UNIVERSIDAD DE MONTERREY

ESCUELA DE INGENIERIA Y TECNOLOGIAS

INTELIGENCIA ARTIFICIAL

Logistic Classification

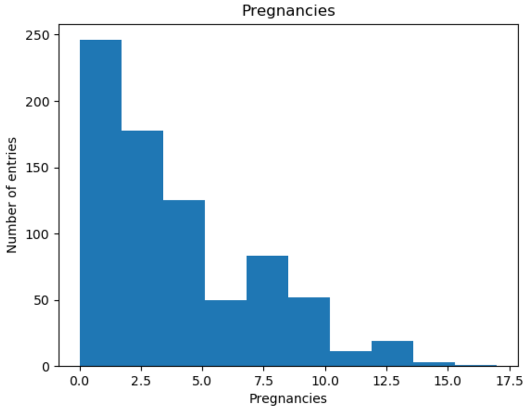
STUDENT NAME: Bernardo Cárdenas Domene

ID: 509883

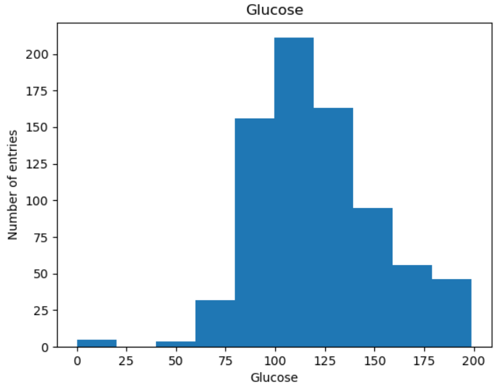
April 20, 2020, San Pedro Garza García, Nuevo León

Logistic classification consists of using a logistic function as a hypothesis function since it produces values between 0 and 1. This way the values from the function can be used for classification where values greater than 0.5 can be classed to a positive class and values lower than 0.5 to a negative class. For this project, this method was applied to a diabetes database in order to be able to predict whether a person has diabetes or not. The dataset contains column values of: number of pregnancies, glucose levels, blood pressure, skin thickness in triceps, insulin levels, BMI, diabetes pedigree function and age with an outcome of 1 if the person has diabetes or 0 if the person does not have diabetes. In order to explore the data, histograms for each column have been created for easy visualization of patterns or trends in the data can be seen. The histograms generated were:

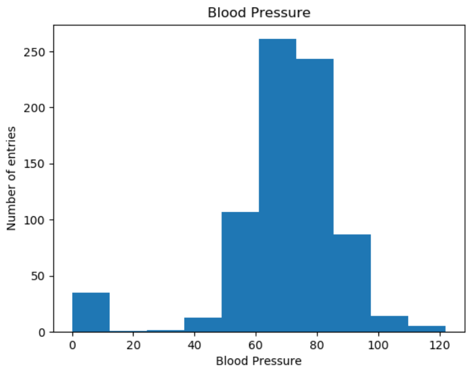
Number of pregnancies: Most rows have 0 pregnancies, and it is less frequent where people have many pregnancies.



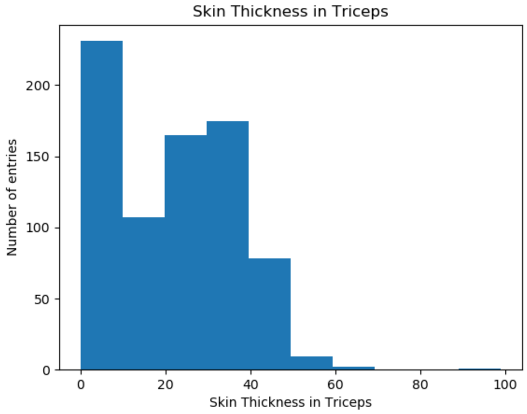
Glucose: A normal distribution can be seen where most people have a level between 100 and 125



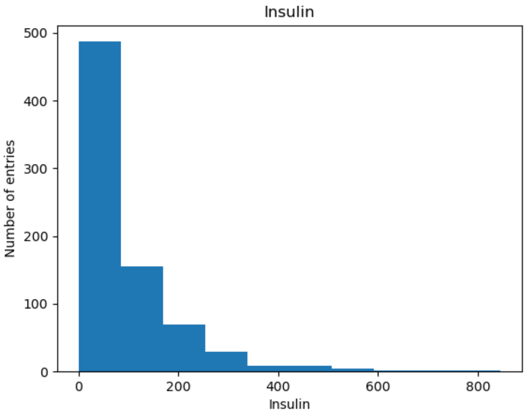
Blood pressure: A normal distribution can be seen in values 60-80. It can also be seen that values for blood pressure 0 are most likely data that is missing.



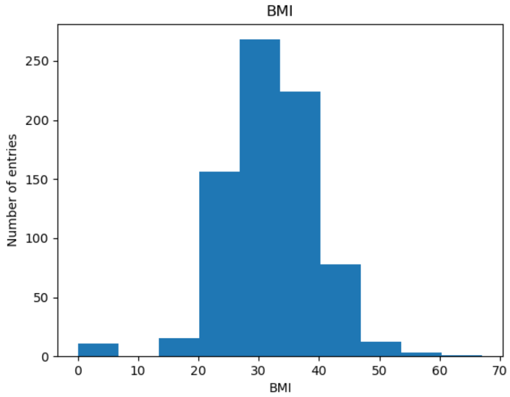
Skin Thickness in Triceps: Here we can see a large part of the data is missing with zero values. The rest of the data shows a normal distribution between values 30-40.



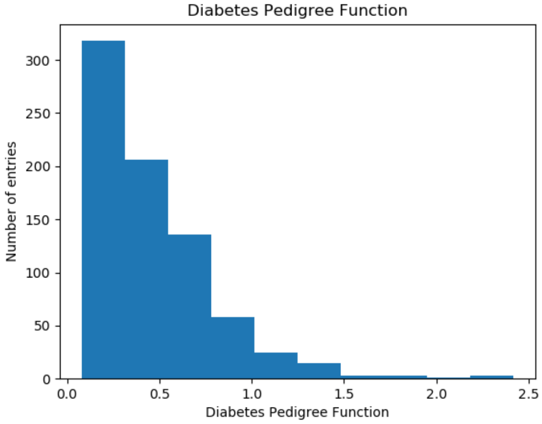
Insulin: Here we can also see a large part of the data has zero values which may indicate that it is missing data.



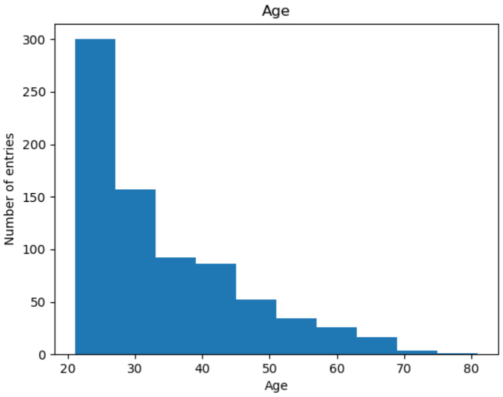
BMI: This histogram shows a normal distribution between values 30-35 which indicates many entries are people classified by the BMI as overweight.



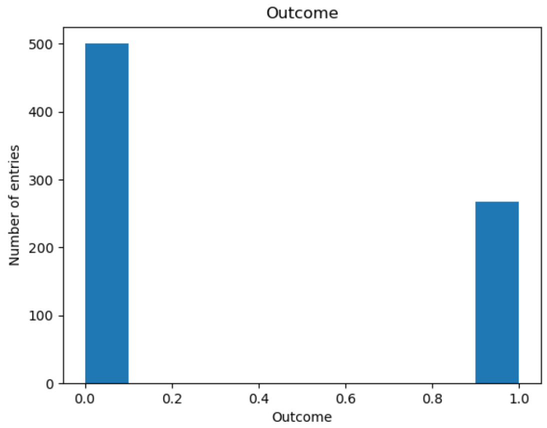
Diabetes Pedigree Function: We can see that smaller values are more common as to higher values.



Age: The graph shows that most people in the data are young between 20-35 years old.



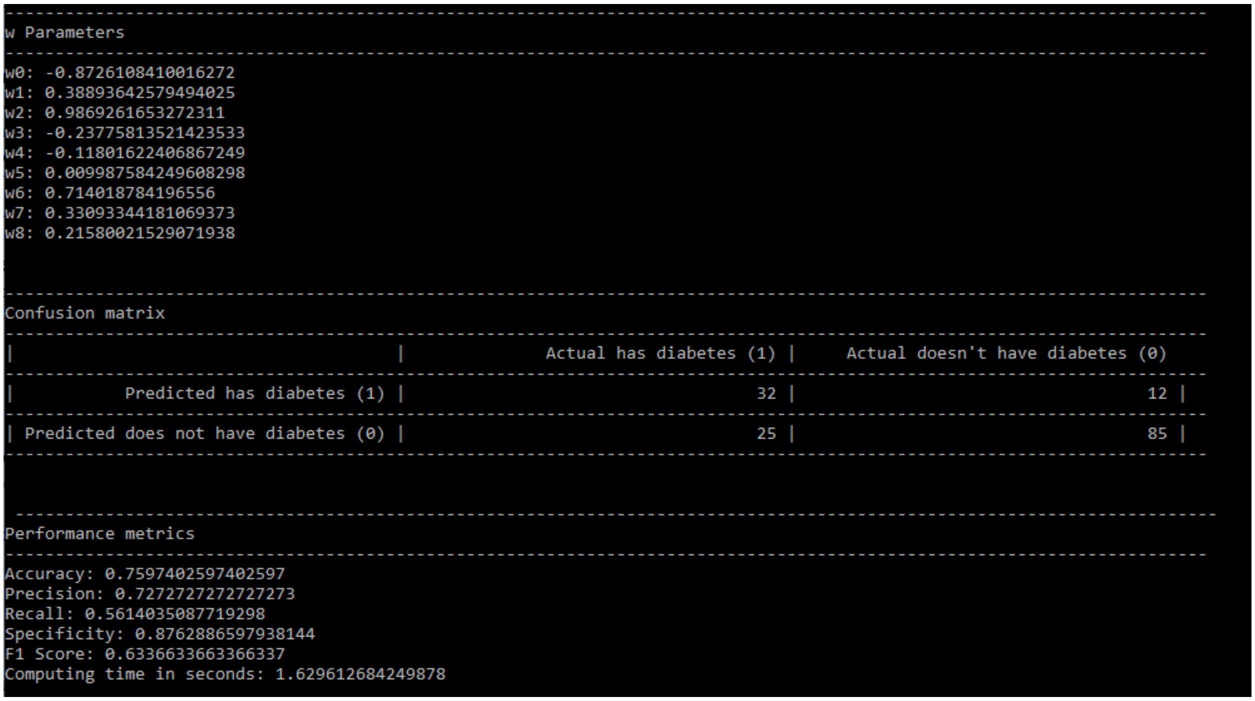
Outcome: We can see that most entries are zeros close to 2/3 being zeros and 1/3 being ones.



After understanding the data from the data sheet, the implementation of a python program for logistic classification was done. First, the training data was split into 2 parts where 80% of the data was used for training and 20% for testing. This was done in order to use the training data to obtain w parameters and then use the testing data to see the results. This process also selects data at random from the data sheet.

The algorithm for obtaining optimal w parameters was done by using the gradient descent algorithm where a logistic hypothesis function was obtained from the data followed by obtaining the gradient of cost function and then applying it to the gradient descent algorithm using a provided learning rate. This process also calculates the L2 norm and repeats itself until this value is less than the provided stopping criteria, so the error is minimized. Finally, optimal w parameters are obtained using the training data. Although scaling data is not necessary for logistic classification, it was done in this program in order to speed up the execution time.

Afterwards, the testing data is used with the obtained w parameters in order to predict weather each row belongs to the positive class (has diabetes) or negative class (does not have diabetes). This was done by applying the logistic function to the training data multiplied by the w parameters. The result is an array of predictions where these were modified by if the values were greater than 0.5 they were replaced by 1 indicating the person has diabetes, and if the value were less than 0.5 they were replaced by 0 indicating the person does not have diabetes. After obtaining the predictions, a confusion matrix was generated using these. Finally using this matrix, performance metrics were calculated in order to determine the performance and accuracy of the algorithm. The results using a learning rate of 0.001 and stopping criteria of 0.01 were:



```
------------------------------------------------------------------------------------------------------------------------
w Parameters
------------------------------------------------------------------------------------------------------------------------
w0: -0.8726108410016272
w1: 0.38893642579494025
w2: 0.9869261653272311
w3: -0.23775813521423533
w4: -0.11801622406867249
w5: 0.009987584249608298
w6: 0.714018784196556
w7: 0.33093344181069373
w8: 0.21580021529071938


------------------------------------------------------------------------------------------------------------------------
Confusion matrix
------------------------------------------------------------------------------------------------------------------------
|                                     |               Actual has diabetes (1) |     Actual doesn't have diabetes (0)
------------------------------------------------------------------------------------------------------------------------
|          Predicted has diabetes (1) |                                    32 |                                    12 |
------------------------------------------------------------------------------------------------------------------------
| Predicted does not have diabetes (0) |                                    25 |                                    85 |
------------------------------------------------------------------------------------------------------------------------


 ------------------------------------------------------------------------------------------------------------------------
Performance metrics
------------------------------------------------------------------------------------------------------------------------
Accuracy: 0.7597402597402597
Precision: 0.7272727272727273
Recall: 0.5614035087719298
Specificity: 0.8762886597938144
F1 Score: 0.6336633663366337
Computing time in seconds: 1.629612684249878
```

Due to the random selection of data from the datasheet, each run of the program shows different results, however the values obtained are very similar and the accuracy varies between 70 and 80 percent. Analyzing the data we can see that from the w parameters we can identify which characteristics (columns) affect the most the outcome of the prediction, for example w2 and w6 show the largest values which means that glucose and BMI seem to affect greatly in the decision to indicate if a person has diabetes or not.

The accuracy performance metric tells us that the algorithm is on average 75% effective in predicting if a person has diabetes or not using the data provided however, as shown in the outcome histogram, most of the data in the datasheet had an outcome of zeros greater than the ones. Since the data does not have balanced classes this metric may not be accurate in describing the performance of the algorithm. In the precision metric we can see that 72% of positive predictions were actually positive and a recall of 56% indicating which predictions were predicted as positive when they were actually positive. The recall value tells us that a large part of the data actually had diabetes, but the algorithm only classified 56% of them this way. Specificity tells us that 87% of predictions were correct as in they were predicted as negative when in reality, they were negative. The F1 score tells us overall how the algorithm performed, in this case we see a value of 0.64.

After analyzing the data, an accuracy ranging between 70-80% was achieved and an F1 score between 0.55 and -0.7 using a learning rate of 0.001 and stopping criteria of 0.01. Modifying these values to a smaller learning rate results in a longer computing time, but a more precise value of accuracy at 78% and F1 score of 0.68. For future works, aspects such as having no missing values in the datasheet would improve the performance metrics in the algorithm.

Honor Code:

I hereby declare that I have done this activity with academic integrity.