## ABSTRACT:

The main objective of the project is to analyse the effectiveness and  compare the predictive abilities of different classification models in the context of Pima Indian Diabetes Dataset. By showing comparison of different performance metrics (e.g. accuracy, precision, recall, area under the ROC curve etc) for different classification models, the project provides a comprehensive overview of the effectiveness of these models while making predictions for a binary classification problem. The project also highlights the analysis of the feature variables to identify the variables that have a significant impact on the outcome variable to predict whether a person shows signs of diabetes or not. After evaluation of different models it is evident that two level ensemble approach (where Generalized Linear Model (GLM), Generalized additive model (GAM), Support Vector machine (SVM) model and Random Forest are trained individually in the first level and the outcome predictions from these models are provided to Neural Network model in second level for making the final prediction of the outcome) show higher accuracy (99%) on the test dataset compared to the other model approach.

## INTRODUCTION:

Diabetes is a long-term metabolic disease characterized by high blood glucose levels, because of insufficient insulin production or ineffective insulin utilization by the body [1]. The Pima Indian Diabetes is selected for this project because the dataset allows investigation of a wide range of demographic and health related characteristics regarding diabetes to obtain insightful information about the key features that impact the risk of diabetes. So, addressing the influence of these factors on the occurrence of diabetes may provide insights into effective approaches for the risk minimization of diabetes.

## DESCRIPTION OF DATASET:

The Pima Indians Diabetes dataset is derived from a study conducted by the National Institute of Diabetes and Digestive and Kidney Disease. The research investigated the health characteristics of 768 adult female Pima Indians living near Phoenix, with specific focus on the prevalence of diabetes and associated factors. The main objective of this dataset is to detect and analyze the key risk parameters associated with diabetes to gain insights into the critical factors impacting the diabetes occurrence among the female Pima Indians.

The dataset contains eight feature predictors to predict the binary dependent variable which indicates the presence or absence of diabetes. The Pima Indians Diabetes dataset consists of the following features:

- **Pregnant:** Represents the number of how many times a female Pima Indian has been pregnant.
- **Glucose:** Indicates plasma glucose concentration at 2 hours in an oral glucose tolerance test. By measuring the amount of glucose, the variable provides information on blood sugar level and metabolic health.
- **Diastolic:** Represents diastolic blood pressure, which is measured in (mmHg). Blood pressure is a physiological measure that indicates the force generated by circulating blood.
- **Triceps:** Represents triceps skinfold thickness, measured in mm.
- **Insulin:** Indicates two-hours serum insulin level, measured in mu U/ml. It represents the concentration of insulin (a hormone) present in the blood two hours after a meal.
- **BMI:** Represents body mass index, which is calculated by dividing the mass of the body by the square of height of a person. So, BMI is determined by using the formula:

$$\text{BMI} = \frac{weight\ in\ kg}{(\text{height in meter})2}$$

- **Diabetes:** Represents diabetes pedigree function which provides the probability of diabetes occurrence of a person by considering the person’s family history of diabetes.
- **Age:** Represents age in years of female Pima Indians.
- **Test:** Represents the binary classification outcome variable to indicate whether a female has diabetes or not, where 1 indicates the presence of diabetes (positive) and 0 indicates the absence of diabetes (negative) in the participants.

In this project, the performance of different models (e.g.: Generalized Linear Models, Random Forest, Neural Networks, etc.) will be evaluated on the basis of its complexity level and compared to obtain insights into the effectiveness of the model to predict the outcome in context of a binary classification problem. The project includes the task of data pre-processing of Pima Indian dataset, analysis of different models based on various performance metrics and comparative assessment of the performance to find the most effective model in predicting whether the individual is diabetic or not.

## DATA-PREPROCESSING:

- **Handling unrealistic values:** Since some of the feature variables (e.g.: glucose, diastolic, triceps, insulin, BMI) logically can’t have 0 value, these 0 values are replaced with NA (Not Available).
- **Handling missing values:** The missing values in the dataset are imputed with the median value of the corresponding column.
- **Handling Outliers:** Box plots and summary statistics are used to investigate and detect whether there are any data points that are significantly unusual from the rest of the data points.

- **Evaluating Feature importance:** Correlation analysis by showing a heatmap is used for obtaining insights into the relationships between variables and for identifying the most important features for predicting the target variable.
- **Scaling features:** Scaling has been applied to numeric feature predictors of the dataset to standardize the values, which leads to effective model training and improved performance.
- **Splitting data in train and test set:** The test dataset is generated by taking a random sample of 100 cases and the remaining cases are assigned to the training dataset.

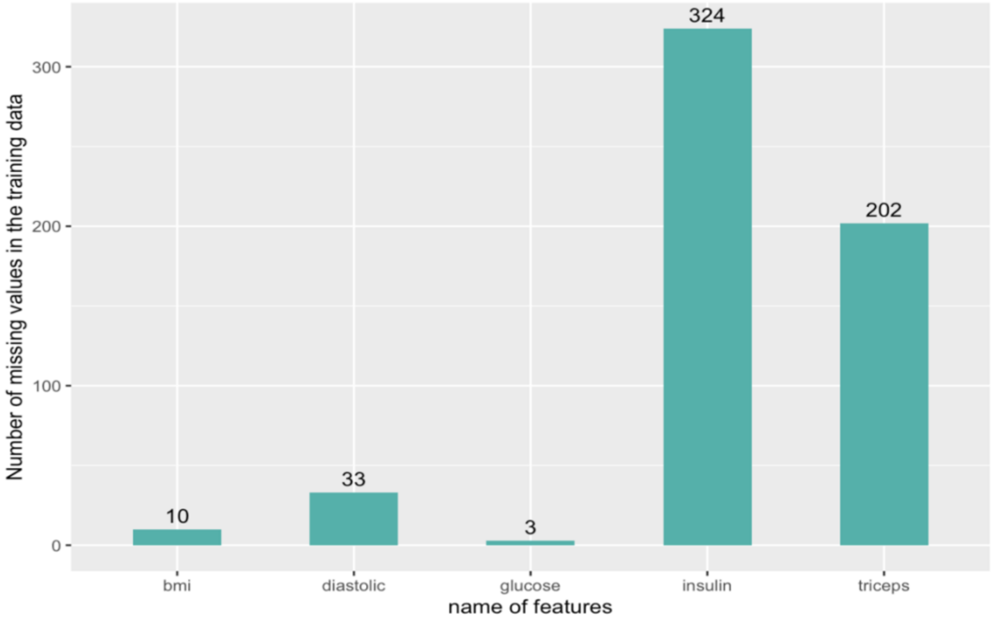

Figure 1.: Features with missing values

Out of 9 features, figure 1 represents 5 features which included BMI, diastolic, glucose, insulin and triceps which contains missing values where insulin was the one with the most frequent missing value, hence handling it by imputing the missing places with median helped deal with the problem to rectify the issue.

Heatmap in figure 2 displayed the intercorrelation among individual features to evaluate how positively or negatively the feature are correlated with each other along with the target variable which is diabetes and deliver valuable information to check how a feature is affecting another feature.

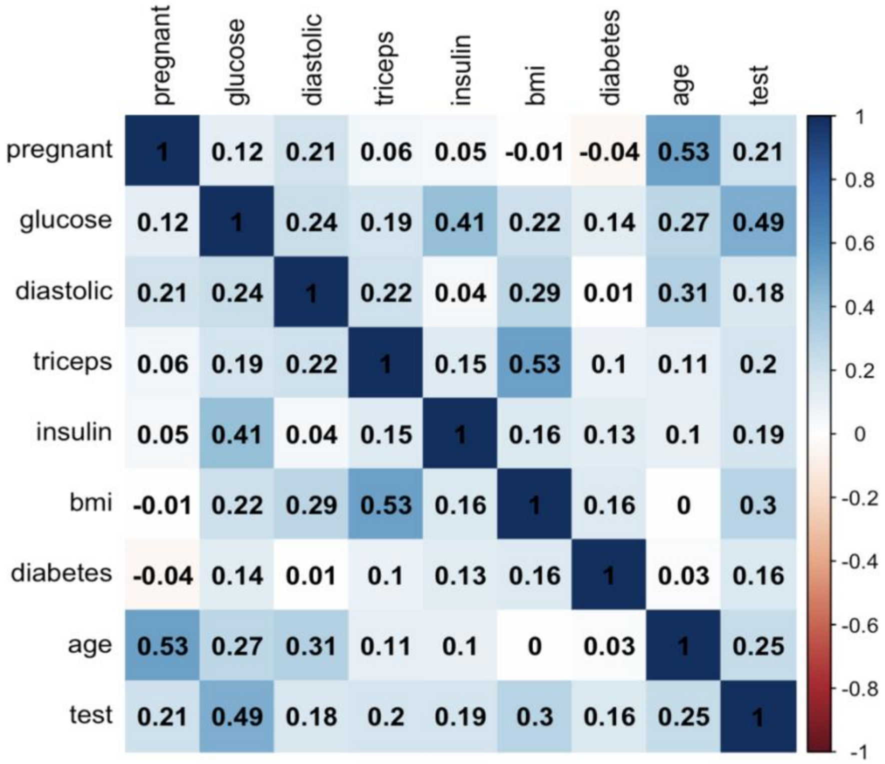

Figure 2: Correlation Analysis

## ANALYSIS OF DIFFERENT MODELS FOR PREDICTION:

- **Generalized Linear Model (GLM)**

  GLMs use logistic regression to predict the outcome of the binary classification problem. Initially, the entire training dataset has been used to train the Generalized Linear Model (GLM) and predict the outcome on the test dataset using a classification threshold of 0.5 to categorize outcomes into positive and negative labels. Then, in order to obtain an improved model performance and prevent overfitting, stepwise AIC variable selection method has been applied on the GLM to get the subset model containing the most relevant feature predictors. Following this, the reduced subset model has been used to predict the outcome and the model performance has been evaluated based on it.

- **Generalized Additive Model (GAM)**

By using smooth functions, a generalized additive model offers enhanced flexibility in modeling of non-linear patterns among variables in the dataset. A generalized additive model has been trained using all the available feature predictors in the training dataset and the model performance has been evaluated on the test dataset.

- **Trees and Random Forest**

  Initially a default tree model has been trained using all the feature variables and a visual representation plot of the tree structure is made. Then, cross-validation is used to select the optimal tree size.

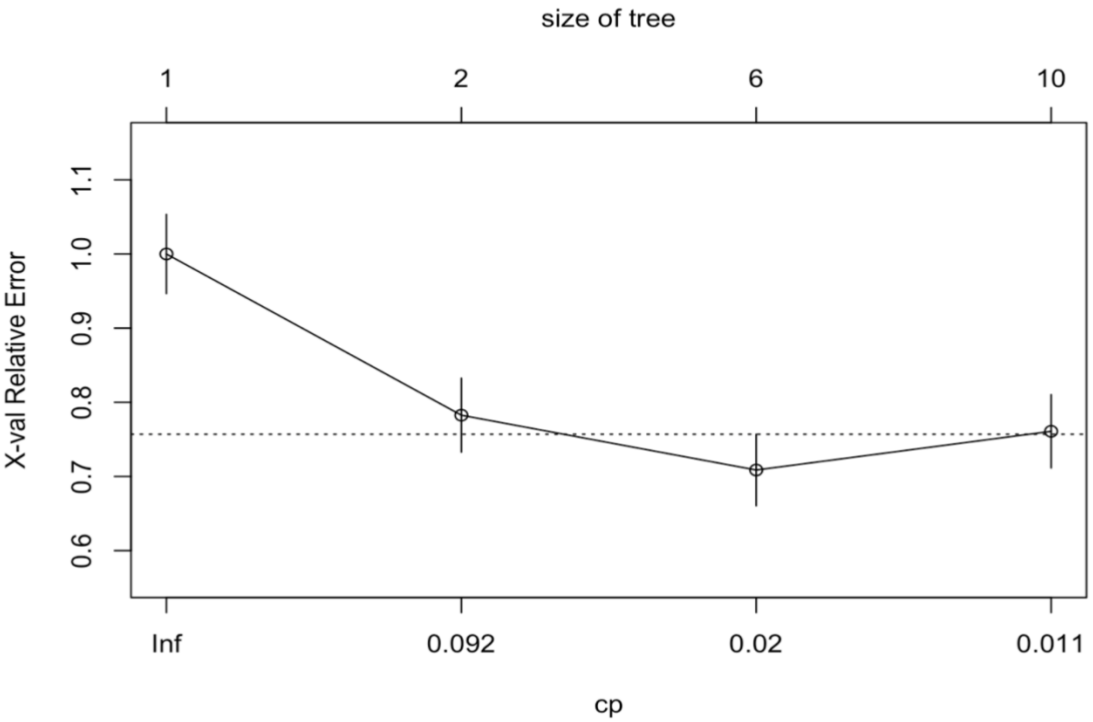

Figure 3: Decision Tree Complexity Parameter Analysis

By visualizing the impact of different complexity parameters, the complexity parameter associated with lowest value of cross-validated error is found, which indicates the optimal tree size. Then pruning (elimination of branches that have lower impact) is applied to get the final simplified tree and predictions are made on the test dataset.

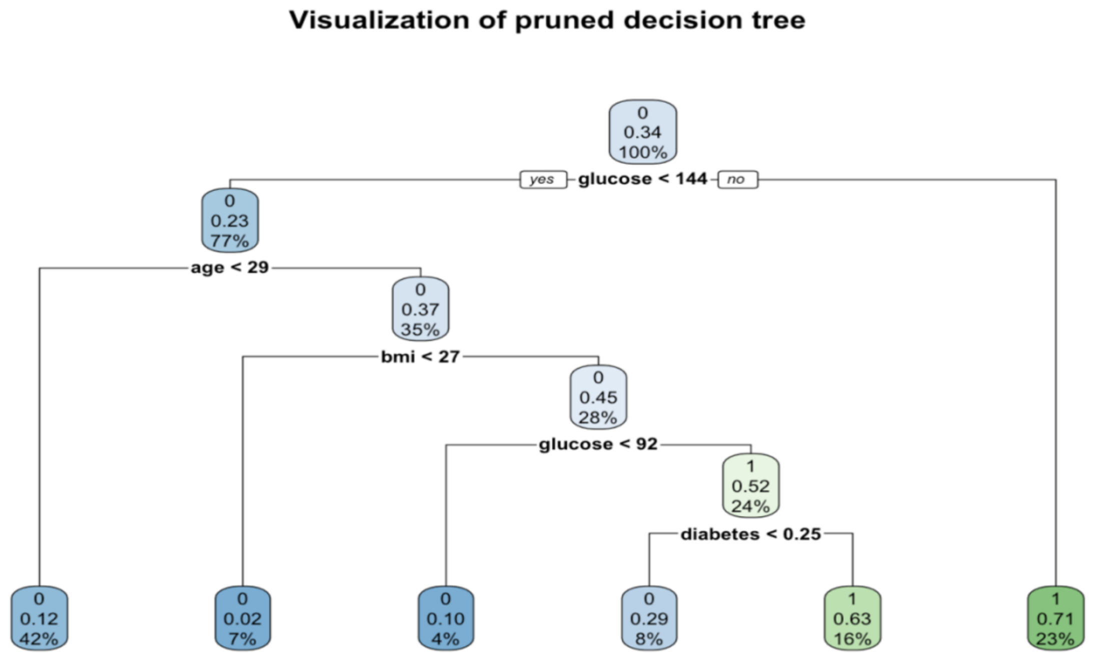

Figure 4: Pruned decision tree model with optimal complexity parameter

In case of a random forest model, the predictions of several decision trees are combined. A random forest model has also been trained to predict the outcomes of Pima Indians diabetes dataset. By visualizing the variable of importance using the random forest model, the most significant predictor can be identified.

- **Support Vector Machine Model (SVM)**

Support vector machine is a supervised learning model that determines an optimal line or hyperplane to effectively separate classes in the dataset. Two approaches, linear and radial SVM, are applied with 15-fold cross validation for model training and then predictions are made on the test dataset for these models. In the case of linear SVM, a hyperplane is used to differentiate two classes and in case of radial SVM, a flexible decision boundary is offered.

- **K-Nearest neighbor (KNN)**

  In the case of the K-nearest neighbor model, which is a supervised learning model, classification for a data point is predicted according to the majority class of its k-nearest neighbors. K-Nearest neighbor model is trained using the scaled train dataset and hyperparameter tuning has been applied to enhance model performance.

- **Neural Network**

  Neural network is a machine learning model consisting of interconnected artificial neurons.

  - **Unscaled Data:** First neural network model is trained using unscaled data and the architecture comprises of two hidden layers.
  - **Scaled Data:** Second Neural network model was trained with a scaled data and 10 hidden layers were present to assess the impact of the difference in scaling and complexity in model performance.

- **Two Level Ensemble Approach**

  In the first level (can refer Figure 5), GLM, GAM, SVM, KNN and random forest are trained independently resulting for the output predictions of these models which are provided as features to train the artificial neural network model in the second level. By predicting outcome on the test dataset and evaluating model performance, the effectiveness of this approach has been assessed.

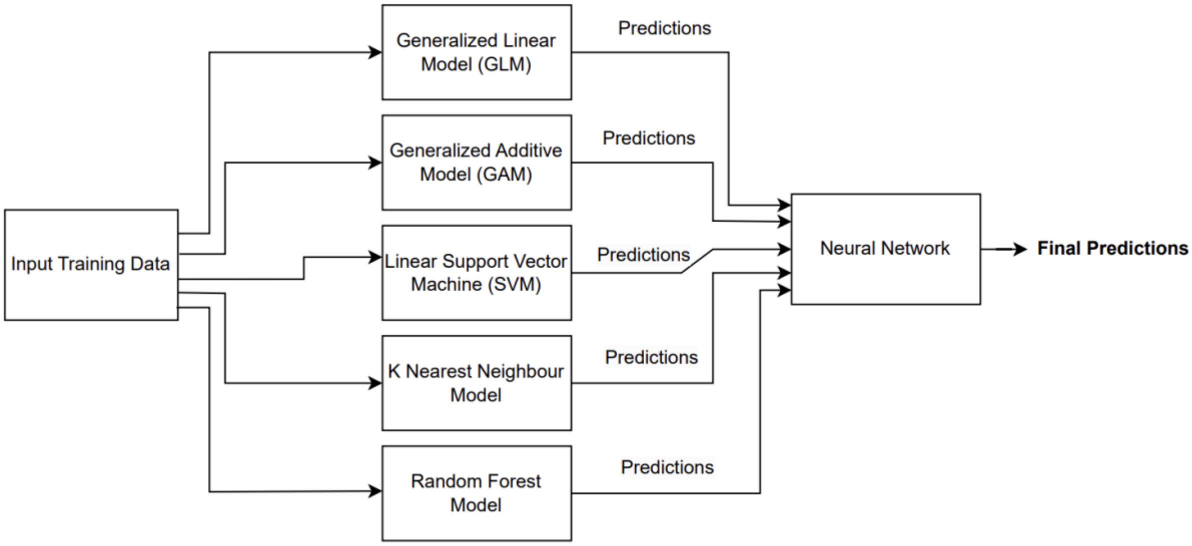

Figure 5: Architecture for two level ensemble Approach

- **Ensemble Approach of combining multiple model predictions**

  In this approach, initially GLM, GAM, SVM, KNN, random forest and neural network models are trained individually and predictions are made on the test dataset independently. Then, the predictions from these models are combined using a majority voting technique, with the most frequent prediction from the trained models is selected as final outcome.

## EVALUATION METRIC OF THE MODEL PERFORMANCE

- **Accuracy**

  Accuracy is a performance metric to measure how effective the model is in predicting the outcome correctly, which is determined by dividing the accurately predicted occurrences by the total number of instances.

- **Confusion Matrix**

  Confusion matrix provides an overview of the overall performance of the model by showing the count of True Negative (correctly predicted negative instances), True Positive (correctly predicted positive class instances), False Negative (incorrectly predicted as negative instances) and False Positive predictions (incorrectly predicted as positive instances) in a table.

- **False Positive Rate and False Negative Rate**

  When a model incorrectly predicts and classifies a negative class as a positive class, it is known as false positive prediction. False positive rate is calculated by dividing the count of false positive predictions by the total count of actual negative instances.

  Again, when a model incorrectly classifies a positive class as a negative class, it is known as false negative prediction. False negative rate is determined as the proportion of the number of false negative predictions to the total count of actual positive instances.

- **Precision and Recall**

  Precision is calculated by dividing the count of true positive predictions by the sum of true positive and false positive instances, which actually represents how accurate the model is when it classifies an instance as a positive class.

  Again, Recall is computed as the proportion of the number of true positive instances to the sum of true positive and false negative predictions, which represents how effective the model is in predicting actual positive class instances.

- **ROC Curve**

  The Receiver Operating Characteristic (ROC) curve provides a visual representation of the model’s capacity to differentiate between two classes by plotting the true positive rate against the false positive rate. To measure the overall performance of a model using ROC curve, the area under ROC curve (AUC) metric is used, where higher value of AUC represents a better model performance.

## RESULTS

**Comparative performance measures of different models**

| Method | Accuracy | False Positive Rate | False Negative Rate | Precision | Recall | ROC-AUC score |
|---|---|---|---|---|---|---|
| Generalized Linear Model (GLM) | 0.76 | 0.076 | 0.542 | 0.923 | 0.759 | 0.6901 |
| Reduced Generalized Linear Model (Subset model) | 0.79 | 0.0769 | 0.457 | 0.923 | 0.789 | 0.7329 |
| Generalized Additive Model (GAM) | 0.78 | 0.1077 | 0.4286 | 0.892 | 0.7945 | 0.7319 |
| Default Tree Model | 0.8 | 0.1846 | 0.2286 | 0.815 | 0.8689 | 0.7934 |
| Pruned Tree Model having optimal tree size | 0.8 | 0.1846 | 0.2286 | 0.815 | 0.8689 | 0.7934 |
| Random Forest (RF) | 0.98 | 0.0153 | 0.0286 | 0.9846 | 0.9846 | 0.978 |
| K- Nearest Neighbor Model (KNN) | 0.82 | 0.0769 | 0.3714 | 0.923 | 0.8219 | 0.7758 |
| Linear Support Vector Machine Model (SVM) | 0.77 | 0.0769 | 0.5142 | 0.923 | 0.769 | 0.7044 |
| Radial Support Vector Machine Model (SVM) | 0.83 | 0.046 | 0.4 | 0.9538 | 0.8158 | 0.7769 |
| Neural Network Model with unscaled Data | 0.62 | 0.323 | 0.4857 | 0.6769 | 0.7213 | 0.5956 |
| Neural Network Model with scaled Data | 0.83 | 0.2154 | 0.0857 | 0.7846 | 0.944 | 0.8495 |
| Two level Ensemble Approach<br>(1st Level: GLM+GAM+SVM+KNN +RF,<br>2nd Level: ANN) | 0.99 | 0 | 0.0286 | 1 | 0.9848 | 0.9857 |
| Ensemble Approach of combining multiple models (Majority Voting) | 0.82 | 0.0769 | 0.3714 | 0.923 | 0.821 | 0.7758 |

## ANALYSIS OF THE MODEL PERFORMANCES

By observing the performance metric on the table, it is evident that the random forest model and two-level ensemble approach displayed higher accuracy compared to the other model approach. As an ensemble learning method, random forest combined several decision trees to minimize overfitting which results in better prediction on the test dataset. Regarding the two-level ensemble approach, since the outcomes of different models are provided as feature predictors to the neural network model to predict the target variable, the neural network can effectively learn from these in the second level of training, which lead to improved overall model performance and accuracy.

Also, In the case of neural network, there is a significant difference in accuracy and model performance between unscaled and scaled data-based training. So, feature scaling has a significant influence on the accuracy of the neural network model since it guarantees balanced contribution of each numeric feature predictors by standardizing the values, which leads to improved model performance.

The reduced GLM subset model has higher accuracy than the GLM model using all the feature predictors implies that more effective predictions can be achieved by training the model with the most relevant features supporting the requirement of less data and better result. The reason behind the better performance of the subset model is that it prevents overfitting and eliminates the influence of unimportant variables by considering only the features that have significant impact on the outcome.

Regarding the Support Vector Machine model, the radial SVM model exhibits higher accuracy compared to the linear SVM because radial SVM offers a flexible decision boundary, which results in effective classification.

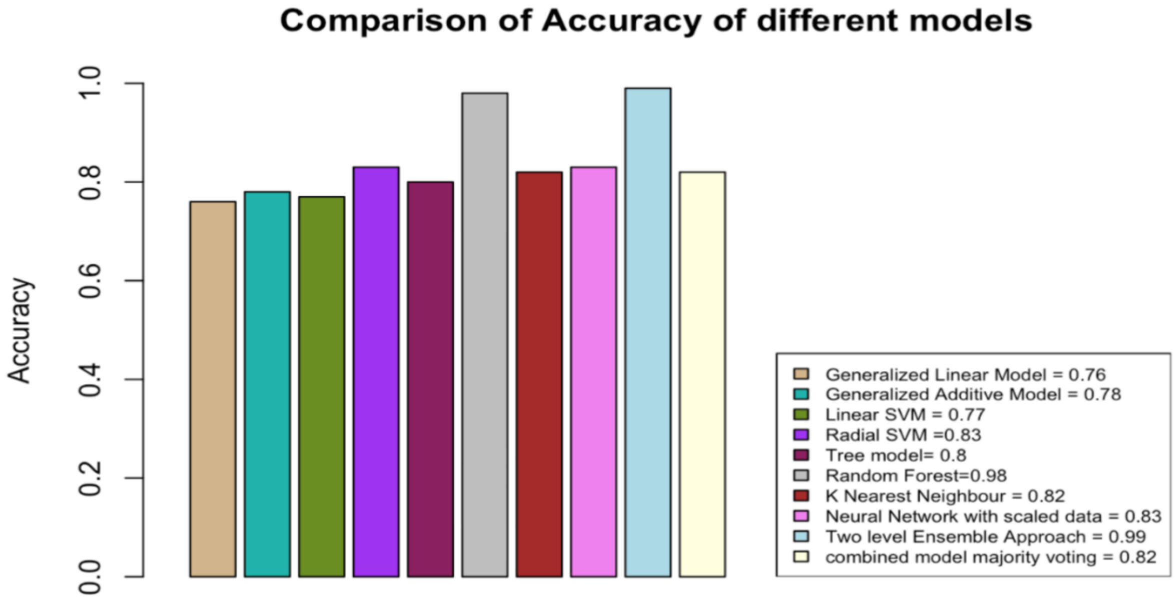

Figure 6: Comparative Analysis of accuracy for different models

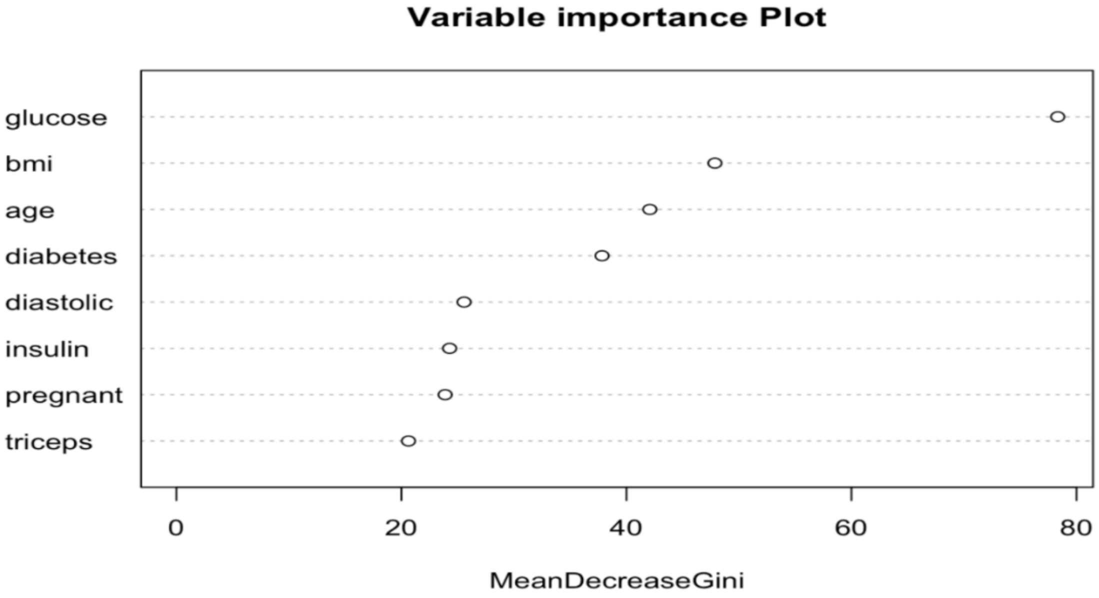

Figure 7: Variable Importance Plot

From the variable importance plot (Figure 7), it is evident that glucose is the most significant predictor feature in the Pima Indian Diabetes dataset to predict the outcome variable.

## CONCLUSION:

After analyzing 13 individual techniques for the prediction of the Pima Indian dataset from the most basic technique which was a simple Generalized Linear Model and increasing the complexity leading to Neural Networks, Two-Level Ensemble Approach can be considered as the best model of all resulting into a 99% of accuracy followed by random forest with overall 98% accuracy on the test dataset.

Taking into account the ROC-AUC curve, where the optimal model achieved a score of 0.98, indicates that the model is making accurate predictions compared to other models. Utilizing predictions from initial models to train neural networks in the second level proved to be more effective than training the neural network with raw data.

Additionally, Glucose seemed to be the most effective predictive feature followed by BMI and Age to predict the target variable.

## REFERENCES:

1) World health organization, Diabetes [accessed on 2023] Retrieved from: https://www.who.int/health-topics/diabetes#tab=tab_1
2) Benarbia, Meriem, "A Machine Learning Approach to Predicting the Onset of Type II Diabetes in a Sample of Pima Indian Women" (2022). *CUNY Academic Works.* https://academicworks.cuny.edu/gc_etds/4895
3) M. Abedini, A. Bijari, and T. Banirostam, “Classification of Pima Indian diabetes dataset using ensemble of decision tree, logistic regression and neural network,” Ijarcce, vol. 9, no. 7, pp. 1–4, 2020, doi: 10.17148/ijarcce.2020.9701.
4) Link to Dataset: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database