Summary of project

Datasets:

1. Primary dataset - health care diabetes.csv

Tasks:

Week 1

Data Exploration:

1. Perform descriptive analysis. Understand the variables and their corresponding values. On the columns below, a value of zero does not make sense and thus indicates missing value:
    - Glucose
    - BloodPressure
    - SkinThickness
    - Insulin
    - BMI
2. Visually explore these variables using histograms. Treat the missing values accordingly.
3. There are integer and float data type variables in this dataset. Create a count (frequency) plot describing the data types and the count of variables.

Data Exploration:

4. Check the balance of the data by plotting the count of outcomes by their value. Describe your findings and plan future course of action.
5. Create scatter charts between the pair of variables to understand the relationships. Describe your findings.
6. Perform correlation analysis. Visually explore it using a heat map.

Week 2

Data Modeling:

1. Devise strategies for model building. It is important to decide the right validation framework. Express your thought process.
2. Apply an appropriate classification algorithm to build a model.

3. Compare various models with the results from KNN algorithm.
4. Create a classification report by analyzing sensitivity, specificity, AUC (ROC curve), etc.

Please be descriptive to explain what values of these parameter you have used.

Data Reporting:

5. Create a dashboard in tableau by choosing appropriate chart types and metrics useful for the business. The dashboard must entail the following:
    - Pie chart to describe the diabetic or non-diabetic population
    - Scatter charts between relevant variables to analyze the relationships
    - Histogram or frequency charts to analyze the distribution of the data
    - Heatmap of correlation analysis among the relevant variables
    - Create bins of these age values: 20-25, 25-30, 30-35, etc. Analyze different variables for these age brackets using a bubble chart.

Steps involved:

1. Pandas library has been used to import data from datafile and performing exploratory data analysis and removing rows with null values.

2. Histogram shows skewed distribution for Pregnancies, Insulin, Age and Diabetes Pedigree Function.

3. Higher values of Pregnancies (>7), Glucose (>124), Blood Pressure (>82), Skin Thickness (>32), Insulin (>125), Age (>31) have higher changes of having diabetes.

4. Higher values of following pairs together have high chances of having diabetes:

    a. Insulin and Glucose.
    b. Age and Glucose.

5. Lower value of Insulin and higher value of Glucose have high chances of having diabetes. Similarly, lower value of Diabetes Pedigree Function and higher value of Glucose have high chances of having diabetes.

6. Heatmap is generated for correlation coefficients between independent variables.

7. Data is split into test and train set and RandomForestClassifier has been used for modelling. Scoring for test data has also been done.

8. The results from RandomForestClassifier model has been compared with KNN model using scoring metric and AUC curve.

9. Dashboard for tableau analysis has been created.