**Pre-diabetes prediction models with Machine learning**

Dongxu Fu, Stephen Omondi, Debra Kernstock

June 15, 2019

**INTRODUCTION**

Diabetes is a chronic disease which is caused by the body’s inability either to use insulin or to produce insulin. It poses a great threat to human health. The characteristic of diabetes is the high level of blood glucose, which it sequences as defective insulin secretion. Diabetes can lead to quite a few chronic damages and dysfunction of various tissues, including the eyes, kidneys, heart, blood vessels and nerves. Diabetes can be divided into two categories, type 1 diabetes (T1D) and type 2 diabetes (T2D). Patients with type 1 diabetes are normally younger, mostly less than 30 years old, and required insulin therapy. Type 2 diabetes occurs more commonly in middle-aged and elderly people, which is often associated with the occurrence of obesity, hypertension, dyslipidemia, arteriosclerosis, and other diseases. People with diabetes are at higher risk of serious health complications including stroke, blindness, kidney disease, etc.  By 2015, diabetes affects over 30 million children and adults in the U.S. That's 1 in 11 Americans. Apart from that, there are another 84 million Americans who are considered prediabetic and are at risk for developing type 2 diabetes - 90% of whom don’t know their condition.

Prediabetes is a serious health condition where blood sugar levels are higher than normal, but not high enough yet to be diagnosed as type 2 diabetes. It is well known in the clinical practice that early diagnosis and treatment can greatly delay the occurrence of diabetic complications and improve the patient’s life quality. However, as most of those people do not have typical diabetes symptoms, such as weight loss, fatigue - subsequently unaware - the diagnosis of prediabetes becomes very challenging. Therefore, the methods to quickly and accurately diagnose and analyze prediabetes are badly needed.

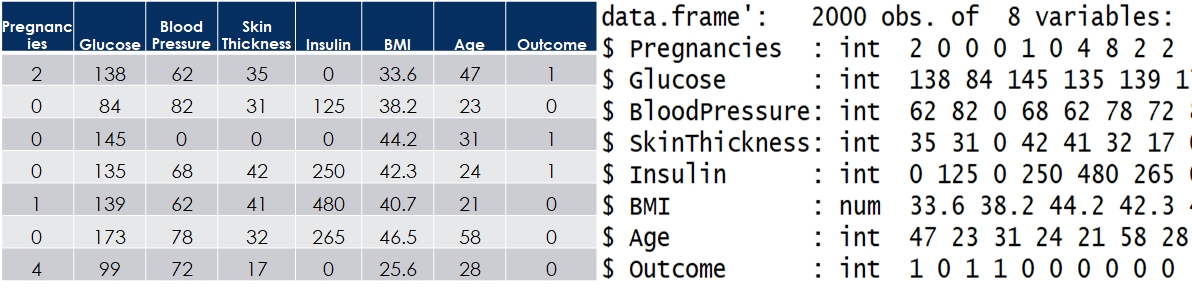
In medicine, diabetes diagnosis mainly depends on fasting blood glucose, glucose tolerance, and random blood glucose levels. Machine learning can help people make a preliminary judgment about diabetes mellitus according to their daily physical examination data, and it can serve as a reference for doctors. However, selection of valid features and the correct classifier are the key in machine learning.

In the last decades, various machine learning methods have been used in predicting diabetes, including support vector machine (SVM), decision tree (DT), logistic regression, etc. Recently, some research groups also proposed a new ensemble approach which combines 30 machine learning methods to optimize the algorithms. However, the results derived from different methods do not always agree with each other, there is no gold standard for which methods should be used for the diabetes prediction. The favorable methods depend on the attributes in the database that was selected. In the current study, a database containing the seven most commonly used diabetes index in the clinical practice was selected alongside four normally used algorithms to determine which one performs the best for predicting diabetes with this type of datasets.

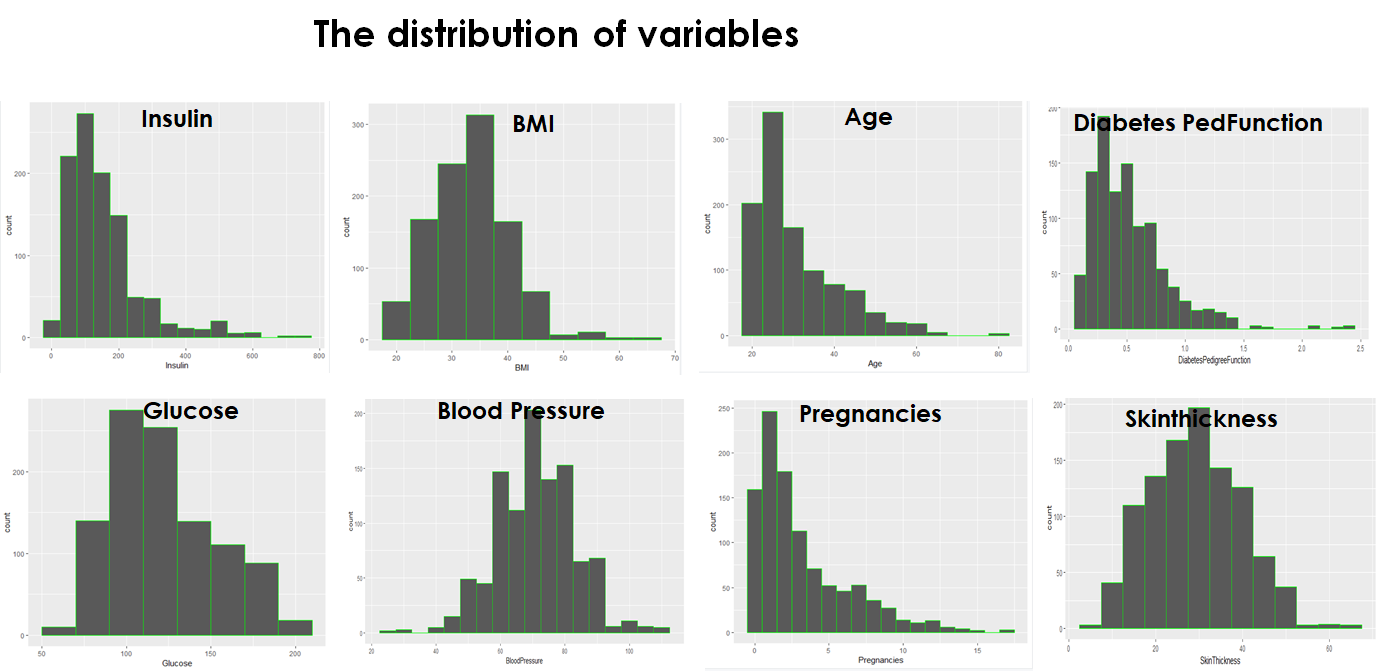
**ABOUT THE DATA**

The dataset used in this study is taken from a hospital in Frankfurt, Germany. The dataset is also available from Kaggle at [https://www.kaggle.com/general/42501#238144](https://www.kaggle.com/general/42501). All patients are females at least 21 years old, and of Pima Indian heritage. The dataset contains 7 attributes namely; times of pregnancy, plasma glucose concentration after a 2-hour oral glucose tolerance test, diastolic blood pressure, triceps skinfold thickness, 2-hour serum insulin, body mass index, diabetes pedigree function, and age. The diagnosis outcome is labeled as 1 or 0, referring pre-diabetes or not pre-diabetes. There is a total of 2000 observations in this dataset, and a partial raw dataset and brief summary of data structures are shown in Figure1. The distribution of each parameter was also analyzed and plotted with bar graphs, as shown in Figure 2.

***Fig1. Part of the raw dataset and brief summary of data structure***



***Fig2. The distribution of each variable***



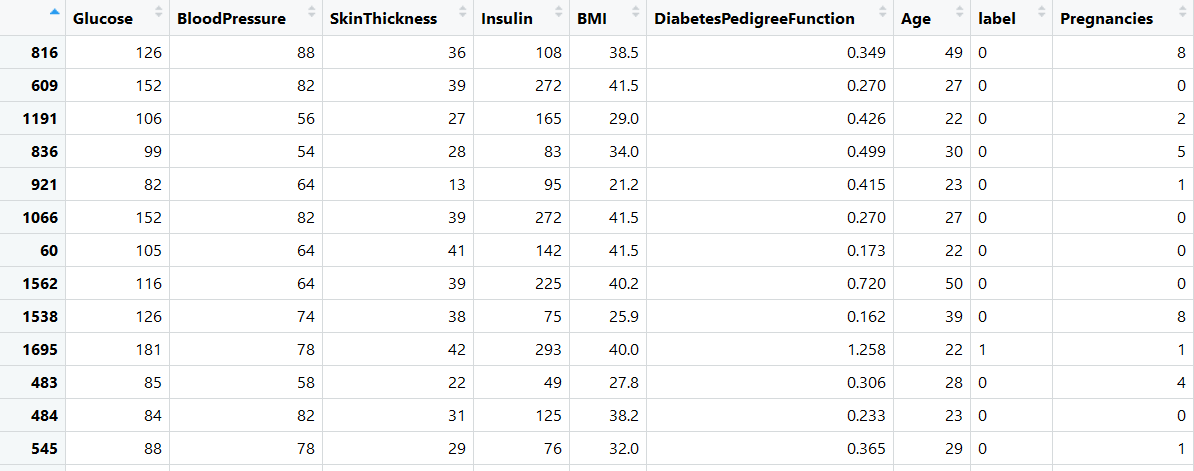
**Data preparation and cleaning**

Data cleaning was performed first before analysis, including checking missing values, incorrect values and outliers. Missing values in this study are referred to any datapoint without values. For numeric values, incorrect values are any values that is zero except in pregnancy and outcome columns, where 0 is a real value, referring to pregnancy time and non-prediabetes accordingly. In this dataset, there are no missing values and outliers. The observations with incorrect values were removed, and the label column was transformed from numeric to factor.

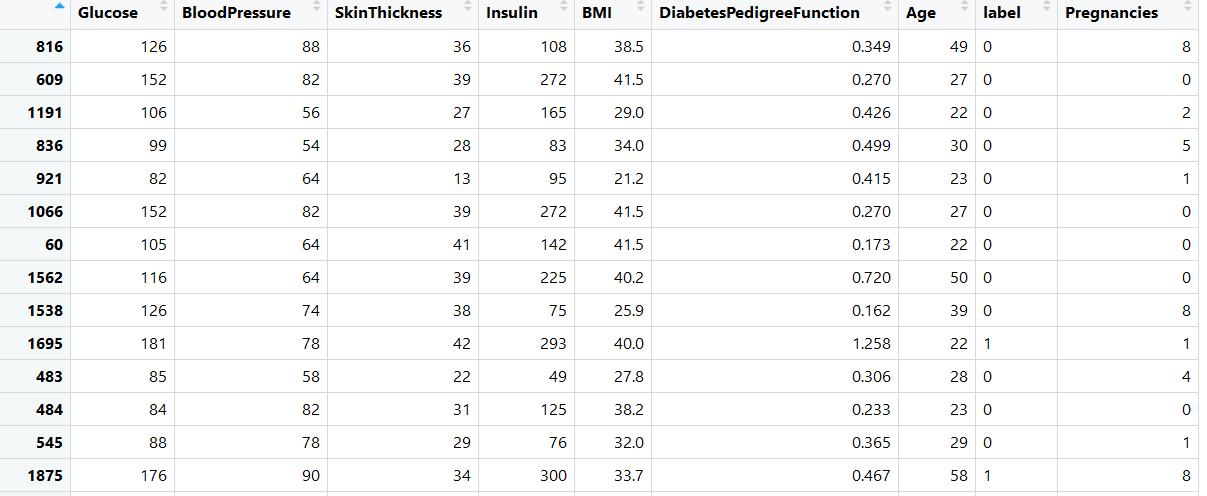
**Splitting training and testing dataset**

Next, the dataset was split into two parts; training data and testing data for modeling purpose, with “cut point=2:8”. In this testing data, the label was removed as well. Sample training set and testing set were shown in table 1 and table 2.

**Table 1. Sample Training Set**



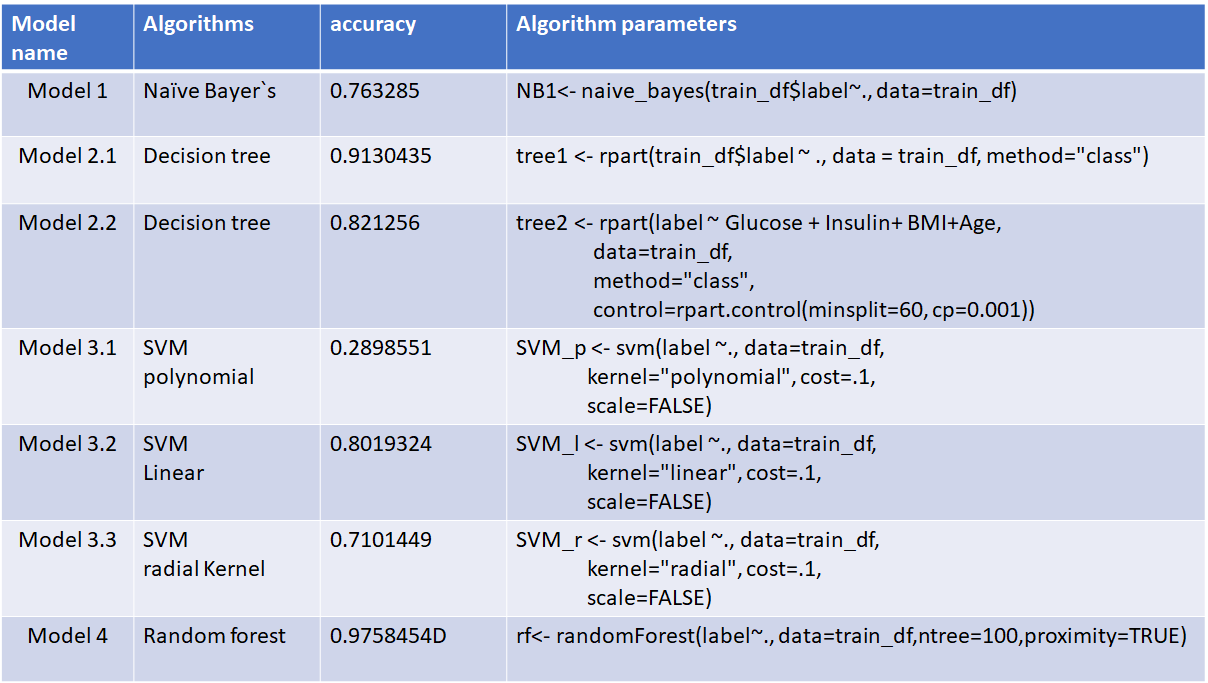
**Table 2. Sample Testing Set**

****

**ANALYSIS & MODELS**

In this project, four most used algorithms namely, Naïve Bayes, Decision Trees, Support Vector Machine and Random Forest (RF) were performed. In support vector machine, different Kennels were tested, including polynomial kernel, linear kernel and radial kernel. The parameters for these models are shown in table 3.

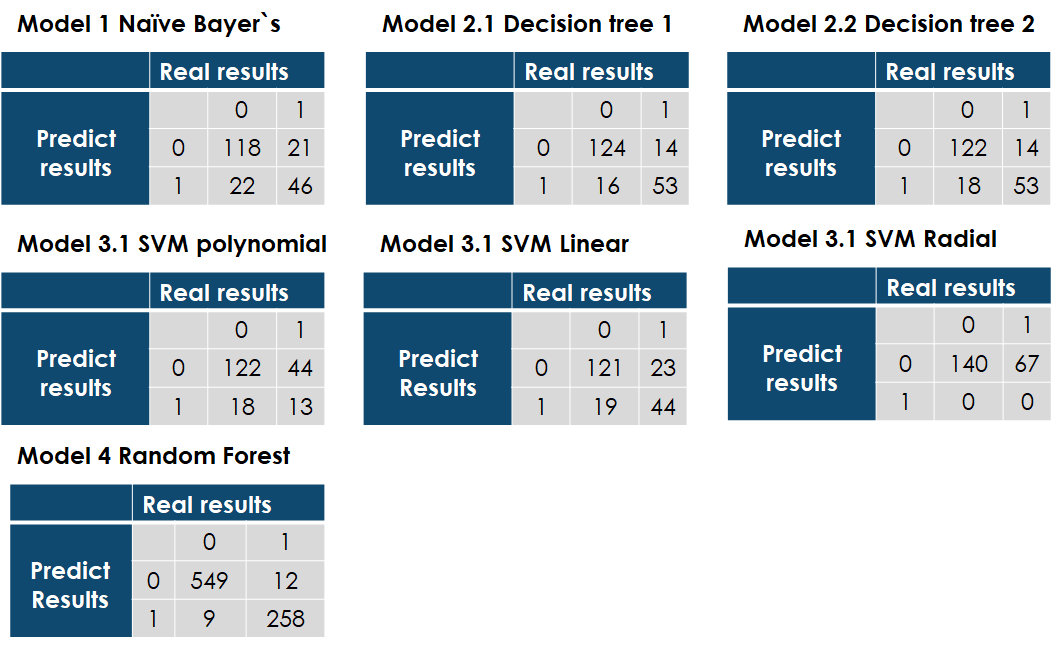
**Table 3. The algorithms of the models in pre-diabetes prediction**

****

**RESULTS**

The prediction results were summarized in the confusion tables, figure 3.

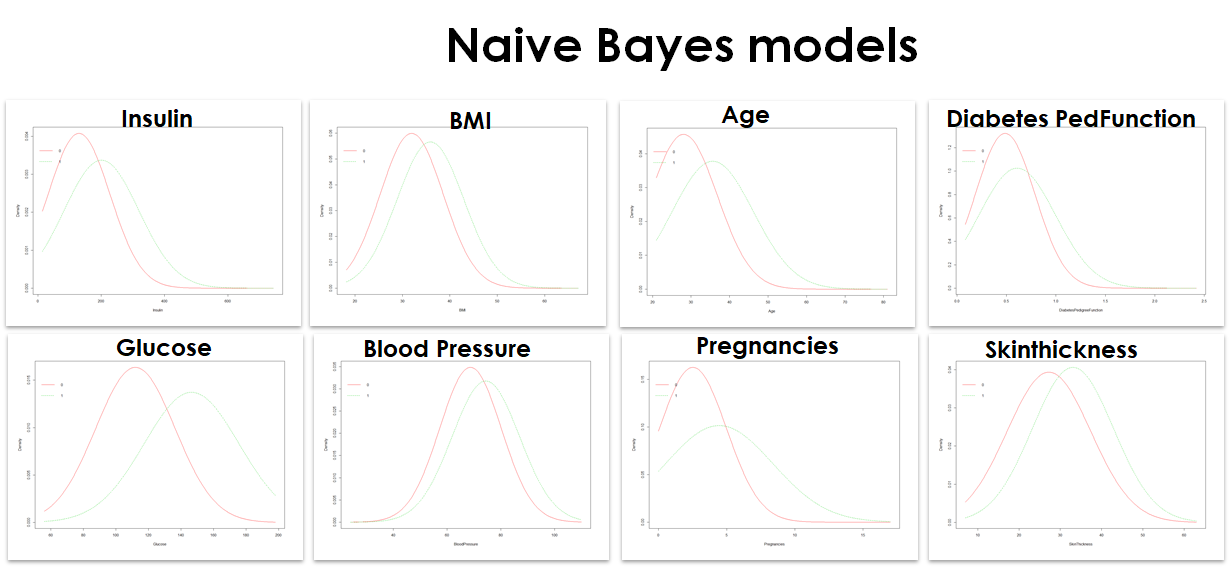
***Fig3. The confusion tables of the models in pre-diabetes prediction***

****

**Naives Bayes**

Naive Bayes predicted 77.29% accurately – showing that 145 people would not be prediabetic while 62 people would be. More details of this model is described in Figure 4, where red lines show the number of participants who are not prediabetic and dotted green line show the predicted number of participants who are prediabetic. On one hand, there are more people who are not at risk, which is a win for their health concern. On the other hand, the model had a below average prediction which leaves room for error. The model does not account for nearly 22.71%, which means additional model parameter tweaking may be needed to make it meaningful.

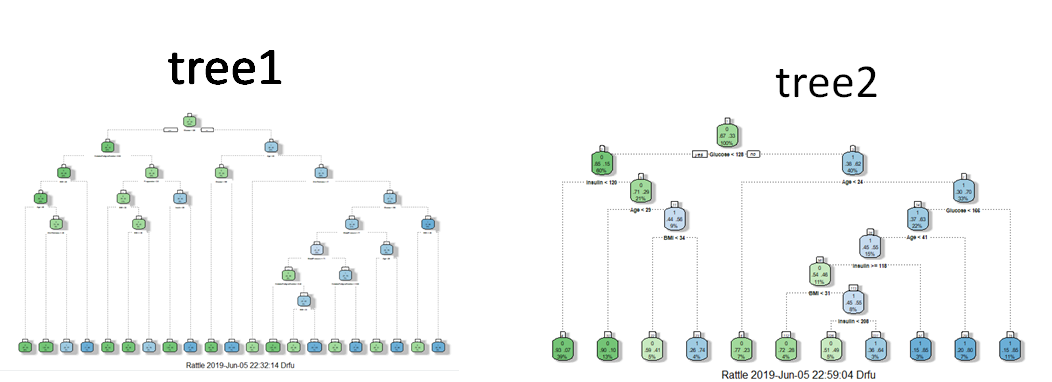
***Fig4. The Naïve Bayes models***

******

**Decision Trees**

Decision trees were used because they offer a great visual understanding for non-technical people. The unpruned decision tree had 247 nodes. The decision tree analysis resulted in an 80.19% accuracy, which is not great but better than using Naive Bayes. The pruned decision tree (Model 2.2) consisted of 227 nodes, which reduced the decision tree slightly and increased the accuracy slightly as well to 82.12%.

***Fig 5. The Naïve Bayes models***

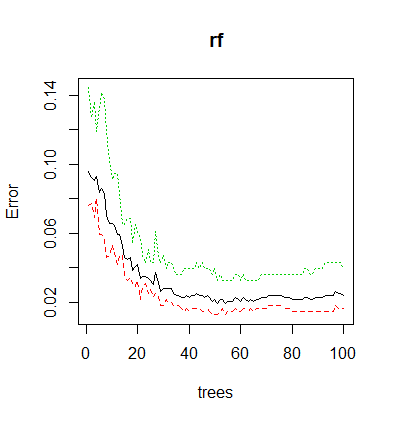
****

**Support Vector Machine**

The next set of model analysis conducted was SVM utilizing polynomial, linear, and radial kernel. The analysis reveals that the highest accuracy achieved was from linear kernel at 77.78% which still does not pass the 80% mark, as shown in table 3 and Figure 3. Additional research is needed to reveal the limitation of this analysis. Another modification of the data will need to be analyzed to see if the model fits better than it did for this iteration.

**Random Forest**

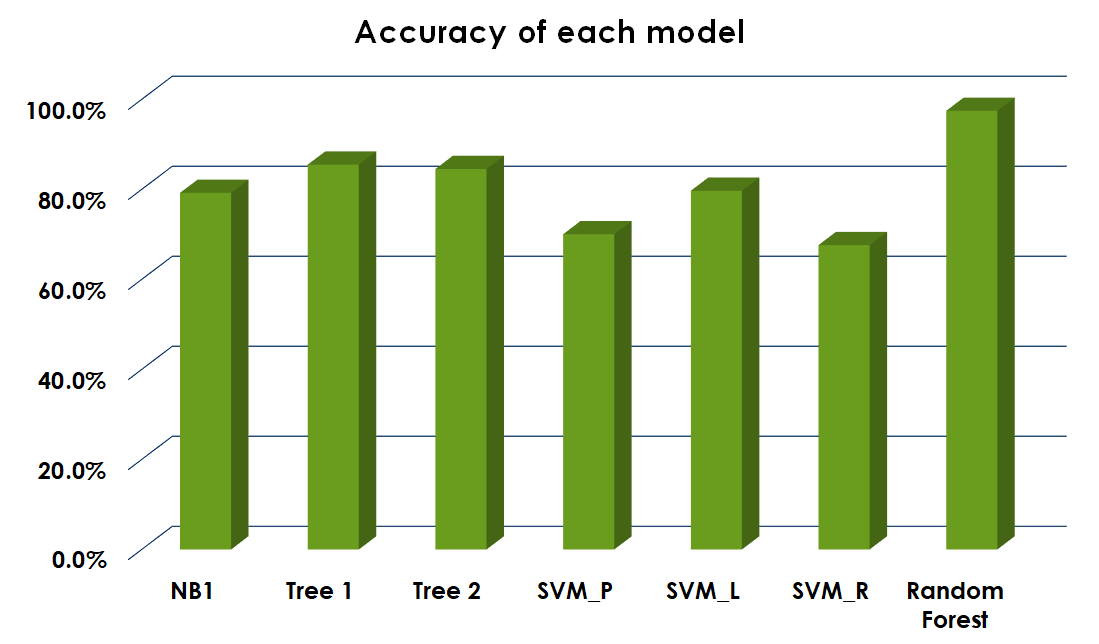
The final analysis conducted for this dataset used Random Forest on the training set. The results were better than the other models as noted. A graph summary is shown in Figure 6, also showing the error rate of this models.

***Fig6. The random forest models*** fffffff****

**The accuracy of each model**

In a summary, Random Forest yielded the best results, with accuracy as high as 97.8%. Decision Tree models were second best with an accuracy slightly higher than 80%. Naïve Bayes which typically performs best in numeric datasets as the one used in this study, did not perform as well as expected. SVM did not yield better results either. In fact, none of the following kernels, polynomial, linear, and radial kernel, have accuracy score over 80%.

***Fig7. The accuracy of each model***

****

**CONCLUSION**

Diabetes is the seventh leading cause of death in the United States. The total direct and indirect estimated cost of diagnosed diabetes in the United States was $245 billion. After adjusting for age group and sex, average medical expenditures among people with diagnosed diabetes were about 2.3 times higher than expenditures for people without diabetes. Currently, there is a consensus in the medical community that early diagnosis is critical to diabetes treatment but effective methods for quickly screening and diagnosing this disease are still lacking.

Strategies can be developed for people at risk for prediabetes and diabetes to achieve healthier lifestyles and longer lives. Diabetes is quite manageable, and, in some cases, it can be reversable. Simple dietary changes and more exercise can make a world of difference. People who have high levels of glucose, insulin, and BMI will eventually become prediabetic and then get diabetes. Wouldn’t it be great if other than seeing a doctor, a person can input how much they weigh, blood sugar levels, and other information on a form which then can see if they are at risk?

Technology is growing at an exponential rate where people can access more information now than ever. As discovered from the multiple models, prediabetes can be predicted from attributes and that is exciting! The implications from this research can be further enhanced and studied as they should be. The information that was learned as a result of experimentation is only a start. The research can be modified to identify different types of diabetes in the future, as to helping the underserved populations with little or no access to diabetes prevention services can go a long way. Health insurance is expensive and not everyone has access to doctor, therefore, knowing simple means of prevention goes a long way.