UDEM

UNIVERSIDAD DE MONTERREY

ESCUELA DE INGENIERIA Y TECNOLOGIAS

INTELIGENCIA ARTIFICIAL

# K Nearest Neighbours

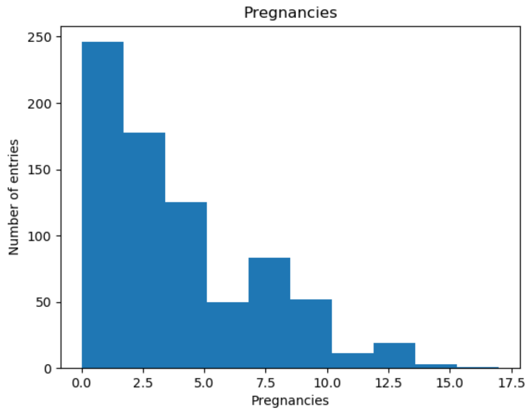
STUDENT NAME: Bernardo Cárdenas Domene

ID: 509883

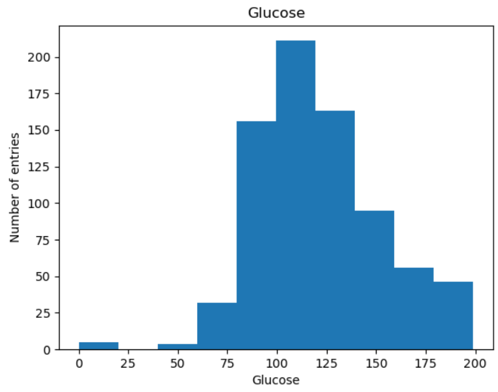
May 9, 2020, San Pedro Garza García, Nuevo León

The K Nearest Neighbours algorithm for classification consists of given a testing point, calculating the nearest k data points in the training data and assigning the class of the testing point to the majority of the nearest data points found. This is done by calculating the Euclidian distance between the point and the x data and determining which class of these is majority by calculating the conditional probability. The highest probability is the one assigned to the testing point in order to predict which class it is most likely to belong to. 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:

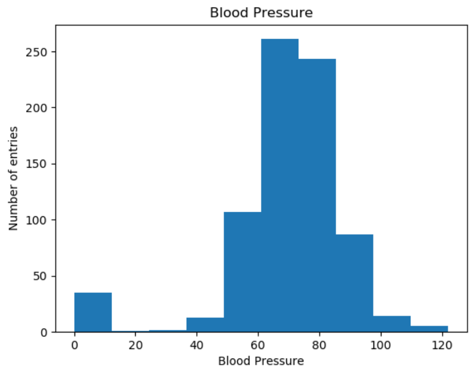
Number of pregnancies: Most rows have 0 pregnancies, and it is less frequent where people have many pregnancies.



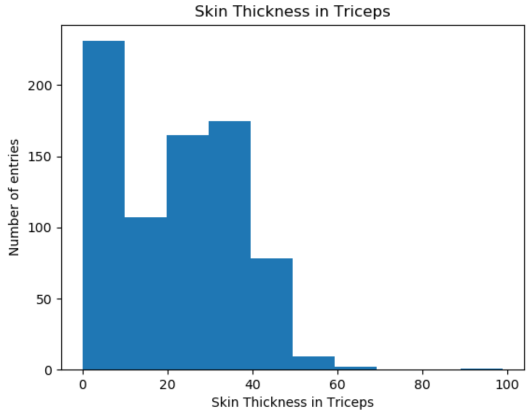
Glucose: A normal distribution can be seen where most people have a level between 100 and 125



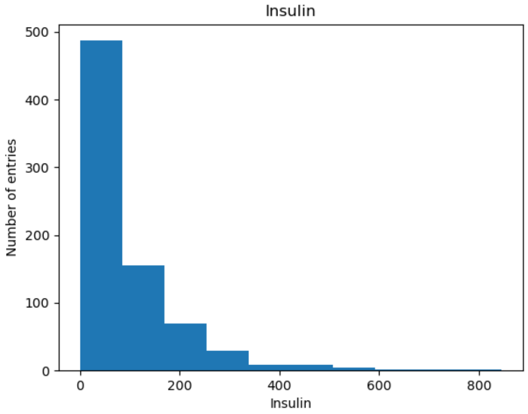
Blood pressure: A normal distribution can be seen in values 60-80. It can also be seen that values for blood pressure 0 are most likely data that is missing.



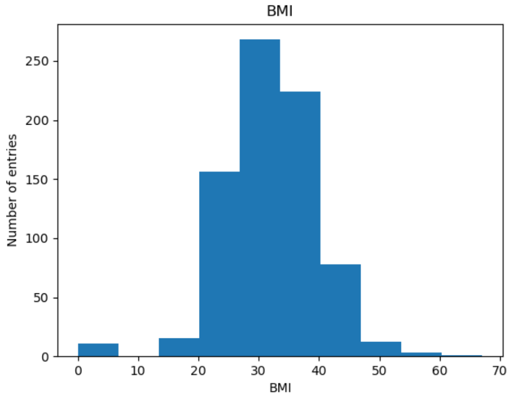
Skin Thickness in Triceps: Here we can see a large part of the data is missing with zero values. The rest of the data shows a normal distribution between values 30-40.



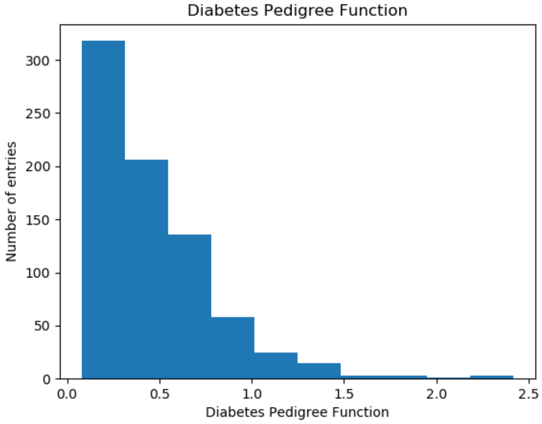
Insulin: Here we can also see a large part of the data has zero values which may indicate that it is missing data.



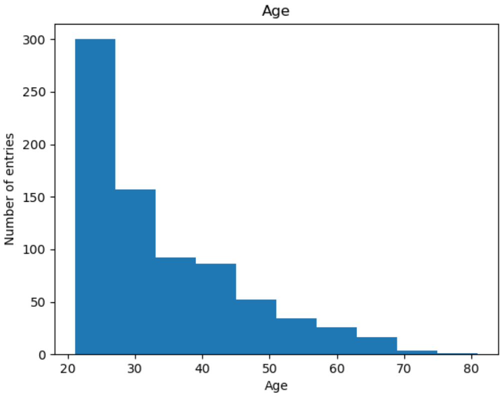
BMI: This histogram shows a normal distribution between values 30-35 which indicates many entries are people classified by the BMI as overweight.



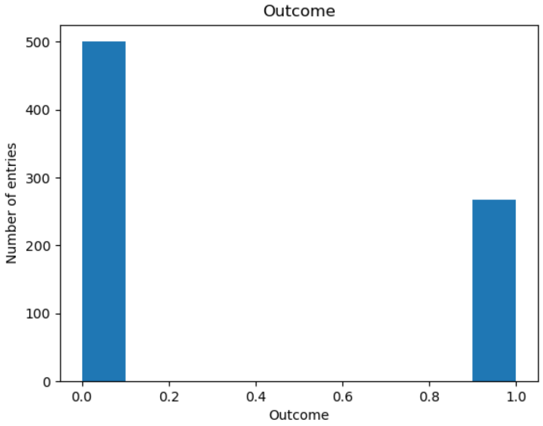
Diabetes Pedigree Function: We can see that smaller values are more common as to higher values.



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



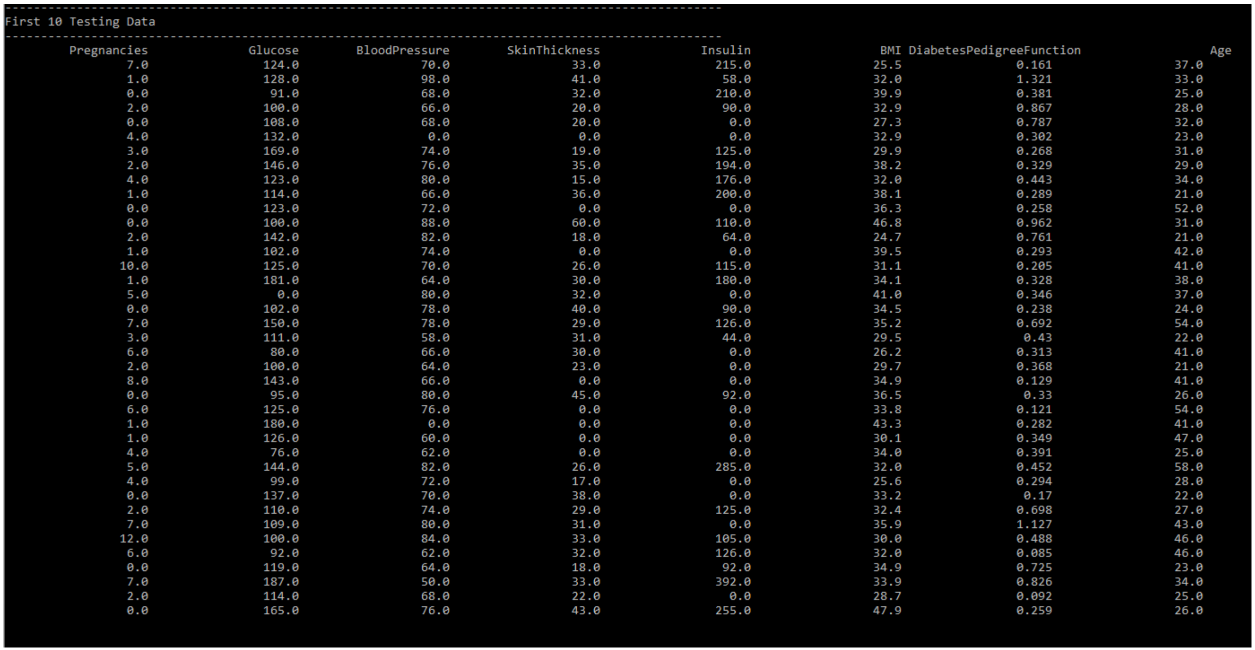
Outcome: We can see that most entries are zeros close to 2/3 being zeros and 1/3 being ones.



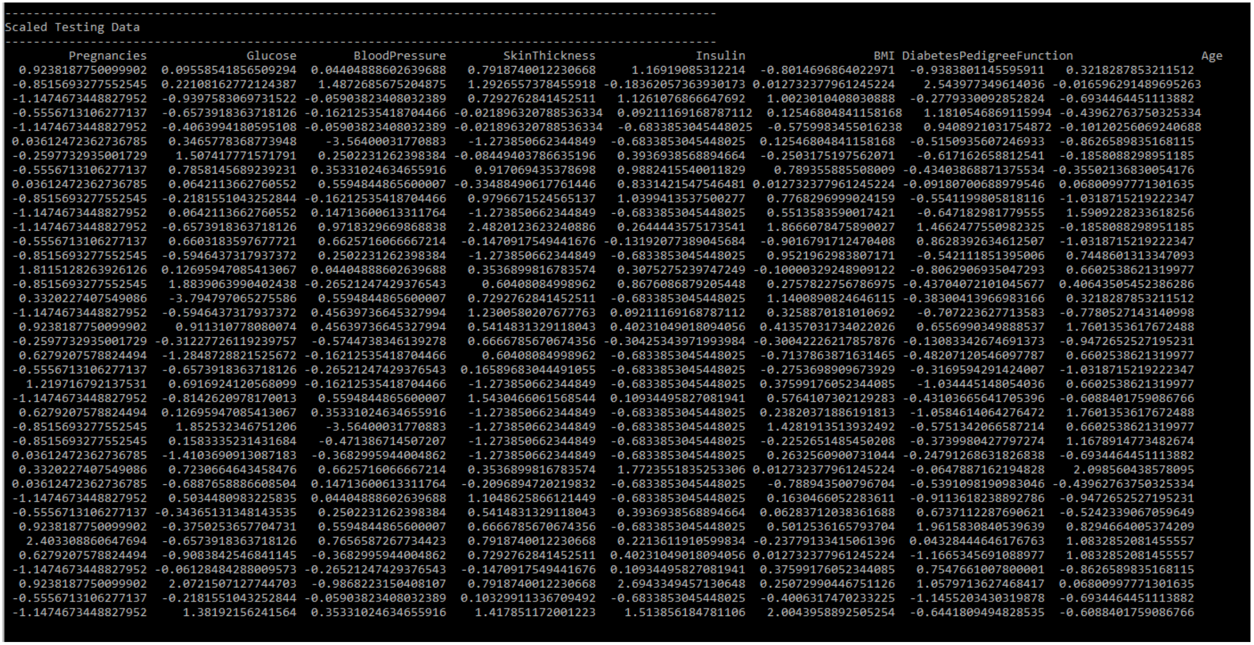
After understanding the data from the data sheet, the implementation of a python program for k nearest neighbors was done. First, the training data was split into 2 parts in a random manner where 95% of the data was used for training and 5% for testing. This was done in order to use the training data to train and then use the testing data to predict and see the results. Afterwards, the testing and training data was scaled so that its values range between 0 and 1.

The algorithm was done by iterating over each x testing data in order to predict if it is most likely to belong to the positive or negative class using the predict method. In each iteration, the predict method calculates the Euclidean distance and conditional probabilities in order to determine this. First the Euclidean distance between the testing point and every x training data is calculated using the compute_euclidean_distance method and then using compute_conditional_probabilities the nearest k elements are found to determine which class is majority and assigning it to the test point. This process repeats until every x testing data's y value has been predicted.

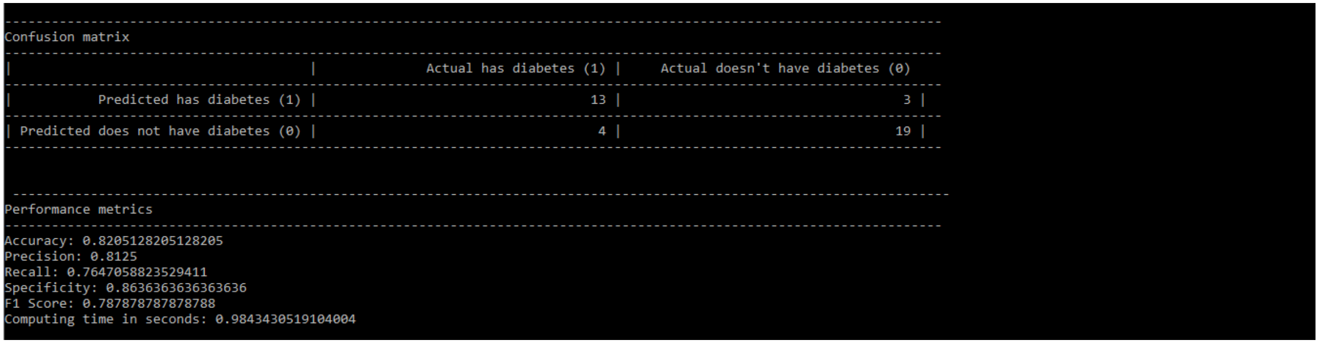
Once calculated the predictions, a confusion matrix was generated using get_confusion_matrix method comparing the predictions to the actual values from y testing data. Finally using this matrix, performance metrics were calculated in order to determine the performance and accuracy of the algorithm. An example of the execution of the program using a k value of 5 is:



```
------------------------------------------------------------------------------------------
First 10 Testing Data
------------------------------------------------------------------------------------------
         Pregnancies              Glucose        BloodPressure         SkinThickness              Insulin                  BMI DiabetesPedigreeFunction                   Age
                 7.0                124.0                 70.0                  33.0                215.0                 25.5                    0.161                  37.0
                 1.0                128.0                 98.0                  41.0                 58.0                 32.0                    1.321                  33.0
                 0.0                 91.0                 68.0                  32.0                210.0                 39.9                    0.381                  25.0
                 2.0                100.0                 66.0                  20.0                 90.0                 32.9                    0.867                  28.0
                 0.0                108.0                 68.0                  20.0                  0.0                 27.3                    0.787                  32.0
                 4.0                132.0                  0.0                   0.0                  0.0                 32.9                    0.302                  23.0
                 3.0                169.0                 74.0                  19.0                125.0                 29.9                    0.268                  31.0
                 2.0                146.0                 76.0                  35.0                194.0                 38.2                    0.329                  29.0
                 4.0                123.0                 80.0                  15.0                176.0                 32.0                    0.443                  34.0
                 1.0                114.0                 66.0                  36.0                200.0                 38.1                    0.289                  21.0
                 0.0                123.0                 72.0                   0.0                  0.0                 36.3                    0.258                  52.0
                 0.0                100.0                 88.0                  60.0                110.0                 46.8                    0.962                  31.0
                 2.0                142.0                 82.0                  18.0                 64.0                 24.7                    0.761                  21.0
                 1.0                102.0                 74.0                   0.0                  0.0                 39.5                    0.293                  42.0
                10.0                125.0                 70.0                  26.0                115.0                 31.1                    0.205                  41.0
                 1.0                181.0                 64.0                  30.0                180.0                 34.1                    0.328                  38.0
                 5.0                  0.0                 80.0                  32.0                  0.0                 41.0                    0.346                  37.0
                 0.0                102.0                 78.0                  40.0                 90.0                 34.5                    0.238                  24.0
                 7.0                150.0                 78.0                  29.0                126.0                 35.2                    0.692                  54.0
                 3.0                111.0                 58.0                  31.0                 44.0                 29.5                     0.43                  22.0
                 6.0                 80.0                 66.0                  30.0                  0.0                 26.2                    0.313                  41.0
                 2.0                100.0                 64.0                  23.0                  0.0                 29.7                    0.368                  21.0
                 8.0                143.0                 66.0                   0.0                  0.0                 34.9                    0.129                  41.0
                 0.0                 95.0                 80.0                  45.0                 92.0                 36.5                     0.33                  26.0
                 6.0                125.0                 76.0                   0.0                  0.0                 33.8                    0.121                  54.0
                 1.0                180.0                  0.0                   0.0                  0.0                 43.3                    0.282                  41.0
                 1.0                126.0                 60.0                   0.0                  0.0                 30.1                    0.349                  47.0
                 4.0                 76.0                 62.0                   0.0                  0.0                 34.0                    0.391                  25.0
                 5.0                144.0                 82.0                  26.0                285.0                 32.0                    0.452                  58.0
                 4.0                 99.0                 72.0                  17.0                  0.0                 25.6                    0.294                  28.0
                 0.0                137.0                 70.0                  38.0                  0.0                 33.2                     0.17                  22.0
                 2.0                110.0                 74.0                  29.0                125.0                 32.4                    0.698                  27.0
                 7.0                109.0                 80.0                  31.0                  0.0                 35.9                    1.127                  43.0
                12.0                100.0                 84.0                  33.0                105.0                 30.0                    0.488                  46.0
                 6.0                 92.0                 62.0                  32.0                126.0                 32.0                    0.085                  46.0
                 0.0                119.0                 64.0                  18.0                 92.0                 34.9                    0.725                  23.0
                 7.0                187.0                 50.0                  33.0                392.0                 33.9                    0.826                  34.0
                 2.0                114.0                 68.0                  22.0                  0.0                 28.7                    0.092                  25.0
                 0.0                165.0                 76.0                  43.0                255.0                 47.9                    0.259                  26.0
```



```
------------------------------------------------------------------------------------------
Scaled Testing Data
------------------------------------------------------------------------------------------
         Pregnancies              Glucose        BloodPressure         SkinThickness              Insulin                  BMI DiabetesPedigreeFunction                   Age
  0.9238187750099902  0.09558541856509294  0.04404888602639688    0.7918740012230668     1.16919085312214  -0.8014696864022971      -0.9383801145595911    0.3218287853211512
 -0.8515693277552545  0.22108162772124387   1.4872685675204875    1.2926557378455918 -0.18362057363930173 0.012732377961245224        2.543977349614036 -0.016596291489695263
 -1.1474673448827952  -0.9397583069731522 -0.05903823408032389    0.7292762841452511   1.1261076866647692   1.0023010408030888      -0.2779330092852824   -0.6934464451113882
 -0.5556713106277137  -0.6573918363718126 -0.16212535418704466 -0.021896320788536334  0.09211169168787112  0.12546804841158168       1.1810546869115994  -0.43962763750325334
 -1.1474673448827952  -0.4063994180595108 -0.05903823408032389 -0.021896320788536334  -0.6833853045448025  -0.5759983455016238       0.9408921031754872  -0.10120256069240688
 0.03612472362736785   0.3465778368773948    -3.56400031770883    -1.273850662344849  -0.6833853045448025  0.12546804841158168      -0.5150935607246933   -0.8626589835168115
 -0.2597732935001729    1.507417771571791   0.2502231262398384  -0.08449403786635196   0.3936938568894664  -0.2503175197562071       -0.617162658812541   -0.1858088298951185
 -0.5556713106277137   0.7858145689239231  0.35331024634655916     0.917069435378698   0.9882415540011829    0.789355885508009     -0.43403868871375534  -0.35502136830054176
 0.03612472362736785   0.0642113662760552   0.5594844865600007  -0.33488490617761446   0.8331421547546481 0.012732377961245224     -0.09180700688979546   0.06800997771301635
 -0.8515693277552545  -0.2181551043252844 -0.16212535418704466    0.9796671524565137   1.0399413537500277   0.7768296999024159      -0.5541199805818116   -1.0318715219222347
 -1.1474673448827952   0.0642113662760552  0.14713600613311764    -1.273850662344849  -0.6833853045448025   0.5513583590017421       -0.647182981779555    1.5909228233618256
 -1.1474673448827952  -0.6573918363718126   0.9718329669868838    2.4820123623240886   0.2644443575173541   1.8666078475890027       1.4662477550982325   -0.1858088298951185
 -0.5556713106277137   0.6603183597677721   0.6625716066667214   -0.1470917549441676 -0.13192077389045684  -0.9016791712470408       0.8628392634612507   -1.0318715219222347
 -0.8515693277552545  -0.5946437317937372   0.2502231262398384    -1.273850662344849  -0.6833853045448025   0.9521962983807171        -0.54211851395006    0.7448601313347093
  1.8115128263926126  0.12695947085413067  0.04404888602639688    0.3536899816783574   0.3075275239747249 -0.10000329248909122      -0.8062906935047293    0.6602538621319977
 -0.8515693277552545   1.8839063990402438 -0.26521247429376543      0.60408084998962   0.8676086879205448   0.2757822756786975     -0.43704072101045677   0.40643505452386286
  0.3320227407549086   -3.794797065275586   0.5594844865600007    0.7292762841452511  -0.6833853045448025   1.1400890824646115     -0.38300413966983166    0.3218287853211512
 -1.1474673448827952  -0.5946437317937372  0.45639736645327994    1.2300580207677763  0.09211169168787112   0.3258870181010692       -0.707223627713583   -0.7780527143140998
  0.9238187750099902    0.911310778080074  0.45639736645327994    0.5414831329118043  0.40231049018094056  0.41357031734022026       0.6556990349888537    1.7601353617672488
 -0.2597732935001729 -0.31227726119239757  -0.5744738346139278    0.6666785670674356 -0.30425343971993984 -0.30042226217857876     -0.13083342674691373   -0.9472652527195231
  0.6279207578824494  -1.2848728821525672 -0.16212535418704466      0.60408084998962  -0.6833853045448025  -0.7137863871631465     -0.48207120546097787    0.6602538621319977
 -0.5556713106277137  -0.6573918363718126 -0.26521247429376543   0.16589683044491055  -0.6833853045448025  -0.2753698909673929      -0.3169594291424007   -1.0318715219222347
   1.219716792137531   0.6916924120568099 -0.16212535418704466    -1.273850662344849  -0.6833853045448025  0.37599176052344085        -1.03445148054036    0.6602538621319977
 -1.1474673448827952  -0.8142620978170013   0.5594844865600007    1.5430466061568544  0.10934495827081941   0.5764107302129283     -0.43103665641705396   -0.6088401759086766
  0.6279207578824494  0.12695947085413067  0.35331024634655916    -1.273850662344849  -0.6833853045448025  0.23820371886191813      -1.0584614064276472    1.7601353617672488
 -0.8515693277552545    1.852532346751206    -3.56400031770883    -1.273850662344849  -0.6833853045448025   1.4281913513932492      -0.5751342066587214    0.6602538621319977
 -0.8515693277552545   0.1583335231431684   -0.471386714507207    -1.273850662344849  -0.6833853045448025  -0.2252651485450208      -0.3739980427797274    1.1678914773482674
 0.03612472362736785  -1.4103690913087183  -0.3682995944004862    -1.273850662344849  -0.6833853045448025   0.2632560900731044     -0.24791268631826838   -0.6934464451113882
  0.3320227407549086   0.7230664643458476   0.6625716066667214    0.3536899816783574   1.7723551835253306 0.012732377961245224      -0.0647887162194828     2.098560438578095
 0.03612472362736785  -0.6887658886608504  0.14713600613311764   -0.2096894720219832  -0.6833853045448025   -0.788943500796704      -0.5391098190983046  -0.43962763750325334
 -1.1474673448827952   0.5034480983225835  0.04404888602639688    1.1048625866121449  -0.6833853045448025   0.1630466052283611      -0.9113618238892786   -0.9472652527195231
 -0.5556713106277137 -0.34365131348143535   0.2502231262398384    0.5414831329118043   0.3936938568894664  0.06283712038361688       0.6737112287690621   -0.5242339067059649
  0.9238187750099902  -0.3750253657704731   0.5594844865600007    0.6666785670674356  -0.6833853045448025   0.5012536165793704       1.9615830840539639    0.8294664005374209
   2.403308860647694  -0.6573918363718126   0.7656587267734423    0.7918740012230668   0.2213611910599834 -0.23779133415061396      0.04328444646176763    1.0832852081455557
  0.6279207578824494  -0.9083842546841145  -0.3682995944004862    0.7292762841452511  0.40231049018094056 0.012732377961245224      -1.1665345691088977    1.0832852081455557
 -1.1474673448827952 -0.06128484288009573 -0.26521247429376543   -0.1470917549441676  0.10934495827081941  0.37599176052344085       0.7547661007800001   -0.8626589835168115
  0.9238187750099902   2.0721507127744703  -0.9868223150408107    0.7918740012230668   2.6943349457130648  0.25072990446751126       1.0579713627468417   0.06800997771301635
 -0.5556713106277137  -0.2181551043252844 -0.05903823408032389   0.10329911336709492  -0.6833853045448025  -0.4006317470233225      -1.1455203430319878   -0.6934464451113882
 -1.1474673448827952     1.38192156241564  0.35331024634655916     1.417851172001223    1.513856184781106   2.0043958892505254      -0.6441809494828535   -0.6088401759086766
```



```
------------------------------------------------------------------------------------------------------------------------
Confusion matrix
------------------------------------------------------------------------------------------------------------------------
|                                       |                   Actual has diabetes (1) |      Actual doesn't have diabetes (0)
------------------------------------------------------------------------------------------------------------------------
|             Predicted has diabetes (1) |                                        13 |                                      3 |
------------------------------------------------------------------------------------------------------------------------
| Predicted does not have diabetes (0) |                                         4 |                                     19 |
------------------------------------------------------------------------------------------------------------------------



  ------------------------------------------------------------------------------------------------------------------------
Performance metrics
------------------------------------------------------------------------------------------------------------------------
Accuracy: 0.8205128205128205
Precision: 0.8125
Recall: 0.7647058823529411
Specificity: 0.8636363636363636
F1 Score: 0.787878787878788
Computing time in seconds: 0.9843430519104004
```

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 75 and 85 percent.

The following table shows performance metrics obtained using different K values:

| **K** | **Accuracy** | **Precision** | **Recall or Sensitivity** | **Specificity** | **F-1 Score** |
|---|---|---|---|---|---|
| 5 | 0.820 | 0.812 | 0.764 | 0.864 | 0.789 |
| 10 | 0.744 | 0.727 | 0.533 | 0.875 | 0.615 |
| 20 | 0.769 | 0.786 | 0.647 | 0.864 | 0.710 |

The table shows that a k value of 5 has the best results with the highest f-1 score. Executing the program it was also noted that using a k value of 10 produced the greatest variation in the performance metrics, where k values of 5 and 20 were more constant.

On average, the accuracy performance metric tells us that the algorithm is on average around 78% 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 have been affected when using larger k values since more of these zeros might be considered as the closest elements. In the precision metric we can see that around 77% of positive predictions were actually positive and a recall of 65% 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 65% 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 an average value of around 0.7.

After analyzing the data, comparing the k nearest neighbor’s method against logistic classification for the same problem we can see that both achieve similar results. Based on the results, k nearest neighbors shows slightly better performance metrics than logistic classification, but has greater variation in these results. Modifying the k value showed that it affected the results but in this specific case of problem, a k value of 5 showed the best results. For future works, aspects such as having no missing values in the datasheet might improve the performance metrics in the algorithm since these could create a bias when finding the nearest neighbors.

**Honor Code:**

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