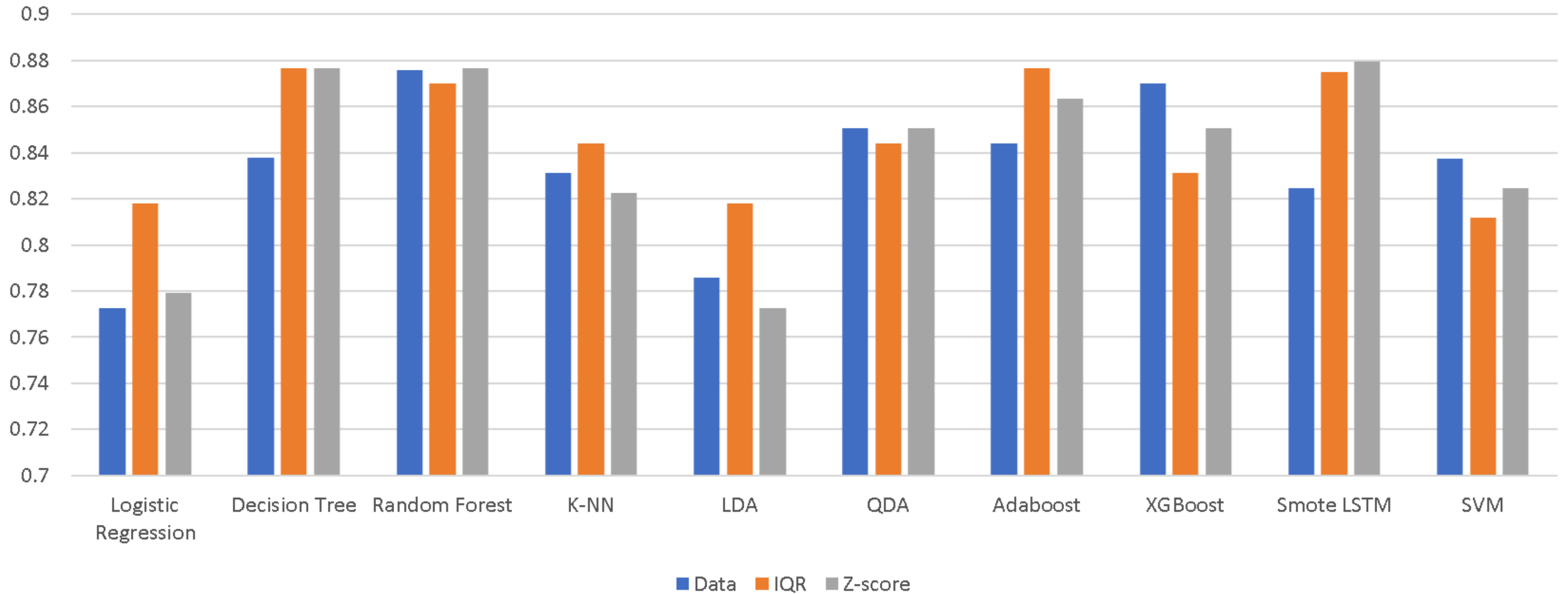
- Base Model Accuracy for this model without EDA is around **83.11%** .
- Model Accuracies after data pre-processing and hyper-parameters tuning using **gridsearch** are **84.41%**(using IQR data), **82.26%**(using Z\_score data) .
- During outlier detection using Z\_score significant data may be lost which results in lower accuracy when compared to that of base model.

## Decision Tree Classifier :-

- Base Model Accuracy for this model without EDA is around **83.76%** .
- Model Accuracies after data pre-processing and hyper-parameters tuning using **gridsearch** are **85.06%**(using IQR data), **87.66%**(using Z\_score data) .
- We can see a considerable increase in the accuracies due to EDA since this model is a weak learner .

# Conclusions


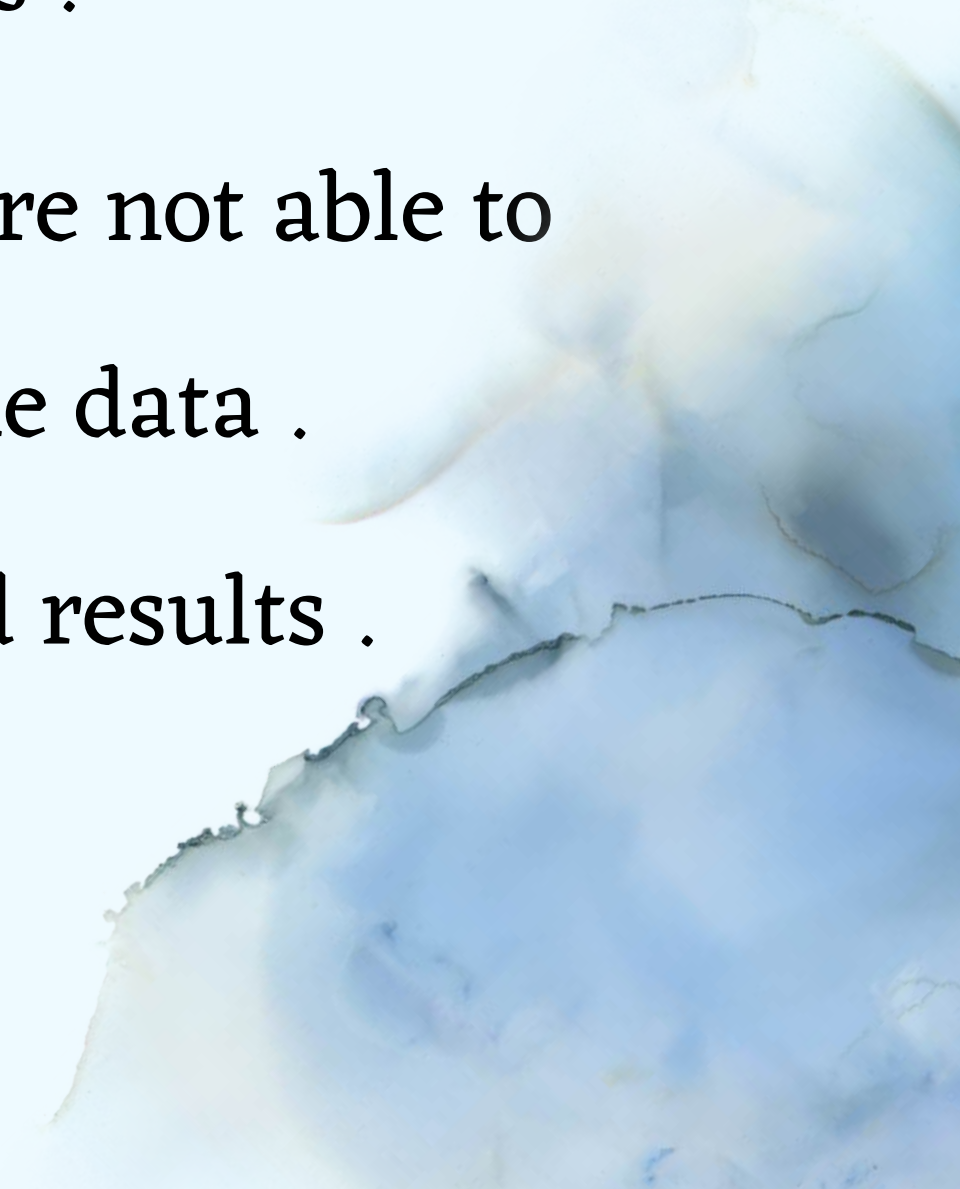
Comparison of Accuracies of All Models



# Conclusions



---

- After comparing accuracies of the all the models, we could see that the best models for our diabetes prediction task are Random Forest Classifiers and SMOTE-based LSTM Classifiers .
  - We could generalise that the normal linear models are not able to discover the dependencies of features within the data .
  - Weak Classifiers trained on processed data gives good results .
- 
- 



# References and Dataset Used :-

- <https://thesai.org/Publications/ViewPaper?Volume=12&Issue=4&Code=IJACSA&SerialNo=66>
- <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

## Contributions :-

- Doddy Subhash : Logistic Classifier, SVM Classifier,  
Stacking
- Varada Mahanth Naidu : Smote based LSTM, Random Forest Classifier,  
Decision Tree Classifier, KNN Classifier
- Preetham Nenavath : Smote based LSTM, Adaboost Classifier,  
XGBoost Classifier, Data Pre-Processing
- Faiz Hashim : Data Pre-Processing, LDA Classifier,  
QDA Classifier, Exploratory Data Analysis



**THANK YOU**