Project 2: Healthcare

Problem Statement:

- The objective is to predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset.
- Build a model to accurately predict whether the patients in the dataset have diabetes or not.

Dataset: Healthcare data.csv

Project Task:

- Data Exploration:
  1. Perform Descriptive Analysis –

     Understanding different variables and their corresponding values. Finding missing values, if any. Replacing a value of zero with the variables mean or median accordingly.

  2. Visually explore these variables using histograms.
  3. Create a count (frequency) plot describing the data types and the count of variables.
  4. Check the balance of the data by plotting the count of outcomes by their value. Describe your findings

     - The countplot shows the values of outcome variable int the dataset. 0

       represents 'no diabetes' whereas 1 represents 'Diabetes'.

     -The dataset is having Outcome variable (label) values as 0 with count 500 and 1 with count 268. It is an imbalanced data to get analyzed.

     - We have to balance the dataset before Data Classification using Imbalanced dataset technique such as SMOTE. This technique will be applied on training and testing dataset.

5. Create scatter charts between the pair of variables to understand the relationships.
6. Perform correlation analysis.
    - A function corr() is used to find the correlation between variables in the dataset.
    - A heatmap is used to visualize the correlation pattern in the dataset.

- Data Modelling

1. Model Building Strategy –

   -The dataset consists of Outcome variable which is a Categorical Variable. 0 represents 'No diabetes' and 1 represents 'Diabetes' for the Outcome column in the dataset.

2. Classification Model - Classification Model that will be applied is Logistic Regression as the Outcome Variable is Binary Categorical Variable.
3. Compare various models with the results from KNN algorithm.

   Other Classification Models:

   - Decision Tree
   - Random Forest Classifier
   - SVM (Support Vector Machine)
   - KNN (K- Nearest Neighbours)

4. Create a classification report by analyzing sensitivity, specificity, AUC (ROC curve), etc.

   Classification Report:

Classification Model	Accuracy	Sensitivity	AUC (ROC Curve)
Logistic Regression (Without SMOTE)	48%	69%	-
Logistic Regression (With SMOTE)	71%	70%	0.83
Decision Tree	57%	67%	0.93
Random Forest Classifier	**72%**	**74%**	**0.98**
SVM	65%	0%	-
KNN	62%	67%	0.86

- Data Reporting:

  Create a dashboard in tableau by choosing appropriate chart types and metrics useful for the business. The dashboard must entail the following:

  1. Pie chart to describe the diabetic or non-diabetic population
  2. Scatter charts between relevant variables to analyze the relationships
  3. Histogram or frequency charts to analyze the distribution of the data
  4. Heatmap of correlation analysis among the relevant variables
  5. Create bins of these age values: 20-25, 25-30, 30-35, etc. Analyze different variables for these age brackets using a bubble chart.

  Link to Tableau dashboard:
  https://public.tableau.com/app/profile/ruhi.nehri/viz/HealthCareCapstaoneProject/FinalReportDsshboard?publish=yes