Big Data Analytics Project

Bruno Hung

Student #: 500457821

Supervisor: Tamer Abdou

Date: 12/3/2021

Table of Contents

[Abstract 3](#_Toc452457337)

Literature Review [4](#_Toc452457338)

[What do you already know about this topic? 4](#_Toc452457339)-5

[What do you have to say critically about what is already known? 5](#_Toc452457339)-17

[Has anyone else done anything that is related?](#_Toc452457339) 18-19

[Where does your work fit in with what has already gone before?](#_Toc452457339) 19-20

Why is your research worth doing in light of what has already been done?21-22

[Shortcomings of the work](#_Toc452457340) 22-23

[Conclusion](#_Toc452457340) 23

[Github Link](#_Toc452457340) 24

[References](#_Toc452457340) 25

# Abstract

This paper presents and overview of machine learning techniques in classification of early stage diabetes. According to Diabetes.ca, the 2019 Diabetes Canada Cost Model found that the rate of diabetes and prediabetes continue to rise. One in three Canadians have diabetes and there is a high chance of developing diabetes as one age further. Diabetes is not something anyone should make light of. It can lead to a lot of complications such as long term disability, heart diseases, kidney diseases, chronic diseases, and etc. This project would explore the chance of detecting diabetes in their early stage. By detecting diabetes early, patients can start treating it earlier to stay as healthy as possible. Treating it earlier can slow down the development of diabetes. The main research idea was to try to use a new dataset and see if there are any relationships between diabetes and the attributes as well as to see if these attributes could be used to build a model to detect diabetes in its early stage. The dataset I would be using is from UCI Machine Learning Repository. The dataset will have 520 instances with 17 attributes such as age, sex, itching, muscle stiffness, and etc. It is collected using direct questionnaires from patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh. As for the machine learning techniques that I would be using, I would be using decision tree technique and logistic regression within the classification method. There was no missing data. From then on, confusion matrix will be created to show the accuracy and determine which model would be the most fit for this research. The modeling will be done using python.

**Literature Review**

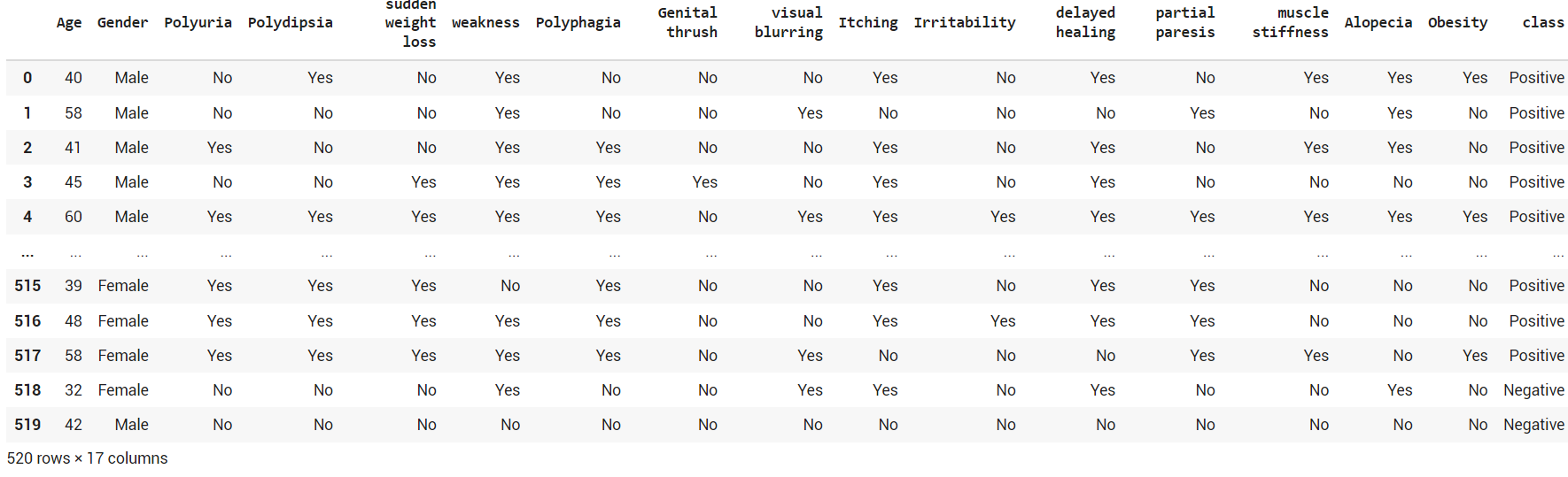
**What do you already know about this topic?**

Generally, as anyone had started to age, the body functions and health would not be as healthy and strong as when they were young. In general, one would assume that this would be the case for diabetes as well. However, according to a study result from International Diabetes Federation Diabetes Atlas, the prevalence of diabetes is increasing worldwide and the adults that were diagnosed with diabetes were getting younger and younger. Half of their estimated death of 4.2 million in 2019 due to diabetes would be to occur in adults younger than 60 years old. In addition, according to another study by Kenneth E Heikes (2008), the total number of cases worldwide was projected to increase from 171 million in 2000 to 366 million by 2030. This meant that there were also a lot of undiagnosed diabetes patients out there whose quality of life and life expectancy would be worsen and be shorten comparing to if they were to get treated. One of the reasons why diabetes was becoming such a problem was because it is very hard to be noticed by diabetes patients in its early stage. For example, in the U.S. in 2002, the prevalence of diabetes was estimated to be 19.3 million, of which about 5.8 million cases were undiagnosed. Furthermore, 41 million individuals were estimated to have pre-diabetes. Prediabetes had an increased risk of gradually developing type 2 diabetes over the years. Sourcing from an online article from Uchicago medical, their professor, Neda Laiteerapong, had said that if one with diabetes is diagnosed early in its development, the treatment could put patients’ health “on a trajectory for the rest of their lives”. By starting to control the blood sugar level earlier, the patients could feel the benefit of their medications or treatment much earlier. Delaying the time of treatment would only delay the time that the patients would feel the benefit of it as battling diabetes was a long process. Laiteerapong also added that it could take up to 10 years before patients could feel the benefit. This literature review examined the ability to detect early stage diabetes. It also explored how the patients’ biological traits or small sickness can be a symptom of early stage diabetes.

**What do you have to say critically about what is already known?**

A lot of researches had been done on diabetes and prediabetes previously. A lot of countries and their researchers and medical doctors had been conducting surveys to collect data on diabetes to help them further understand the cause, implications and symptoms of it. For the dataset (Figure 1) that was being examined in this literature reviews, the dataset was collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet. Figure 1 gave a high level preview of what the entire dataset look like.

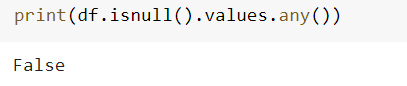
Figure



The dataset was multivariate. It had 520 instances and 17 attribute which were the signs and symptoms of newly diabetic or would be diabetic patients. Some of these attributes were some of the very common stroke symptoms such as weakness, obesity, visual blur, muscle stiffness, and delayed healing. A study was done by April Carson (2012) indicating that in their population-based study, almost one in four individuals with diabetes reported stroke symptoms. Therefore, these attributes were very reasonable as the symptoms of diabetes were often sharing symptoms of other complications such as kidney disease and heart disease and could be confused as other complications to people that were not as aware.

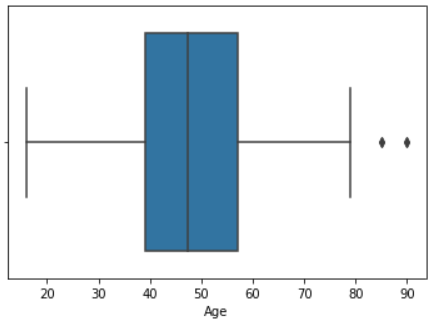
Within this dataset, the attributes were mostly an object type. Out of the 17 attributes, only age was an integer type. This information would be important to know when it comes to working on the statistical model. In addition, there were no missing data in the dataset. This was known from the calculation (Figure 2) being done in python.

Figure



To investigate further, out of the 520 people, the oldest person was 90 while the youngest was 16. These were most likely to be outliers but to confirm, a boxplot (Figure 3) had been used to help visualize the distribution.

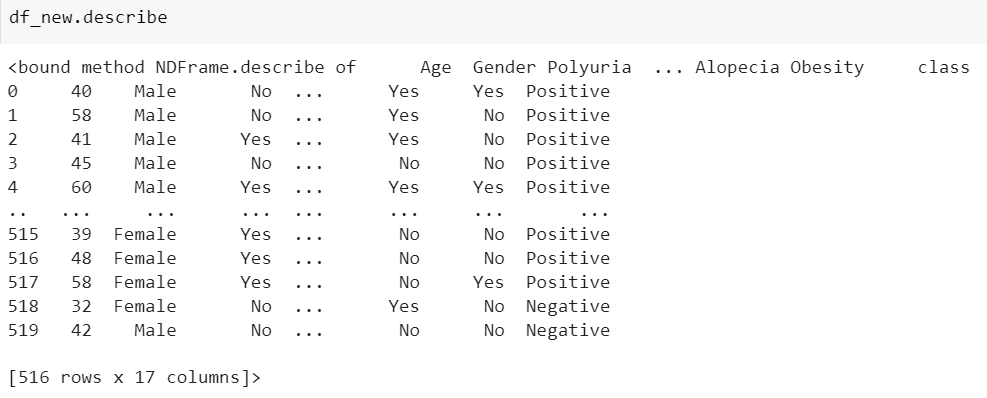
Figure



The boxplot (Figure 3) indicated that any person that was over 80 would be considered an outlier. To proceed, the outliers would be removed as part of the data cleaning. Figure 4 would be the dataset description when the outliers were removed. The dataset would be left with 516 rows after removing 4 data point with age being over or equal to 80.

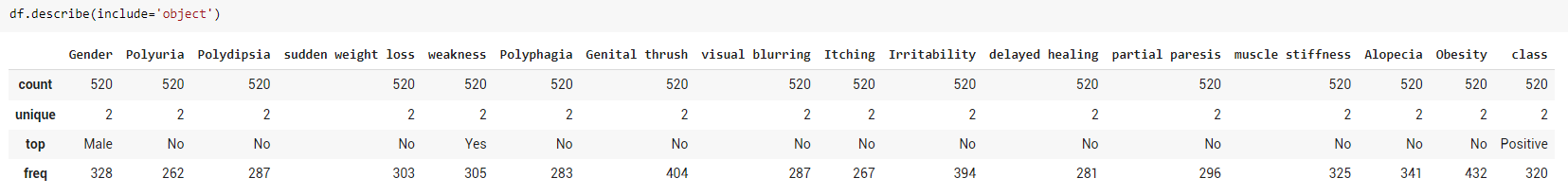


Figure



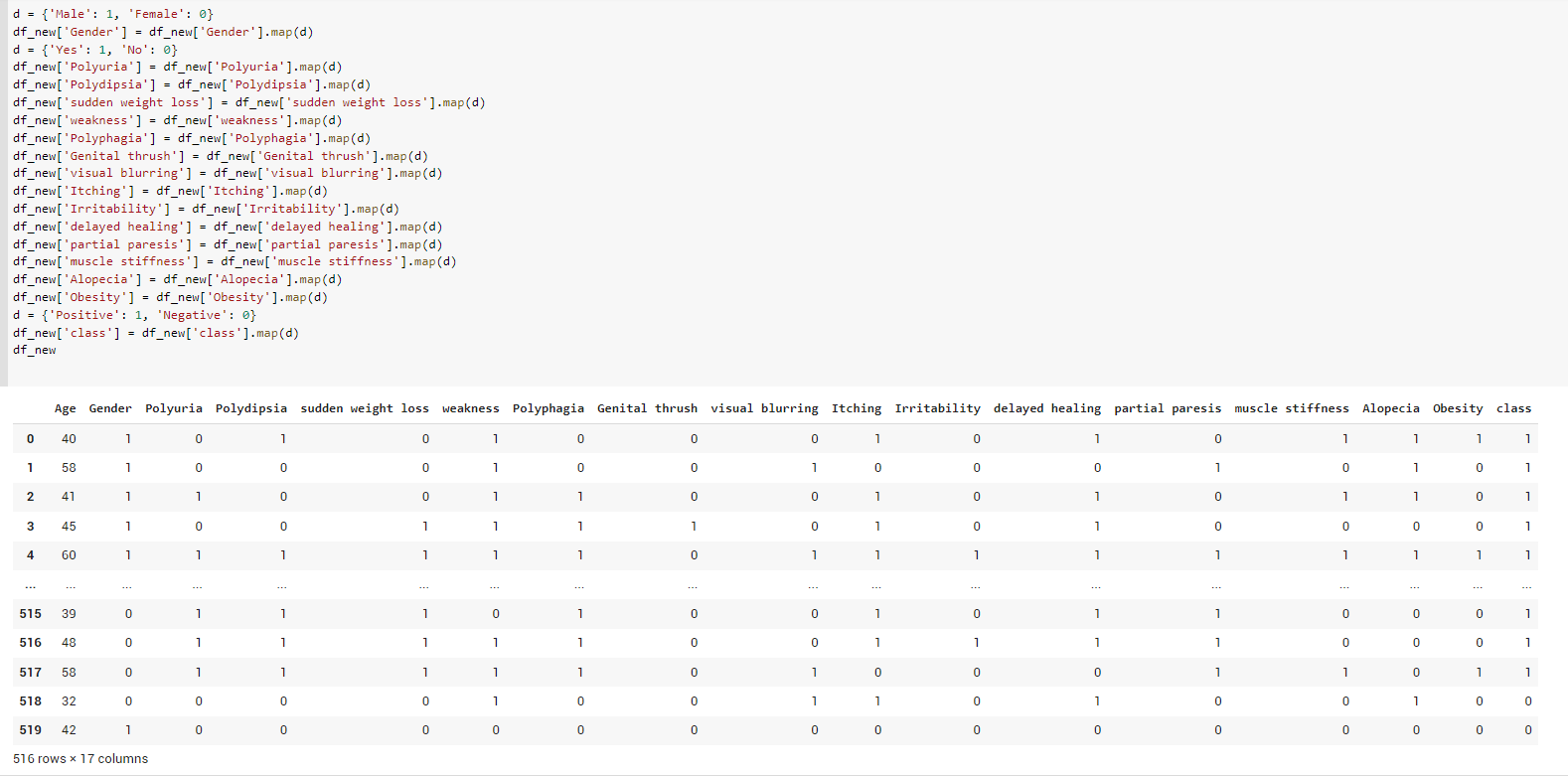
After removing the outliers, the median for age was 47, mode was 35 and mean was 47.7. The distribution of age is slightly skewed to the right. This was not the only attribute that was not normally distributed. For example, in Figure 5, gender was not as normally distributed as there were 326 males and 190 females in the dataset. Furthermore, there was about half the people in the dataset would have polyuria and itching. There were more people leaning towards not having sudden weight loss, weakness, genital thrush, irritability, obesity, stiffness and etc.

Figure



However, this does not give information on their relationship with diabetes, hence, a pair plot, heat map and boxplots would be needed. Therefore, replacing the string values in the dataset with corresponding values was needed. This would allow much more sophisticated data analysis later on when utilizing the different classification models. Value such as “Yes” would be replaced as 1 while “No” would be replaced as 0, Male would be replaced as 1 while Female would be 0 and Positive would be replaced as 1 while Negative would be replaced as 0 as shown in Figure 6.

Figure

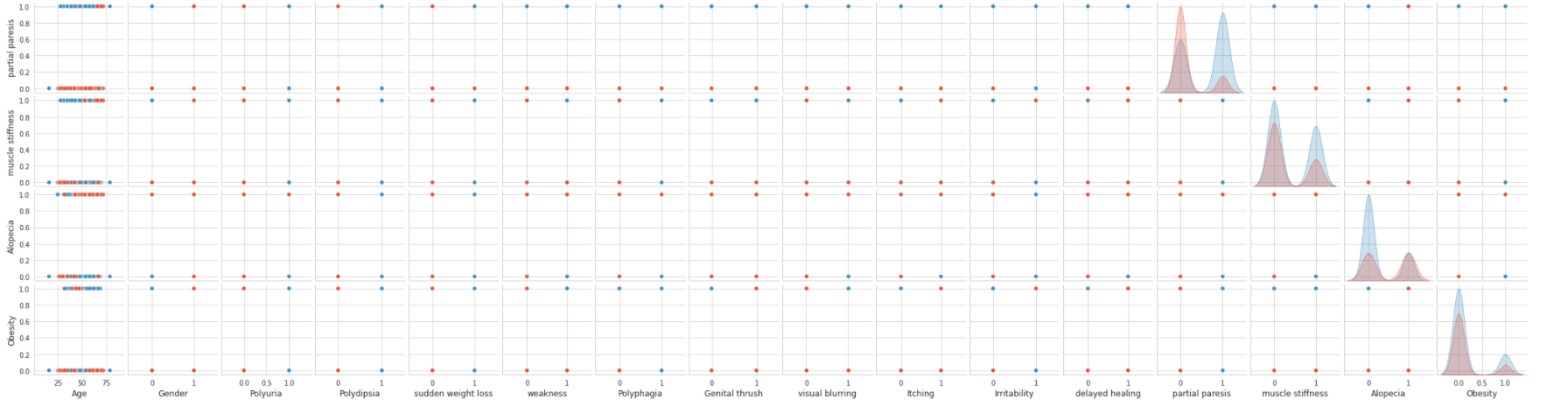


After the data conversion, a new dataset would get created and be named as df\_new. Then, the pair plot (Figure 7), heat map (Figure 8) and boxplots (Figure 9, 10, 11, 12) would be created.

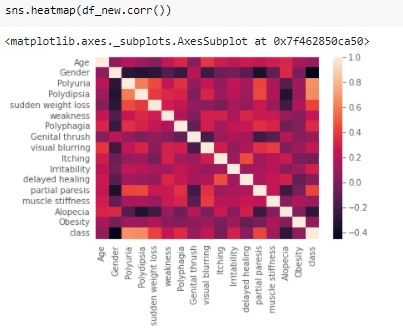
Figure



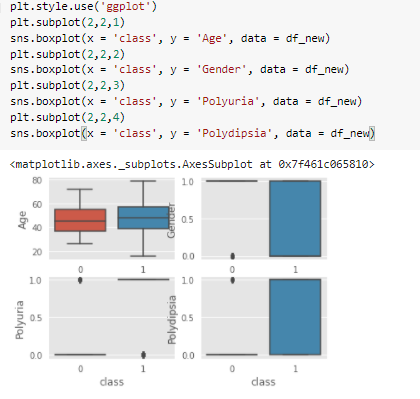




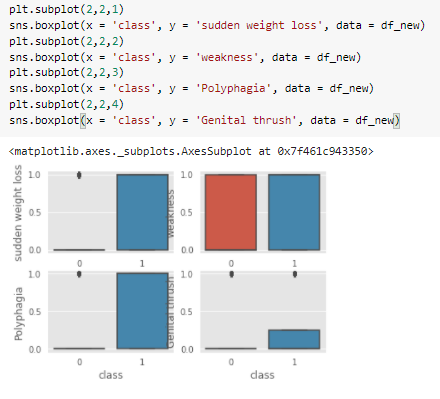
Figure



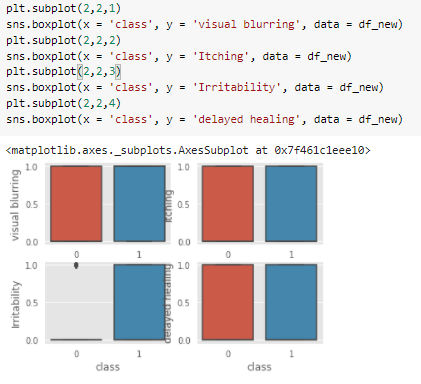
Figure



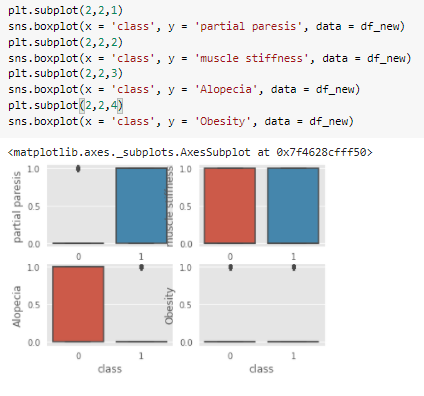
Figure



Figure



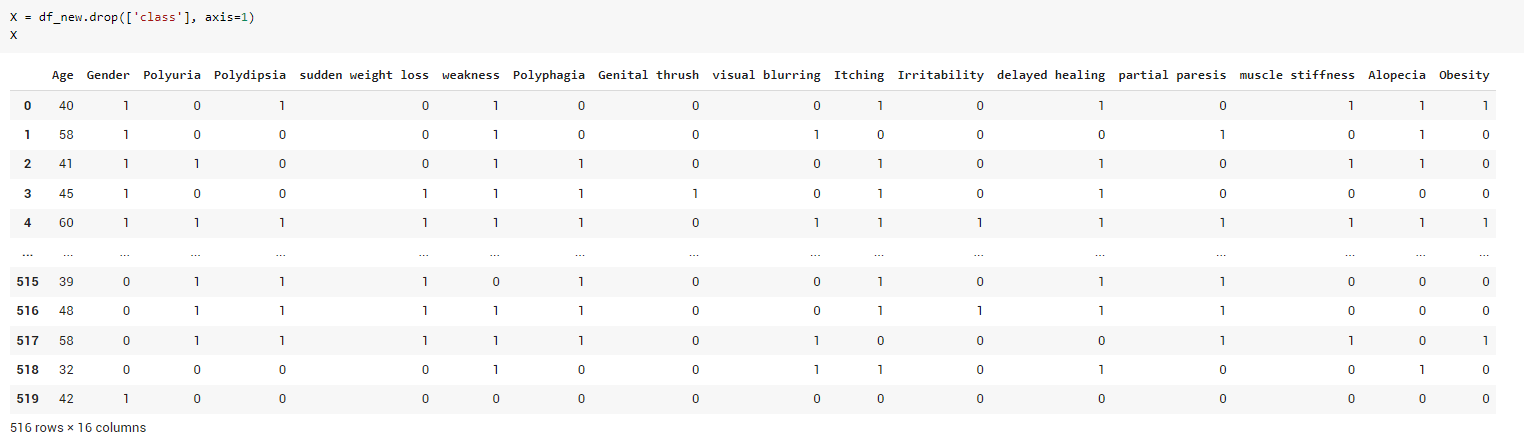
Figure

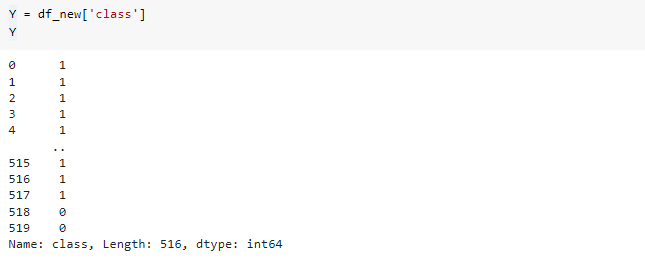


From these data visualization, there were a few attributes that showed medium strength relationship with diabetes. They were Alopecia, partial paresis, irritability, visual blurring, polyphagia, sudden weight loss, polydipsia, polyuria and gender.

Now that the data had been prepared, creating the statistical models was the next step. This would be a classification problem as this project would want to find a tool to help identify diabetes in their early stage. There were 3 methods being explored: 1. Decision Tree Classifier 2. Random Forest Classifier. 3. Logistic Regression. In order to test the performance metrics of the model, dividing up the dataset into training set and testing set was needed. Let’s set the X and Y parameters (Figure 13) for these models.

Figure





For these models, 33% would be set for the test set size and random state would be 1 (Figure 14). These parameters would be kept the same for consistency across all the models.

Figure



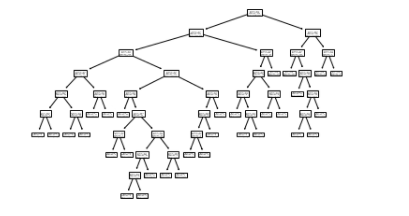
Next, import the Decision Tree Classifier (Figure 15) for the first model.

Figure



By plotting the tree, Figure 16 would get generated. That was the decision tree for the first model.

Figure



After the tree was generated, this model would be used to predict the new Y values (Figure 17).

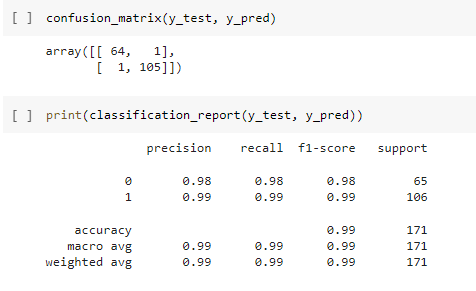
Figure



Next, testing the accuracy score, precision and recall with confusion matrix and classification report (Figure 18) was needed to understand the usefulness of the decision tree model.

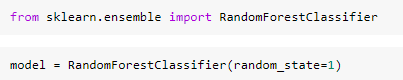


Figure



Next, import the Random Forest Classifier and create the model (Figure 19). Random Forest Classifier is an extension of a simple decision tree. It would provide the results of the best available decision tree.

Figure



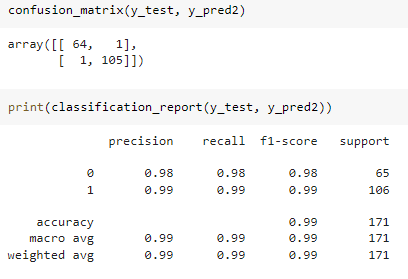
After the random forest model was generated, the model would be used to predict the new Y values (Figure 20).

Figure



Next, testing the accuracy score, precision and recall with confusion matrix and classification report (Figure 21) was needed to understand the usefulness of the random forest model.

Figure



Since the numbers are exactly the same between the first two models and since Random Forest Classifier chose the best decision tree to present the numbers, the two models generated the same tree. Therefore, only the comparison between the Decision Tree and the Logistic Regression was needed. The logistic Regression model (Figure 22) was created below.

Figure





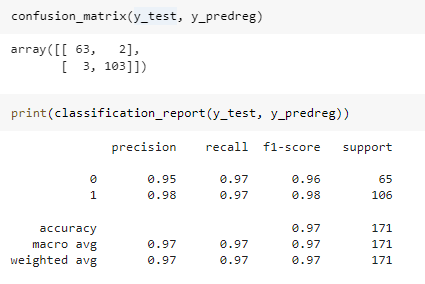
After the random forest model was generated, the model would be used to predict the new Y values (Figure 23).

Figure



Next, testing the accuracy score, precision and recall with confusion matrix and classification report (Figure 24) was needed to understand the usefulness of the random forest model.

Figure



To determine the better model to use, comparing the 2 classification reports (Figure 18 and Figure 24) from the decision tree and the logistic regression was essential. In this comparison, let’s focus on 2 key metrics – Accuracy and Recall. Accuracy would give information on the quality of the model while Recall would determine which model would be the best model to answer the research question. In this case, the accuracy of the decision tree was 99% while the accuracy of the logistic regression was 97%. This showed that the decision tree is of a higher quality model. For this research purpose, the model would be used to look for the undiagnosed diabetes patients. Trying to limit the number of false negative would be crucial. Therefore, recall rate would be looked at. The recall rate for the decision tree was 99% while the recall rate for the logistic regression was 97%. Since the decision tree was having a higher accuracy rate and higher recall rate, it was decided that the decision tree was the best statistical out of the 3 proposed methods.

#### Has anyone else done anything that is exactly the same?

There had been a lot of similar research done before on diabetes and prediabetes. These researches all had similar ideas to explore the symptoms of diabetes and to detect them in their early stage. They were not exactly the same as what this literature review would be addressing. The main difference was the method used and the attributes involved. This literature review would be using supervised machine learning technique, of which, the decision method will be conducted for classification. Other researches were also consulted to strength the support of this literature review. Some of those researches were using classification models as well while some were using regression. For example, a study done by Shield Study group (2007) called “Symptoms of Diabetes and Their Association with the Risk and Presence of Diabetes: Findings from the Study to Help Improve Early evaluation and management of risk factors leading to Diabetes” used pairwise comparisons. Shield was a 5-year longitudinal observational study of individuals with or at risk for a diagnosis of diabetes. It had three phases with an initial screening phase to draw interest to participate in the study following up with a survey to identify cases with questions about health status, health knowledge behaviours and treatments. In the last phase, four additional annual surveys was done to follow disease progression in those with diagnosed diabetes and follow the rate of transition from at risk to a diagnosis of diabetes. In addition, the pairwise comparisons were made between type 1 and type 2 diabetes and between high risk and low risk, followed by between type 2 diabetes and high and low risk. Multiple comparisons test was used to test differences across groups with different variables. To determine whether ADA symptoms were independently associated with type 2 diabetes diagnosis, multivariate stepwise logistic regression model was also constructed to adjust for other key risk factors.

A study was also done by Alexandra Garcia (2019) called “Mexican Americans' diabetes symptom prevalence, burden, and clusters”. Mexican Americans were recruited from the community to participate in the research. The data they collected included demographics, medical diseases, height, weight, quality of life and etc. They then conduct one-way ANOVAs with Bonferroni [post hoc comparisons](https://www-sciencedirect-com.ezproxy.lib.ryerson.ca/topics/nursing-and-health-professions/post-hoc-analysis) to assess differences across different groups. In addition, they also conducted different clustering methods such as hierarchical and K-means to determine the best clustering method.

Another study done by Catherinie Cowie (2010) called “Prevalence of Diabetes and High Risk for Diabetes Using A1C Criteria in the U.S. Population in 1988-2006” used A1C criteria. There was also a study done by Kenneth Heikes (2008) called “Diabetic Risk Calculator: A simple tool for detecting undiagnosed diabetes and prediabetes”. He compared both linear regression model and classification tree model for his data and he came to the decision that the classification tree performed slightly better than the logistic regression model for undiagnosed diabetes. In Kenneth’s model, he used attribute such as history of diabetes in family, medication being used, and etc which are different from the attributes in the dataset of this literature review.

## Has anyone else done anything that is related?

To continue further discussion from the previous mention on the researches consulted in this literature review, they were conducted at different locations targeting different groups of people to collect their data. In addition, their research topics are different. For example, in Shield group’s study (2007), their research topic was to examine prevalence of ADA symptoms and their association with diabetes diagnosis. Their conclusion of their research is that the occurrence of ADA symptoms alone may not be sufficient to identify those who should be diagnosed with type 2 diabetes and that they should find other combinations or other addition symptoms to be included in their next evaluation. On another research by Kenneth E Heikes (2008), his research objective was to develop a tool to calculate the probability that an individual is prediabetes or undiagnosed with diabetes. This tool was a self-administered and simple tool that anyone could use. They created the calculator using some of the attribute they collected. There was also similar research done by Catherine Cowie (2010) examining the prevalence of diabetes and undiagnosed diabetes using A1C criteria. Her final conclusion is that the prevalence of diabetes was disproportionately affected among the elderly and minority groups. While Catherine had done researches on ethnicity and its relationship with diabetes, Andrew J Karter (2013) had also done research on the prevalence of diabetes in Pacific Islanders and Asian subgroups. His conclusion was that there was a much lower rate of diabetic patients among the Chinese and several other Asian subgroups while there was a higher rate of diabetic patients among Pacific Islanders, South Asians, and Filipinos. Overall, these researches were all related to diabetes, its symptoms and its prevalence. They all gathered a lot of data and information from different sources to tackle the objective of understanding diabetes and undiagnosed diabetes.

## Where does your work fit and in with what has already gone before?

This literature review fit in what has gone before because this literature review is looking to create a decision tree model that would increase the rate of detecting diabetes in the early stage and prediabetes. It shares similar objective as the objectives of other researches. Their topics were very relatable as well as they were all focusing on understanding the symptoms and causes of diabetes. This literature review came in from a different angle to explore this topic. The dataset being used had some attributes that were not used in any researches found in the database. The dataset was donated back in July in 2020 which was fairly new. Therefore, there might be new information in there to be discovered. This literature review could determine the accuracy of a decision tree classification model on detecting diabetes using this new dataset. If the accuracy is high then it means that the attributes in the database are crucial towards the studies. Other researchers could start collecting data with those attributes and start progressing towards their own discovery. Other researchers could continue to work towards this direction and contribute more towards a greater cause of helping potential patients earlier and saving more lives.

**Why is your research worth doing in light of what has already been done?**

This research is worth doing in the light of what has already been done because there are still a lot of questions unanswered. While diabetes is getting more and more common for the general population, it is still very troublesome to deal with when it comes to detecting the undiagnosed diabetes and prediabetes. There are many factors and attribute or in combination of the two that could contribute to diabetes. For instance, living condition, mental health, genes, food, household income and social status could all become attributes towards any studies. They could be the tangibles and the intangibles. There is a reason why after many years of research on diabetes, the scientists and doctors are still continuing their investigation. It is simply because diabetes is still not sufficiently understood. The implications of undiagnosed diabetes are too devastating to ignore. Researchers need to continue to analyze it from different angles and discover more factors to be attributes in their studies. This is also what this literature review is trying to accomplish. By utilizing a fairly new dataset with new attributes, it is possible to shed some light onto the research of undiagnosed diabetes and prediabetes and help find different ways and methods to detect problems before it takes a turn for worse.

**The shortcomings of the work**

There were a few shortcomings that could potentially affect the quality of research paper and also affect the ability to answer the research question. The limitation of this research could be broken down into 4 areas which were the implementation of data collection method, sample size, time constraint and the statistical methodologies. For implementation of data collection, this dataset was collected from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh through direct questionnaires. The data was most likely to be on a lot of Bangladesh patients that lived in Bangladesh. This limited the data because of the geography, culture, and the genetics of the people would have less variability. There are also other factors such as air quality and etc could become fixed as well due to the data collected from one location only. The data was also collected from a diabetes hospital which would increase the amount of positive diagnosed people in the questionnaires. Using questionnaires to collect data could also be another limitation as questionnaires could have dishonest answers. It also lacks flexibility when it comes to the provided answers and the given questions in the questionnaires. Second of all, sample size is another limitation. This dataset only consisted 520 data points with 17 attributes. This dataset would be considered relatively small when it comes to a research paper. Therefore, the result of this research paper might not be convincing as results might change with a bigger dataset. Smaller dataset leads to a higher variability, which might lead to bias. In addition, another limitation is time constraint. The time for some parts needed to be completed for review might not be sufficient. If there was more time given, there would be opportunities to explore more datasets on diabetes and prediabetes. There would be more time to explore other resources for other research papers for references. Last but not least, it would be statistical methodologies. This would be a limitation because the current methodology this research paper was suggesting and implementing was decision tree. And decision tree has its advantage and disadvantage. Decision tree tends to lead to overfitting of the data. It could lose its generalization capabilities by generating too many new nodes. New data can cause decision tree to be altered in a big way, especially since the current dataset used in this research paper was relatively small. A new decision tree would need to be recreated and recalculated.

**Conclusion**

Based on the results from this research paper by referencing different research articles, looking for relationships among the attributes and comparing different statistical methodologies like decision tree versus logistic regression, this research paper could conclude that the model was a good tool to use when predicting diabetes in its early stage for Bangladesh people. However, as far as using it as a tool to predict diabetes in its early stage for a broader population for a more popular use, this research paper would not be sufficient to prove that yet due to a number of limitations like sample size, collection method, time constraints and methodologies. This research paper did not fully accomplish its research purpose to create a model that would help the generation population; however, the decision tree was able to use some of the attributes that were not previously used in some of the other research articles. This was a step towards understanding diabetes and prediabetes more.

GitHub Link

<https://github.com/bskhung/Data-Project.git>

**References**

*Early stage diabetes risk prediction dataset. Data Set*. UCI Machine Learning Repository: Early stage diabetes risk prediction dataset. data set. (n.d.). Retrieved October 14, 2021, from http://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.#.

*One in three Canadians is living with diabetes or prediabetes, yet knowledge of risk and complications of disease remains low*. DiabetesCanadaWebsite. (n.d.). Retrieved October 14, 2021, from https://www.diabetes.ca/media-room/press-releases/one-in-three-canadians-is-living-with-diabetes-or-prediabetes,-yet-knowledge-of-risk-and-complicatio?gclid=CjwKCAjwh5qLBhALEiwAioods7KhEpfwDkItJ7JFm3C8\_mTnwI2uX9MVtgMhPXswxqIrkAUugSiIKBoC1SEQAvD\_BwE.

Carson, April P et al.(2012), Association of Prediabetes and Diabetes with Stroke Syptoms: The Reason for Geographic and Racial Differences in Stroke study, American Diabetes Association, from <https://www.proquest.com/docview/1039306075?accountid=13631&pq-origsite=summon>

Pouya Saeedi et al.(2020), Mortality attributable to diabetes in 20-79 years old adults, 2019 estimates: Results from the International Diabetes Federation Diabetes Atlas, 9th edition, Elsevier B.V, from <https://www-sciencedirect-com.ezproxy.lib.ryerson.ca/science/article/pii/S016882272030139X>

SHIELD Study Group et al.(2007), Symptoms of Diabetes and Their Association With the Risk and Presence of Diabetes: Findings from the Study to Help Improve Early evaluation and management of risk factors Leading to Diabetes(SHIELD), American Diabetes Association, from <https://www.proquest.com/docview/223032756?accountid=13631&pq-origsite=summon>

Alexandra A. Garcia et al.(2019), Mexican Americans’ diabetes symptom prevalence, burden, and clusters, Elsevier Inc, from <https://www-sciencedirect-com.ezproxy.lib.ryerson.ca/science/article/pii/S0897189718307092>

Karter, Andrew J et al.(2013), Elevated Rates of Diabetes in Pacific Islanders and Asian SubgroupsL The Diabetes Study of Northern California (DISTANCE), American Diabetes Association, from <https://www.proquest.com/docview/1318732812?accountid=13631&pq-origsite=summon>

Heikes, Kenneth et al.(2008), Diabetes Risk Calculator: A simple tool for detecting undiagnosed diabetes and pre-diabetes, American Diabetes Association, from <https://www.proquest.com/docview/223025601?accountid=13631&pq-origsite=summon>

Cowie, Catherine et al.(2010), Prevalence of Diabetes and High Risk for Diabetes Using A1C Criteria in U.S. Population in 1988-2006, American Diabetes Association, from <https://www.proquest.com/docview/223032527?accountid=13631&pq-origsite=summon>