Early Stage Diabetes Risk Prediction Dataset Analysis Progress report

Nathan Didier, Shehnaz Islam, Thuy Tran

York University

Data Plan:

The dataset selected was the Early Stage Diabetes Risk Prediction Dataset. As diabetes affects such a large majority of persons, with 7.3 % of Canadian populations being diabetic based of the statistics from the health fact sheet from Statics Canada. The team saw it as a beneficial opportunity create a model which could quickly and accurately predict if a patient has diabetes faster than current methods used. We believe if we can create this model persons who are in high risk of diabetes can be identified and warned earlier using the machine learning model.

Problem and Goals:

Aim(What we are looking to solve):

Our aim as group is to create a model which can predict if a patient is likely to have diabetes based on existing symptoms.

Business Questions Seeking to Answer:

We are looking to identify what are the symptoms which a strongly associated with diabetes. If these symptoms can be identified the more accurate the prediction model will be.

Expected Results:

We are expected factors such as **age , delayed healing, obesity, weakness** to be strongly associated with diabetes in patients. With less known symptoms to be partially associated with diabetes in patients. Once the associated factors were identified an accurate prediction model would then be possible to be created from this knowledge.

Dataset Summary:

The dataset contains 570 individual tuples with 17 attributes with 1 target Class Attribute. 16 out of the 17 attributes are nominal, binary attributes with 1 attribute is numeric. The following variable are correlated to the target class **Gender, Polyuria, Polydipsia, Sudden Weight Loss, Weakness, Polyphagia, Visual Blurring, Irritability, Partial Paresis, Muscle Stiffness, Alopecia**. Itching, Delayed Healing, Obesity and Genital Thrush all are uncorrelated to the target class.

Classification Application/Problem:

After discussion, the team decided to use Naïve Bayesian Classifier, Ada Boost, and Artificial Neural Network algorithms for classification applications. Coupled with ROC & AUC, Error Rate, and Speed for evaluations.

Project Schedule:

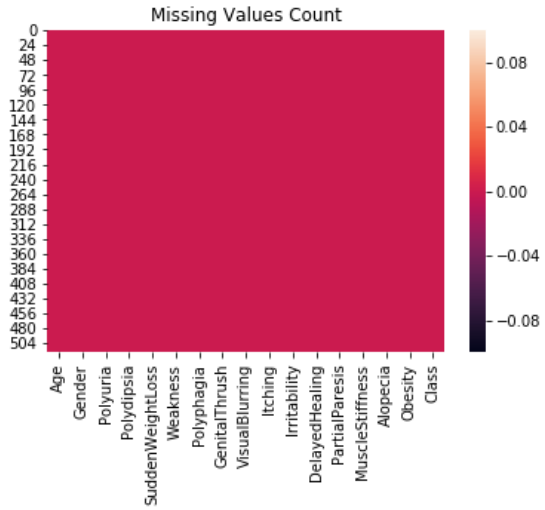
Task ID	Task Description	Task Duration	Start Date	End Date
1	First Group Meeting (Determine Project Roles and Objectives)	1	19 September 2020	19 September 2020
2	Data Cleaning	1	20 September 2020	20 September 2020
3	Second Group Meeting (Discuss Findings)	1	21 September 2020	21 September 2020
4	EDA Creation	6	22 September 2020	28 September 2020
5	Create Report for Second Checkpoint	4	28 September 2020	02 October 2020
6	Create Presentation for Second Checkpoint	2	02 October 2020	04 October 2020
7	Third Group Meeting (Discuss Model Creation Based of EDA findings)	1	12 November 2020	13 November 2020
8	Create Prediction Model	10	13 November 2020	23 November 2020
9	Fourth Group Meeting (Discuss Prediction Model)	1	23 November 2020	24 November 2020

Project Progress:

Data Cleaning:

We have completed all data cleaning preparations identifying 17 attributes: 16 categorical and 1 numeric attribute. With 520 tuples and no missing data values.

```
    ....
    ...: rawData.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 520 entries, 0 to 519
Data columns (total 17 columns):
Age                 520 non-null int64
Gender              520 non-null object
Polyuria            520 non-null object
Polydipsia          520 non-null object
SuddenWeightLoss    520 non-null object
Weakness            520 non-null object
Polyphagia          520 non-null object
GenitalThrush       520 non-null object
VisualBlurring      520 non-null object
Itching             520 non-null object
Irritability        520 non-null object
DelayedHealing      520 non-null object
PartialParesis      520 non-null object
MuscleStiffness     520 non-null object
Alopecia            520 non-null object
Obesity             520 non-null object
Class               520 non-null object
dtypes: int64(1), object(16)
memory usage: 69.1+ KB
```



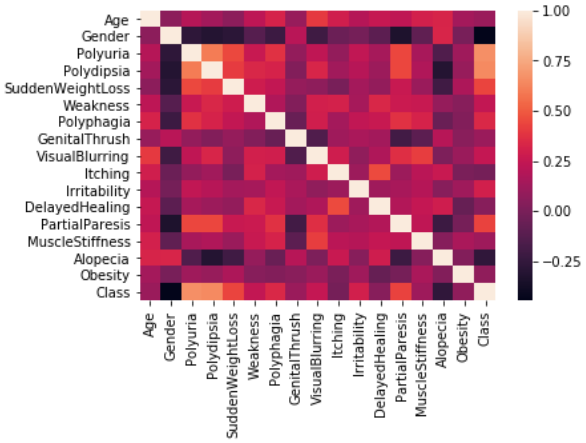
All values also are also consistent with no noisy values.

```
Male      328
Female    192
Name: Gender, dtype: int64
No      262
Yes     258
Name: Polyuria, dtype: int64
No      287
Yes     233
Name: Polydipsia, dtype: int64
No      303
Yes     217
Name: SuddenWeightLoss, dtype: int64
Yes     305
No      215
Name: Weakness, dtype: int64
No      283
Yes     237
Name: Polyphagia, dtype: int64
No      404
Yes     116
Name: GenitalThrush, dtype: int64
No      287
Yes     233
Name: VisualBlurring, dtype: int64
```

```
No      267
Yes     253
Name: Itching, dtype: int64
No      394
Yes     126
Name: Irritability, dtype: int64
No      281
Yes     239
Name: DelayedHealing, dtype: int64
No      296
Yes     224
Name: PartialParesis, dtype: int64
No      325
Yes     195
Name: MuscleStiffness, dtype: int64
No      341
Yes     179
Name: Alopecia, dtype: int64
No      432
Yes      88
Name: Obesity, dtype: int64
Positive    320
Negative    200
Name: Class, dtype: int64
```

Data Reduction:

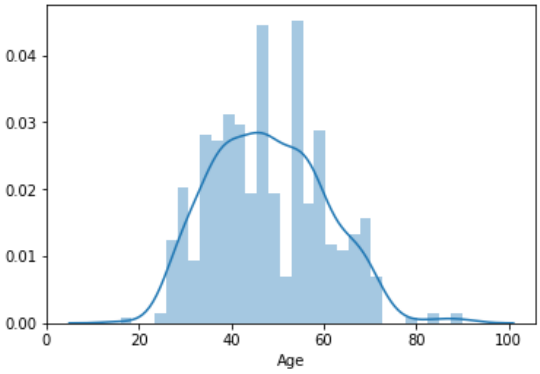
Chi-Square Tests were done to determine the correlated values and determine the unsignificant attributes. Using a significance level of 1% ‘Itching’, ‘Delayed Healing’, ‘Obesity’, and ‘Genital Thrush’ were determined as not statistically significant.

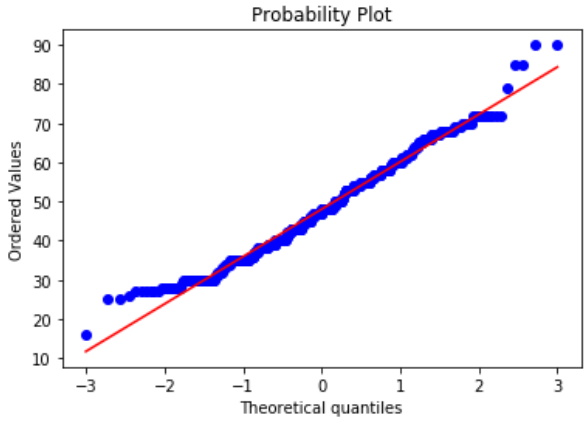



```
In [192]: for x, y in zip(list, p_value_list):
     ...:     print(str(x)+'----------'+str(y))
Gender----------3.289703730553317e-24
Polyuria----------1.7409117803442155e-51
Polydipsia----------6.1870096408863144e-49
SuddenWeightLoss----------5.969166262549937e-23
Weakness----------4.869843446585542e-08
Polyphagia----------1.1651584346409174e-14
GenitalThrush----------0.016097902991938178
VisualBlurring----------1.7015036753241226e-08
Itching----------0.8297483959485009
Irritability----------1.7714831493959365e-11
DelayedHealing----------0.32665993771439944
PartialParesis----------1.565289071056334e-22
MuscleStiffness----------0.006939095697923978
Alopecia----------1.9092794963634e-09
Obesity----------0.12710799319896815
Class----------3.498537810092432e-114
```

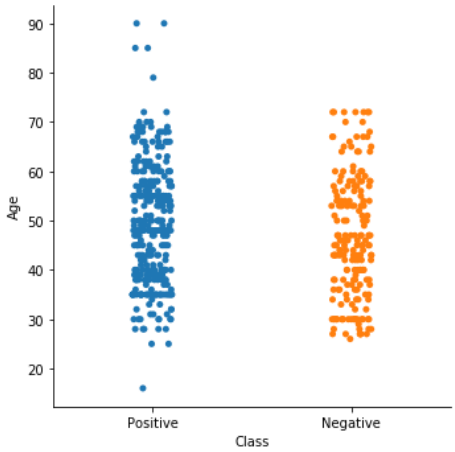
Explorative Data Analysis

During the explorative data analysis for the numeric attribute Age data distribution was normal with a skewness of 0.33 and Kurtosis of -0.19. 50% of all age values were in the range of 39 to 57. Based on the data we also found that patients above the age of 70 were more likely to be diagnosed with diabetes.

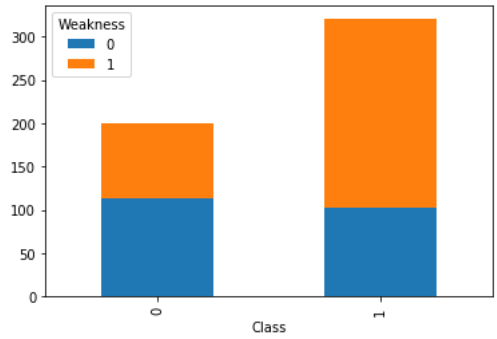


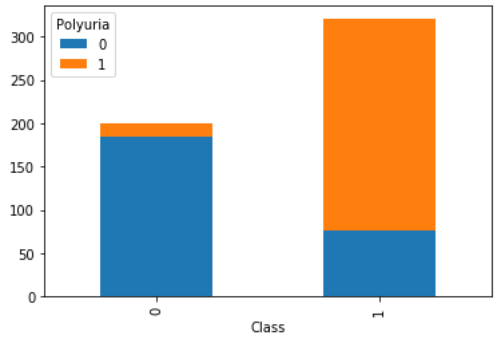


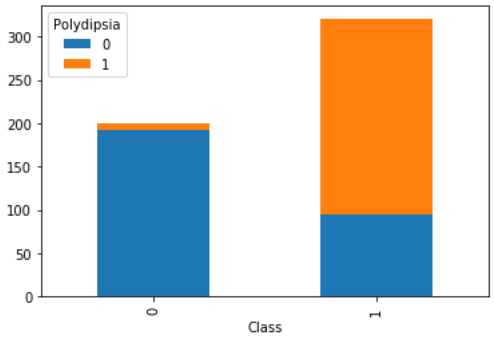
During the explorative da utes we found that despite males were counted fem We learnt that there

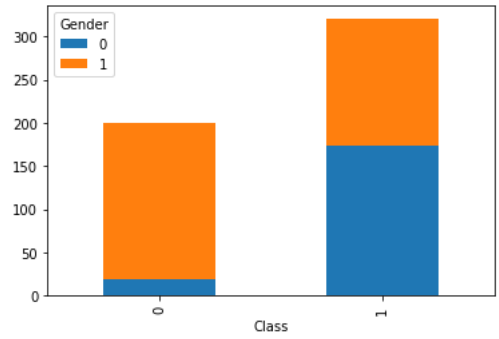


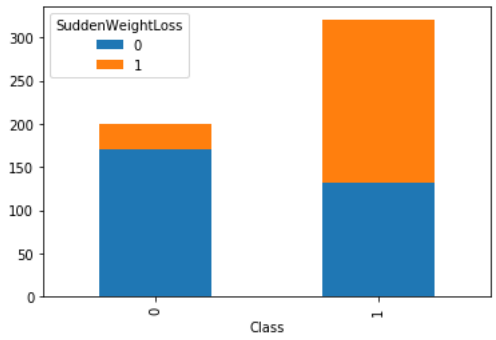
were 120 more diabetic cases over non-diabetic cases. Diabetic patients were also more to experience Polyuria, Sudden Weight Loss, Weakness, and Polydipsia.

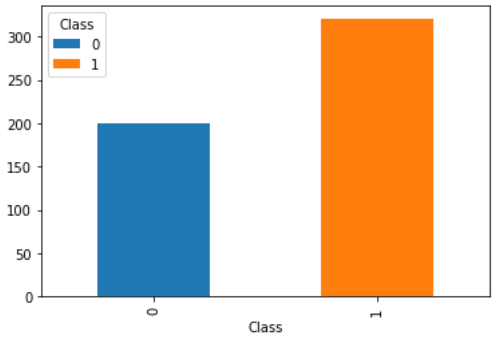




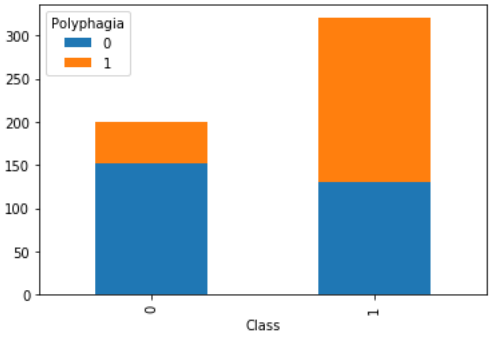


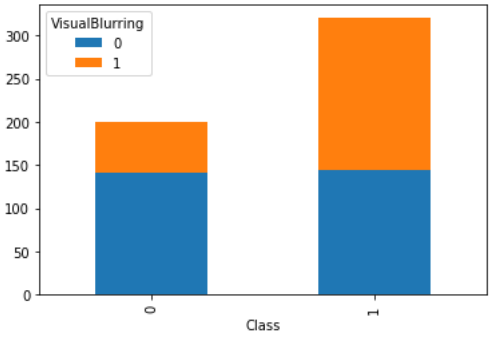


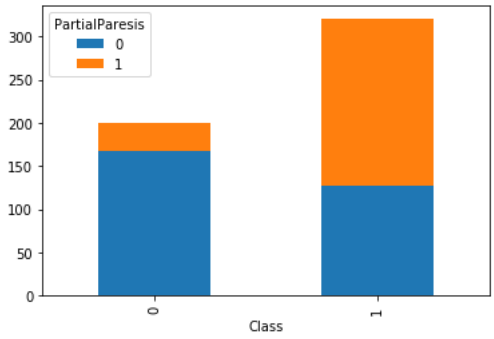


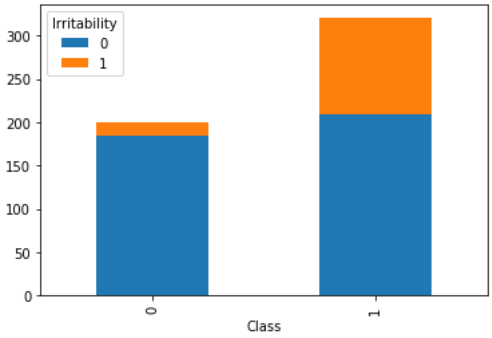


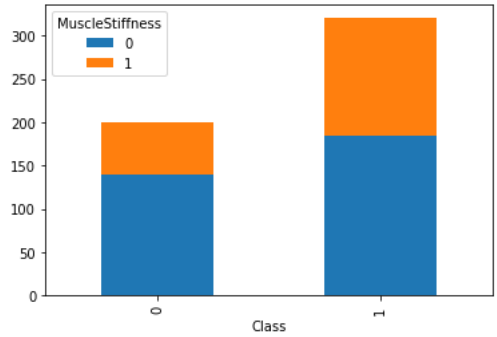
It was also found that diabetic patients were less likely to experience Alopecia than non-diabetic patient and more likely to experience Visual Blurring, Polyphagia, Muscle Stiffness, and partial Paresis.

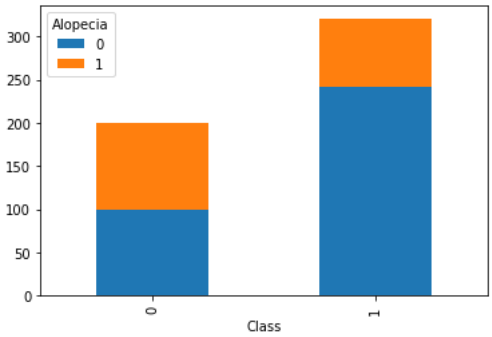












Data Transformation

Due to the fact machine learning algorithms cannot be used on categorical data directly the data was converted into numbers. Allowing us to perform algorithms on the new indicator variable created from the categorical variables.

Difficulties and Issues:

- Finding the description of certain attributes within the dataset.
- Determining if we had enough data to make an accurate prediction model.
- Uncertainty of the accuracy of the dataset due to misleading dataset descriptions.
- Learn new programming language, Python.
- Insufficient understanding about health problems, i.e. Diabetes
    - Symptoms leading to diabetes
    - Different stages of diabetes

Solutions:

Additional research was done regarding the attributes in question. An attribute definition document was created shared amongst members to gain better understand of what each attribute represents.

We assumed that the dataset was collected using proper data collection practices and confirmed that all the data is accurate after the pre-processing of the data was done.

Data Camps, W3C, Python Documentation, and Class Notes were all used for training team members to better at using python for completing this project.