EARLY PREDICTION OF DIABETES USING CLASSIFICATION ALGORITHMS

Sahana Mukherjee, Harshith Nadendla, Anusha Narthu

EAS595 Final Project, SP21

Introduction

Diabetes Mellitus, commonly known as simply diabetes, is a chronic metabolic disorder, which is characterized by high sugar level, primarily arising out of the abnormal function of the pancreatic cells. The islets of Langerhans cells of the pancreatic gland are responsible for producing insulin: the chief hormone which regulates the amount of glucose level in the blood.

Three main types of diabetes are commonly found:

1. Type 1: Caused due to auto immune response where the pancreatic cells fail to produce enough insulin.
2. Type 2: A condition in which the cells of the body fail to respond to insulin properly.
3. Gestational Diabetes: Occurs in pregnant women without any prior history of diabetes, may develop significantly high blood sugar, giving rise to this third form of diabetes.

Symptoms of diabetes often encompass a wide range, starting from frequent urination, increased thirst, alopecia, obesity, increased appetite and many more. If left untreated, this disorder could result in serious health complications, including cardiovascular disease, which accounts for 75% of deaths due to this disease.

Below we list some of the major health issues arising out of diabetes, if left untreated for a prolonged time.

Health Complications due to diabetes

1. Coronary artery disease: chest pain, heart attack, stroke, atherosclerosis.
2. Neuropathy: nerve damage, paralysis.
3. Nephropathy: Kidney damage
4. Diabetic Retinopathy: Damage to the blood vessels of the retina, potentially leading to blindness.
5. Alzheimer's disease: mainly caused due to type 2.
6. Skin diseases: including bacterial and fungal infections.
7. Hearing impairment

Early detection of diabetes is essential as, if left untreated for a prolonged time, this disease can eventually lead to death, even though could seem trivial in the initial days.

Various signs and symptoms start quite early, and if proper evaluation of these symptoms can be done, the disease can be predicted at its early stages of development, hence aiding in expeditious treatment to avoid impediments in the later future.

This project leverages the power of Machine Learning Algorithms, to classify data, collected from patients, based on the varied symptoms they developed, and tries to predict whether an individual has the potential to develop the disorder. The data set deals with 520 patient records, with a total of 16 features each, the features representing the symptoms each record showed. Our objective is to take into account these 16 features, train a classification algorithm, which tries to learn, whether or not the individual has the disease, and eventually, we deploy the trained algorithm on unforeseen data, and evaluate how well it could do the predictions and also, we aim to realize which are the most significant symptoms, that could potentially contribute maximum to the development of this disorder. This entire exercise is achieved in multiple steps, starting from curating the dataset, data cleaning and necessary manipulations, selection of machine learning models based on the data set, training the models, validation and evaluation of the model performances and finally comparing the models to perceive the best fit model which can describe our objective to the maximum possible extent.

As mentioned above, there are multiple signs and symptoms that could arise in an individual, helping us to prognosticate the disease.

Among the many symptoms, the data set used in this exercise uses 16 features as listed below:

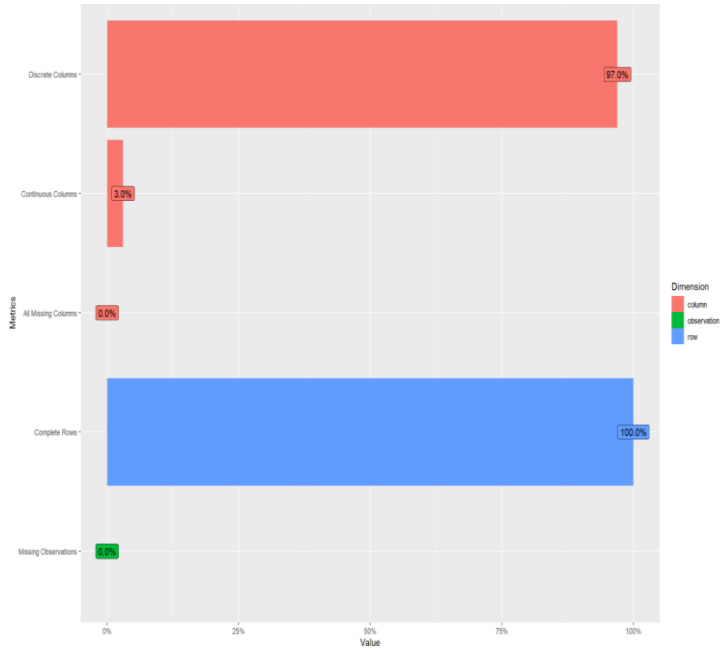
Symptoms to consider in concerned dataset

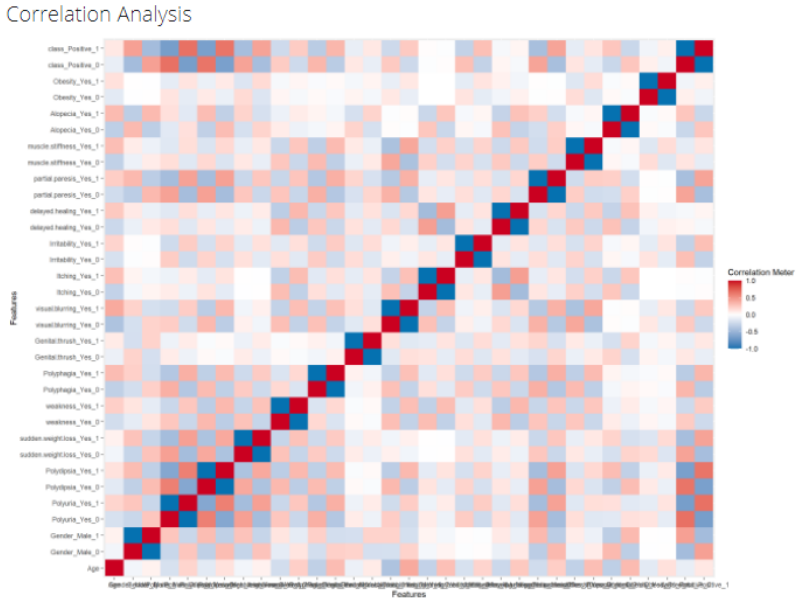
1. Age
2. Gender
3. Polyuria (frequent urination)
4. Polydipsia (increased thirst)
5. Sudden weight loss
6. Weakness
7. Polyphagia (increased appetite)
8. Genital thrush
9. Visual blurring
10. Itching
11. Irritability
12. Delayed healing
13. Partial paresis
14. Muscle stiffness
15. Alopecia
16. Obesity

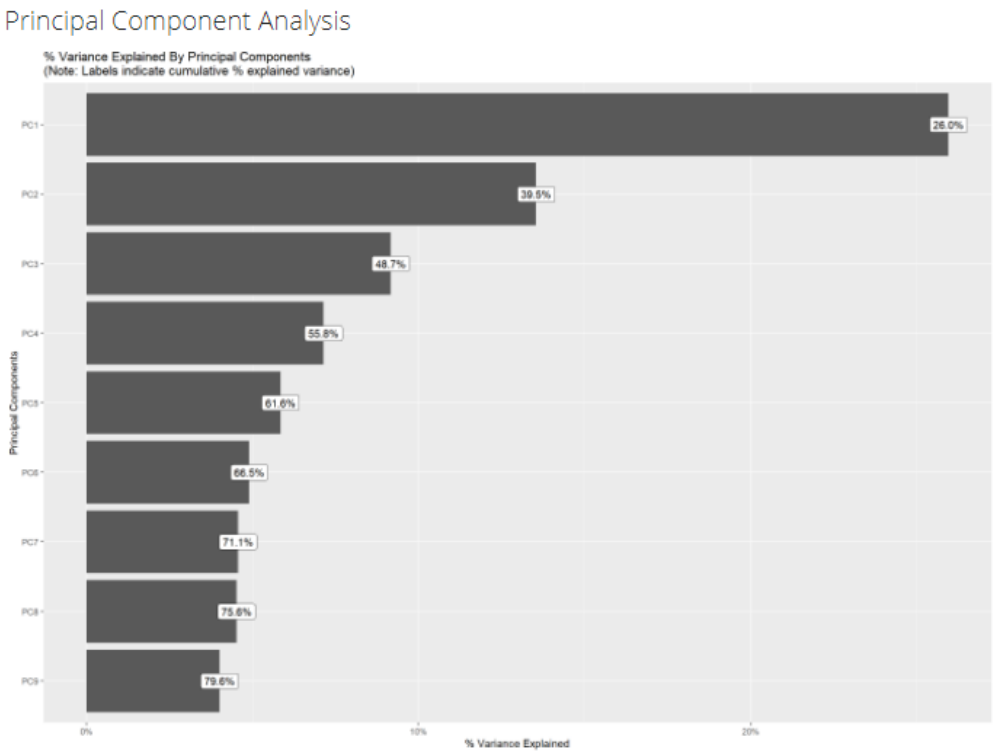
Data Exploration, Curation and Preparation

As stated previously, the concerned dataset contains 520 records and 16 features, representing the symptoms to consider.

Below is a brief pictorial overview of how the data set that we are about to analyze, looks like after converting the features to factor variables (with two levels 0 and 1) and applying one hot encoding for the categorical features.







Some inferences drawn after data exploration:

- There are no missing values in our data set.
- There is also no significant amount of correlation among the features, so we can fairly assume that the 16 features are independent of each other.
- Principal component analysis reports that 9 principal components, which are basically formed by the weighted linear combination of the independent features, are sufficient to describe around 80% variability in the data set. As the data set is fairly manageable with 16 features, and also there is no major correlation among the features, we do not opt for principal component analysis on the data set for model evaluation at this moment.

Classification of the data using Machine Learning Algorithms

The concerned data set is well balanced with two labels: the condition of having the disease is labelled as 1 and the condition of not having the disease is labelled as 0.

As the data happens to be fairly linearly separable, and the features can be fairly assumed to be independent of each other, we deploy the following classification algorithms on our data set: Logistic Regression, Naïve Bayes Classifier, Random Forest Classifier and Boosted Decision Tress.

Logistic Regression

The data has been split into two halves, using the conventional train test split of 70-30, where 70% of the data is used for training and remaining 30% is kept aside for testing phase.

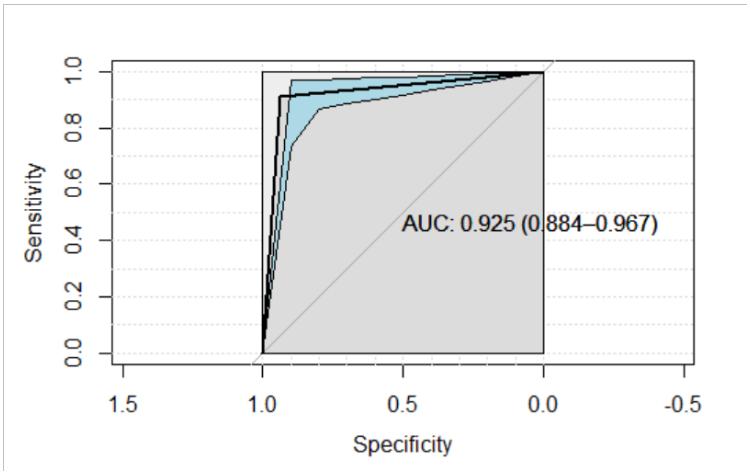
Threshold value used for probability estimation is 0.5, which indicates, that, if the calculated probability that the individual has the disease based on the input features, is greater than 50%, the model considers it has been labelled as 1, i.e., as the condition of having the disease and anything below 0.5 is being labelled as 0 by the model.

This threshold value is a tunable parameter for logistic regression, and depends on the business requirements and available data sets.

Below is the confusion matrix generated by the model, after the trained model is tested on the 30% data reserved for testing.

	Actual	
	0	1
predictions		
0	62	8
1	4	82

The ROC curve gives the model's performance, by plotting the True Positive Rate and False Positive Rate.



AUC is calculated as 92.5%, which is significantly high and depicts that the model has been able to distinguish between the two classes quite well, labelling 0s as 0s and 1s and 1s, most accurately with few exceptions which is represented by the contingency table above.

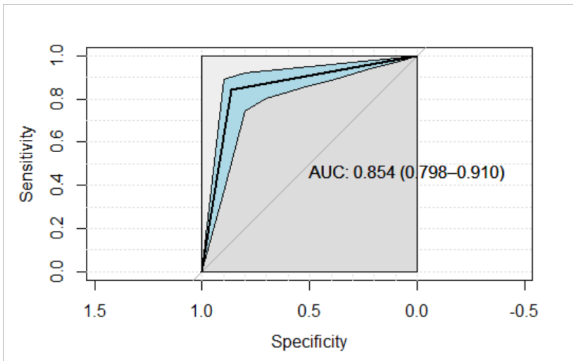
Naïve Bayes Classifier

Similar approach of 70-30 train test split is followed and the model is trained using Naïve Bayes classifier. The features, being assumed to be fairly independent of each other, with no major correlation among them, NB Classification is presumed to perform well, as this algorithm considers the input variables to be independent for easy computation.

Below is the confusion matrix generated by the model, after the trained model is tested on the 30% data reserved for testing.

	Actual	
	0	1
predictions		
0	57	14
1	9	76

The ROC curve is plotted and area under the curve is found to be around 85%.



Naïve Bayes is the only probabilistic approach used in this exercise. The performance of the Naïve Bayes Classifier can be explored more by playing around with the hyperparameter of 'type', which indicates what kind of probability (conditional a-posterior probability, or maximal probability) to return for each class. The former type is used in this exercise.

Random Forest and Boosted Decision Trees

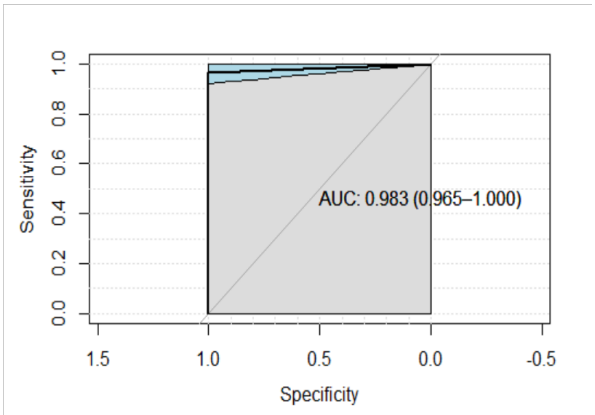
Confusion Matrix for Random Forest:

	Actual	
	0	1
predictions		
0	66	3
1	0	87

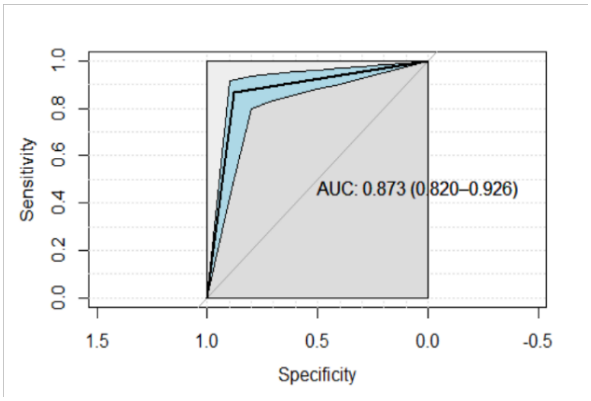
Confusion Matrix for Boosted Decision Trees

	Actual	
	0	1
predictions		
0	58	12
1	8	78

AUC for Random Forest



AUC for Boosted Decision Trees

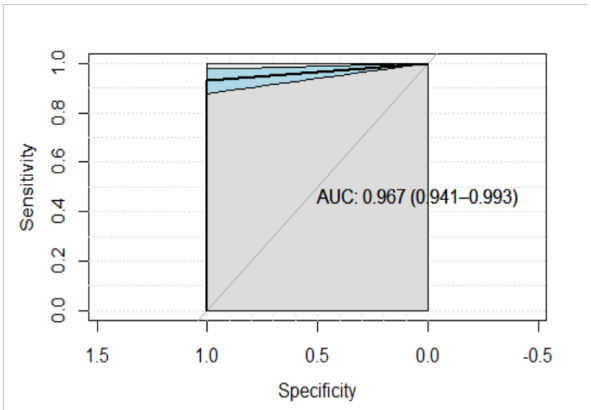


Use of Cross-Validation to estimate the skill of trained model in general on unforeseen data:

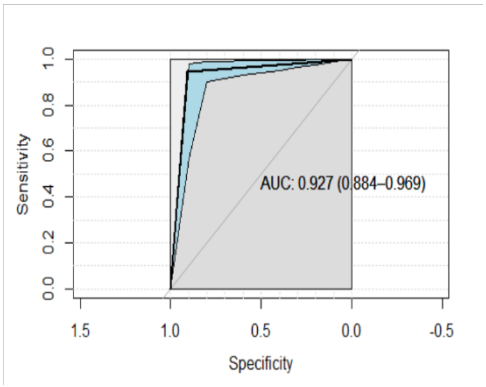
Cross validation is incorporated to discover the ability of the model to perform in general when used to make predictions on data not used during the training of the model. This helps to forecast the generalizability of the model, and how much variance is hidden in it. Lesser the variance, better is the model, i.e., the performance of the model should not change drastically, with the change in the dataset, provided evaluation metrics remain the same.

Below are the AUCs obtained after evaluating the model based on repeated cross validation.

Random Forest Classification after using CV



Logistic Regression after CV

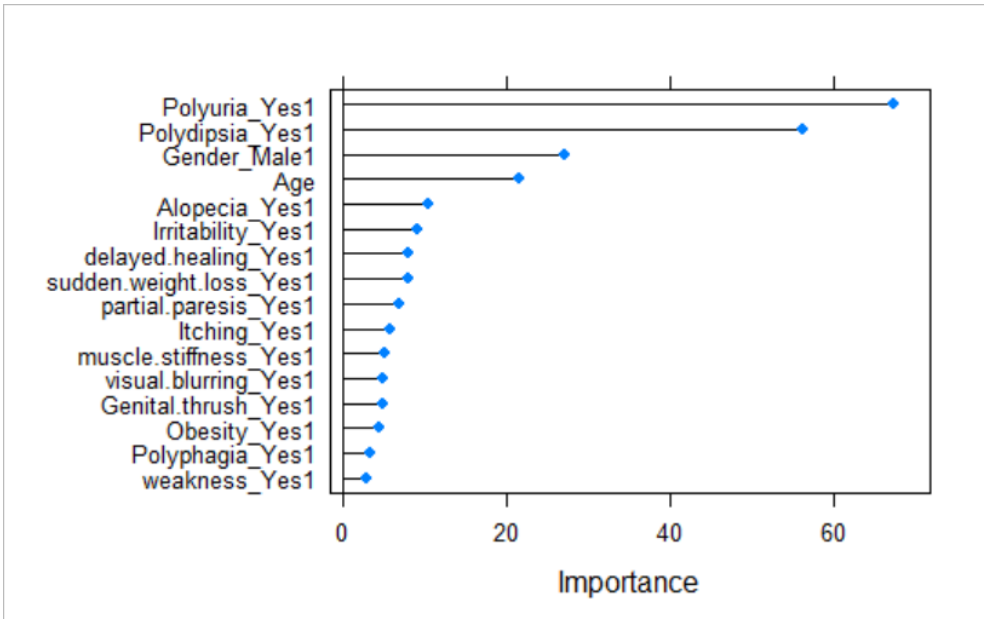


Extracting the importance of features and training the model on top 5 features

Among the 16 independent features, now the task is to explore which are the most important features, i.e., which are the features that are contributing most to the predictions.

Method Random Forest Classifier is used for feature selection. The more a feature can decrease the impurity, the more importance the feature is. In random forests, the impurity decrease from each feature is averaged across several decision trees to determine the final importance of the variable. The importance is calculated based on different parameters of which Gini index, entropy gain are significant and widely used in practice.

Features according to their importance or contribution to making predictions:



The above analysis shows that Polyuria (frequent urination), Polydipsia (increased thirst), age, gender and Alopecia are the top 5 features, which lead to the possible inference that these symptoms could potentially be early indicators of diabetes and should not be neglected, rather taken care of to avoid health complications in the future.

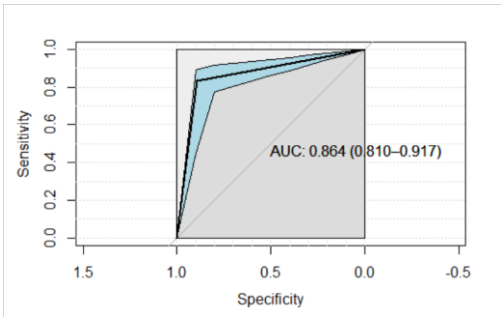
Now, we train the model on the top 5 features, instead of all the 16, and observe the performance metrics.

Logistic Regression using top 5 features:

Confusion Matrix:

	Actual	
	0	1
predictions		
0	59	15
1	7	75

AUC Plot

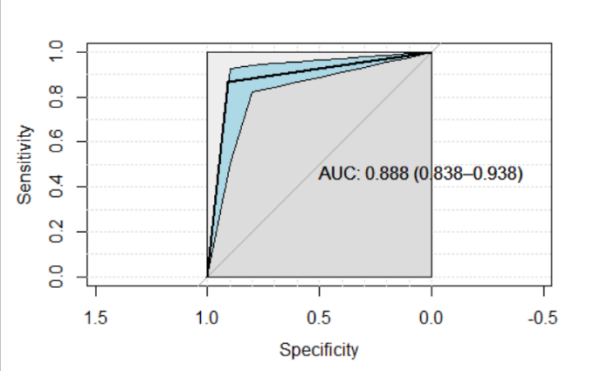


Random Forest Classifier using top 5 features:

Confusion Matrix

	Actual	
	0	1
predictions		
0	57	9
1	9	81

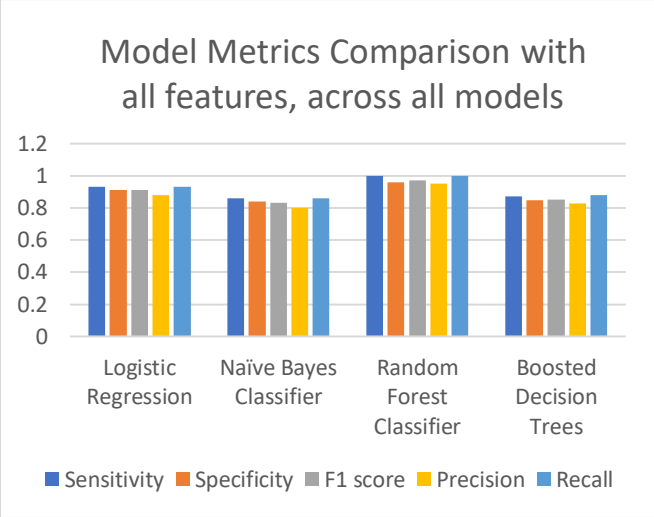
AUC Plot

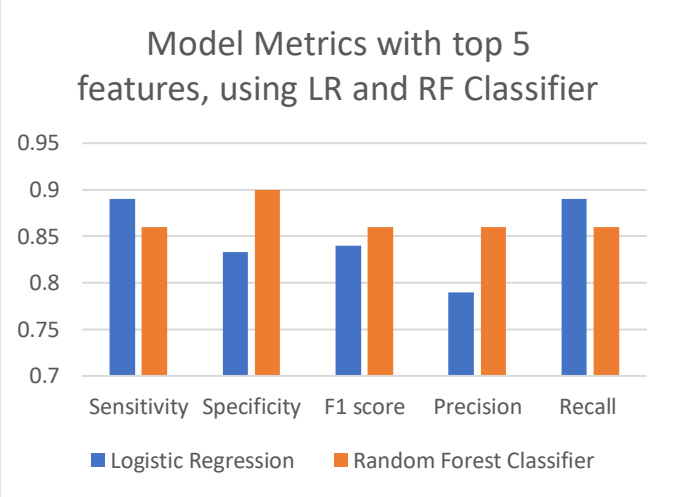


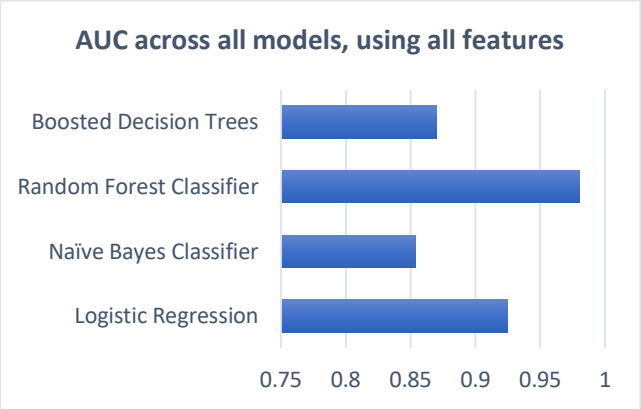
Inference from model evaluation using top 5 features:

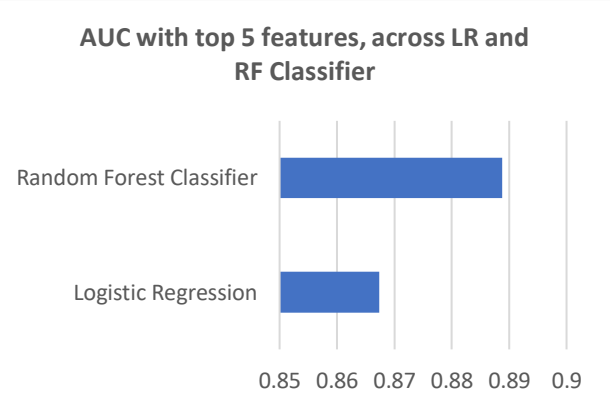
The model performance slightly decreases when the top 5 features are used, provided the evaluation metrics remain the same. This could be possibly because of the curse of dimensionality, too many features could increase the model performance, but it could lead to overfitting and thus inaccurate predictions. Again, using the top 5 features, we are losing some information, there might be some important information captured in the remaining features which are not being considered in top 5. So, a tradeoff is necessary between selection of top features, and information loss, and this again will depend upon specific business requirements.

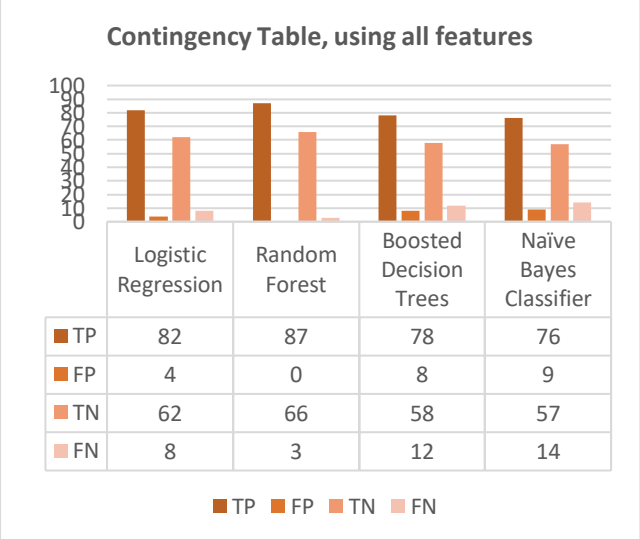
Comparison and evaluation of performance metrics across all deployed models

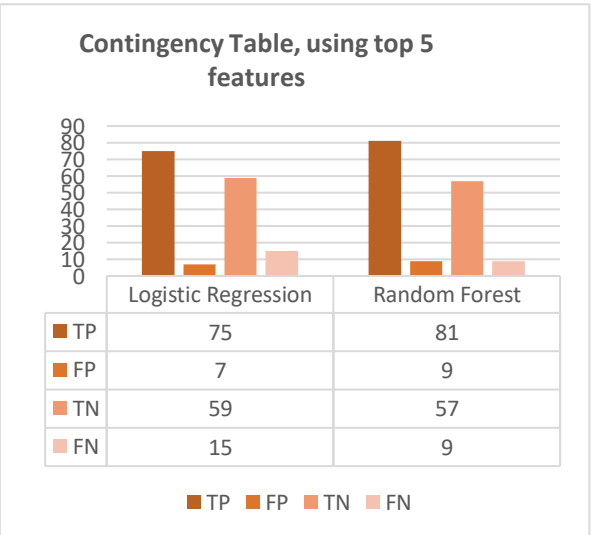












Conclusion

The exercise aims to predict whether or not an individual is a potential candidate of having developed the disorder, based on historic data of patients, and 16 symptoms or features.

Several models like Logistic Regression, Random Forest Classifier, Boosted Decision Trees and Naïve Bayes Classifier have been deployed to achieve the purpose. Of all the models, inference has been drawn that Random Forest Classifier, followed by Logistic Regression, performs possibly the best, with maximum area covered under the ROC. The dataset being almost linearly separable, unbiased, and relatively low correlation among input variables are the key factors that lead these two models give quite good predictions.

Considering that this exercise is being performed to provide early detection of the complication, in order to avoid serious health issues in the future, the FALSE NEGATIVES, obtained from the models, are quite significant metrics to deal with, and the model should be selected such that it has least number of false negatives, as false negatives indicate that the individual has the symptoms that could potentially lead to the disease, but the model predicts the label as 0, i.e., the condition of not having the disease. In this respect also, it is found that Random Forest Classifier performs quite well, with only 3 instances of False Negatives.

Cross Validation is performed to assess how general the model can be, and feature importance is extracted to discover the most important features in the data set, to reduce the inaccuracy due to overfitting. The two best performing models, logistic regression and Random Forest Classifier are again trained using the subset of top 5 features and performance metrics are observed. This analysis leads to increased number of false negative instances, which is not desirable. So, there should be a tradeoff between these issues, and it can be concluded, that Random Forest and Logistic Regression can be performed to achieve the purpose, with a little tradeoff between false negatives and overfitting issues.

Future Scope

The dataset not being considerably large enough, neural networks were not observed to provide better performance metrices, as when the dataset is fairly linearly separable with only two classes, a neural network will mostly behave like a simple perceptron, which essentially learns like logistic regression with sigmoid activation function.

Future work would include several tasks of which tuning the hyper parameters of the models would be crucial. Repeated manipulation of variable parameters would enable us to assess if there could be some value of those parameters which could come up with better performance, with lower misclassification rate. The hyper parameter optimization approach has already been incorporated in this exercise, including Grid Search and Random Search. Aim will be to dedicate more time and efforts to these two methods, but would not be limited to them. Approaches like Bayesian Optimization, Gradient Based Optimization would also be explored in the near future, to yield the possibly most optimal machine learning model with minimal misclassification rate. The dataset being impressively interesting, attempts would be made if more data can be collected, on this topic, so that more information can be captured about the leading symptoms, and the analysis can dig deeper into the vast domain of diabetes-research and contribute its bit towards a healthier and safer tomorrow.

Contributions:

Sahana, Harshith, and Anusha together materialized this work. Anusha performed detailed exploratory data analysis, including one hot encoding and principal component analysis. Sahana performed Naïve Bayes Classification and Logistic Regression (including feature selection) on the dataset. Harshith took care of the two ensemble methods: Random Forests and Boosted Decision Trees. Cross Validation with Logistic Regression and Random Forests were taken up by Anusha. Visualizations were taken care by Sahana. Documentation of introduction, model performances, conclusion and future scope were collaboratively taken care of by all.