Problem statement:

Use Decision Trees to prepare a model on fraud data treating those who have taxable income <= 30000 as "Risky" and others are "Good"

Column names	Type of data	Relevance
Undergrad	Binary	a person is under-graduated or not
Marital.Status	categorical	marital status of a person
Taxable.Income	Continuous	Taxable income is the amount of how much tax an individual owes to the government
Work Experience	Continuous	Work experience of an individual person
Urban	binary	Whether that person belongs to the urban area or not
City population	continuous	Number of people in the city by category

Data Pre-processing

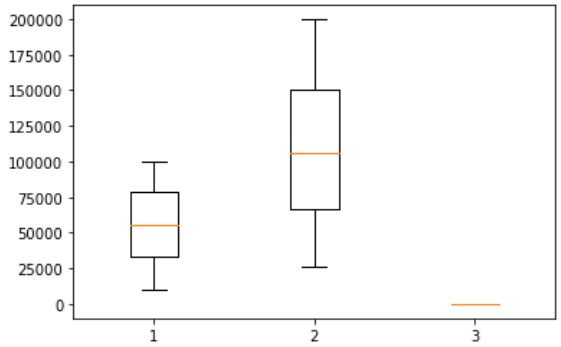
- Data with both objective and numeric

- Checking if there are any null, nan, or duplicated values. the data type is numeric or objective
- We did label encoding for categorical and binary columns
- We created bins for good, and risky based on their income.

Exploratory Data Analysis

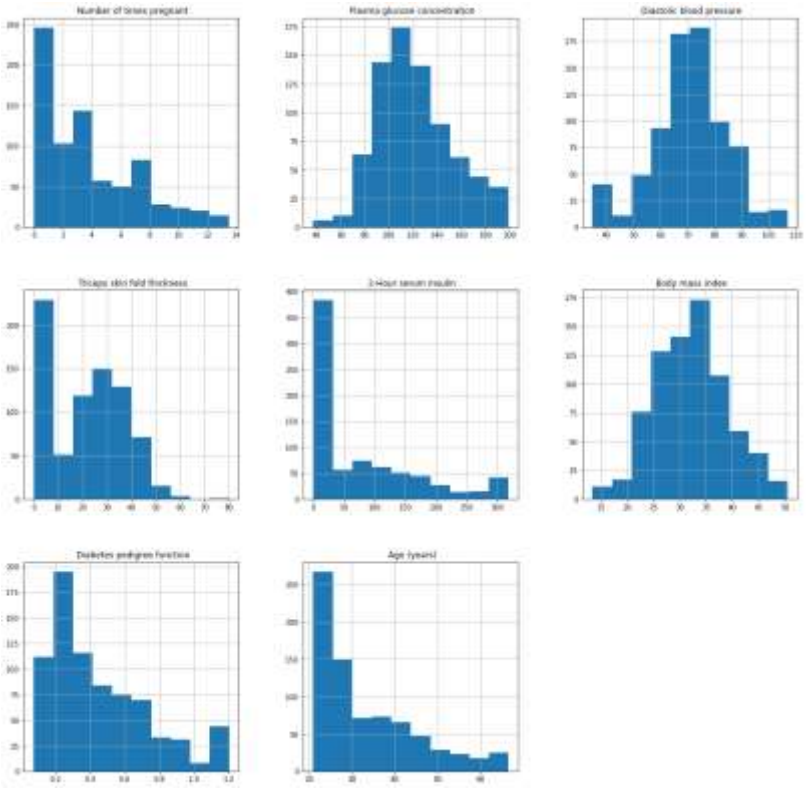
Univariate Analysis

- By using a boxplot we get outliers.

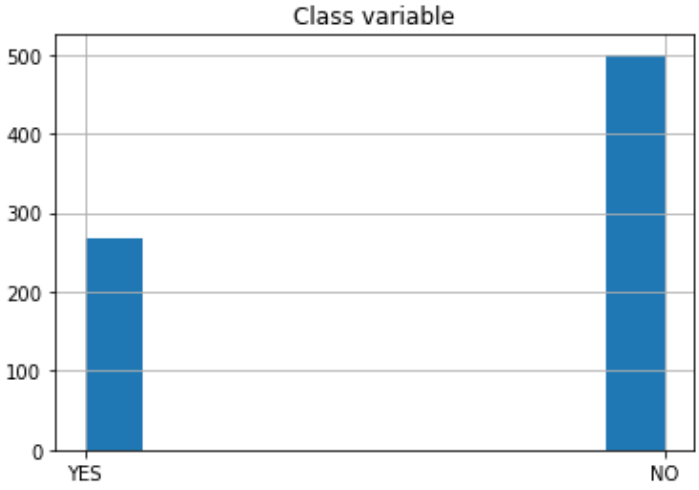


- We din't get any outliers for numerical data

- By using a boxplot we can visually observe which columns have the outliers



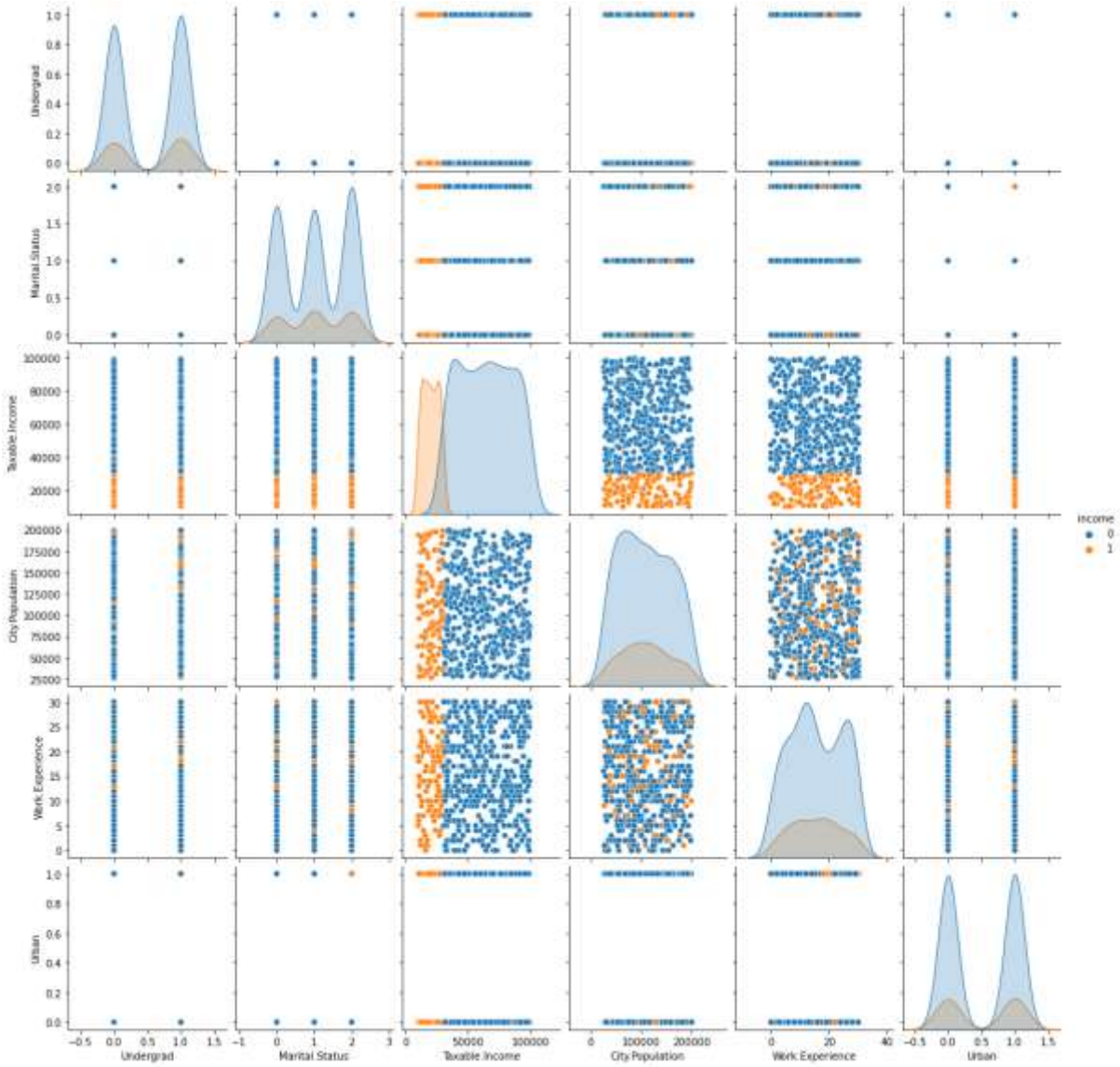
- By using the histogram shape of each column visually.
- The hist plot to see the distribution of the dataset but this time we are using this visualization to see the changes that we can see after those null values are removed from the dataset and we can clearly see the difference



- Here from the above graph it is clearly visible that our dataset is completely imbalanced in fact the number of patients who are diabetic is half of the patients who are non-diabetic.
- For outliers I performed winterization to compress the values in the 5^{th} and 95^{th} percentile.

Bivariate analysis.

- Pair plot helps to understand the best set of features to explain a relationship between two variables or to form the most separated clusters.



Model Building

- In the model building I took the target Urban as a class variable column and the predicted variable as the remaining columns. Then I split it into training and testing it at 70% &30%.
- I built the decision tree for a max depth is 5. I predicted the data.
- for training predicted accuracy data I got 48%
- for testing predicted accuracy is 65%

Model for Random forest

GridSearch CV

- Now after building the model let's check the accuracy of the model on the training dataset.
- In this model for the test predicted accuracy is 50%.
- For training predicted accuracy was 100%. In this model half of the data shows wrong predictions.

Randomized search cv

- by using the random search cv model for classification model
- for test accuracy of prediction is 43%
- for training predicted accuracy is 66%. So, this model has a high variance

Conclusion:

- After The analysis reveals that despite employing machine learning models such as random forests and decision trees, the accuracy in predicting fraudulent behavior remains low. This indicates the complexity and challenges involved in effectively identifying instances of tax evasion.

- In a dataset comprising 500 individuals, the models achieved the highest accuracy. However, upon closer examination of specific cases, discrepancies emerge. For example, the model inaccurately classified the first individual—a non-graduate and single person with a high income—as not being in a risky position, highlighting a misclassification rate of 50%.

- Similarly, the twelfth individual—a non-graduate, divorced individual with 14 years of work experience and a taxable income of 11k—was deemed to be in a risky position by the model. However, it's uncertain whether this prediction is accurate, emphasizing the limitations and uncertainties inherent in predictive analytics for fraud detection.

- These findings underscore the ongoing challenges in fraud detection and the need for more refined and accurate methodologies. While machine learning models offer valuable insights, their effectiveness can be limited, necessitating further research and development to enhance fraud detection capabilities.

Impact

- The impact of this analysis serves as a stark reminder of the ongoing challenges in fraud detection and the imperative for continuous innovation and improvement in detection methodologies. It highlights the need for interdisciplinary collaboration and investment in research to develop more effective strategies for combating financial misconduct and safeguarding against fraudulent activities.