Big Data Analytics Project

Bruno Hung

Student #: 500457821

Supervisor: Tamer Abdou

Date: 11/8/2021

Ryerson University

Table of Contents

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. In this project, I want to explore the chance of detecting diabetes in their early stage. I want to detect it earlier so that 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 is how likely I would be able to detect diabetes in its early stage from the patients' biological traits and symptoms. 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 within the classification method. There is no missing data. The reason why I would be using the decision tree is that normalizing the dataset is not needed. It is also known to have a higher relative accuracy than many other classification algorithms. From then on, I would use the data generated to create a confusion matrix to let me know the result of my findings. 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 projected to increase from 171 million in 2000 to 366 million by 2030. There were also a lot of undiagnosed diabetes patients out there. 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. Battling diabetes was a long process. Laiteerapong also said 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 before. 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 that was being examined in this literature reviews, the dataset was collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet. 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, out of the 520 people, the oldest person was 90 while the youngest was 16. The median for age was 47.5, mode was 35 and mean was 48. The age is fairly normally distributed in the dataset. However, gender was not as normally distributed as there were 328 males and 192 females in the dataset. To further analyze the dataset, the supervised machine learning technique would be used. Afterwards, the decision tree method was being opted into when it came to the specific method while taking everything into consideration.

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 is 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. A study done by Alexandra Garcia (2019) called "Mexican Americans' diabetes symptom prevalence, burden, and clusters" used Agglomerative hierarchical and k-means clustering analyses were performed on a Gower matrix. A 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. While taking all this information into consideration, this literature review would also choose the classification model. In Kennth's model, he used attribute such as history of diabetes in family, medication being used 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?

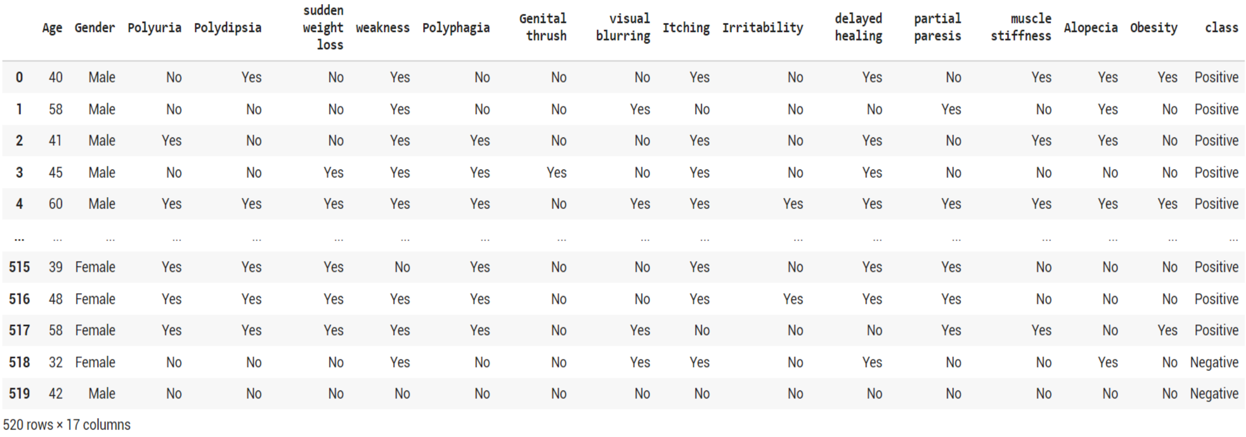
This literature review fit in what has gone before because this literature review is looking to create a decision tree model that would increase the rate of detecting diabetes in the early stage and prediabetes. It shares similar objective as the objectives of other researches. Their topics were very 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.

Descriptive Statistics

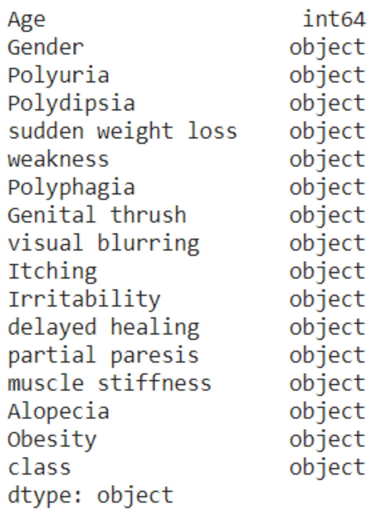
Table 1



	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
0	40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes	Positive
1	58	Male	No	No	No	Yes	No	No	Yes	No	No	No	Yes	No	Yes	No	Positive
2	41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	Yes	No	Positive
3	45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No	Positive
4	60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
515	39	Female	Yes	Yes	Yes	No	Yes	No	No	Yes	No	Yes	Yes	No	No	No	Positive
516	48	Female	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	No	No	No	Positive
517	58	Female	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No	No	Yes	Yes	No	Yes	Positive
518	32	Female	No	No	No	Yes	No	No	Yes	Yes	No	Yes	No	No	Yes	No	Negative
519	42	Male	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Negative

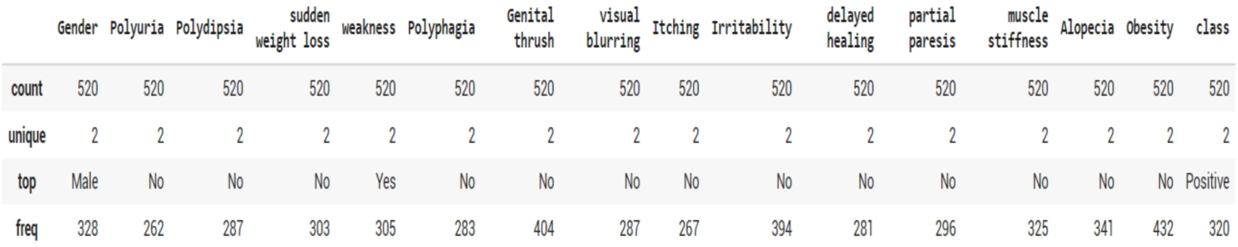
520 rows × 17 columns

Table 2



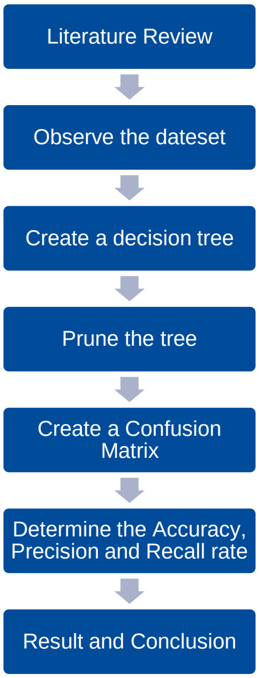
```
Age                       int64
Gender                   object
Polyuria                 object
Polydipsia               object
sudden weight loss       object
weakness                 object
Polyphagia               object
Genital thrush           object
visual blurring          object
Itching                  object
Irritability             object
delayed healing          object
partial paresis          object
muscle stiffness         object
Alopecia                 object
Obesity                  object
class                    object
dtype: object
```

Table 3



	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
count	520	520	520	520	520	520	520	520	520	520	520	520	520	520	520	520
unique	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
top	Male	No	No	No	Yes	No	No	No	No	No	No	No	No	No	No	Positive
freq	328	262	287	303	305	283	404	287	267	394	281	296	325	341	432	320

Tentative Graph of Methodology



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