**Steps taken:**

1. It consists of 768 row \* 9 columns
2. All the columns are made of continuous non-binary data sets except the outcome which is made of discrete binary data set.   
   The target or our dependent variable is the outcome, while the rest columns are features/independent variables.
3. This is a summary of the data provided.



1. Checked for null values and found none.
2. Checked for zero values in “BloodPressure”, “Glucose”, “SkinThickness”, “BMI” and “Insulin” and replace by the median since it is not normal for a living human being to have any of those variables zero.
3. Outliers were checked in every column except in the Outcome (bec it is the target) and it was treated.
4. The distribution of the data then was understood using graphs, the following findings were made:
   * There are high number allocated at zer0, one and two pregnancies, then decline
   * Most of data is allocated between 85 -145 glucose
   * the data is concentrated at 75 but it is moderate at 60, 65 and 85.
   * most of the data is concentrate at 22 skin thinckness
   * Also here most of the data concentrate at 25 insulin and few more is concentrate at 255 insulin while the rest of the data is distrusted between the two values.
   * The highest concentration of BMI is at 32.5 and then the decrease going away from this point.
   * There are high number of people that score lower than 0.4 (diabetes pedigree function) and even drop almost the half at score of 0.78 but it pick up again at 1.2.
   * A high number of sample ages between 20 and 30 and the rest varies between 40 -60
5. Afterwards, correlations were examine and represented by heatmap; the lighter the color, the stronger the relation.
   * The glucose level (from all the variables) has the strongest relation with the outcome. While the skin thickness has the weakest relationship.
   * It was also found that age and number of pregnancies have a moderate relationship, that’s why it would not be treated as multi-collinearity case therefore no need to drop any of the features.
6. Different models were used which are KNN, DT & SVC. Those specific models were chosen since it issue in hand is supervised and classification.
7. The results:



Precision measures the proportion of positively predicted labels that are actually correct Recall measure the proportion of positively predicted labels out of actual positives.  
F1-score represent the model score as a function of precision and recall score; giving equal weight to them.

SVC scored the highest score in the recall, making it the best model in this case

For Tableau:

https://public.tableau.com/app/profile/sameriah.girgis