PROJECT REPORT

Group No.: Group 13
Student Names: Namrata Bhartiya and Richa Talaty

Abstract

Our goal was to analyze the Census Income dataset from the UCI Machine Learning Repository and solve the classification problem of identifying whether a person has an income of $50K per year or more.

Introduction

The intention of this Project was to get a good understanding of the problem definition and the various attributes of this dataset to build the Classification models that could help in achieving the desired results. The appropriate data wrangling steps were performed to handle inconsistency, missing values, duplicates, and outliers. In Exploratory Data Analysis, the univariate analysis of numerical and categorical predictors was carried out separately to understand the distribution of the data and draw meaningful insights. The correlation analysis of the predictors against the target variable yielded information about the underlying relationships, resulting in removal of the redundant and uncorrelated attributes. To deal with the class imbalance problem, the Synthetic Minority Oversampling Technique (SMOTE) was used by oversampling the minority class. Before the training process, the data was rescaled using Standardization to ensure equal importance was given to all predictors in terms of variability. Thereafter, the pre-processed data was used for training the Logistic Regression, Hard Margin SVM, K Nearest Neighbours and Neural Network classification models, and their performance was evaluated using measures such as Accuracy, Precision and Recall.

Data description

Link - https://archive.ics.uci.edu/ml/datasets/Census+Income
The US Adult Census dataset is a repository of 48,842 entries extracted from the 1994 US Census database. This dataset comprises a total of 14 attributes. There are 5 attributes which are continuous in nature, and 8 attributes which are categorical in nature. The target variable is Income, which is binary-categorical in nature with categories ‘<=50 K’ and ‘>50 K’.

Attributes

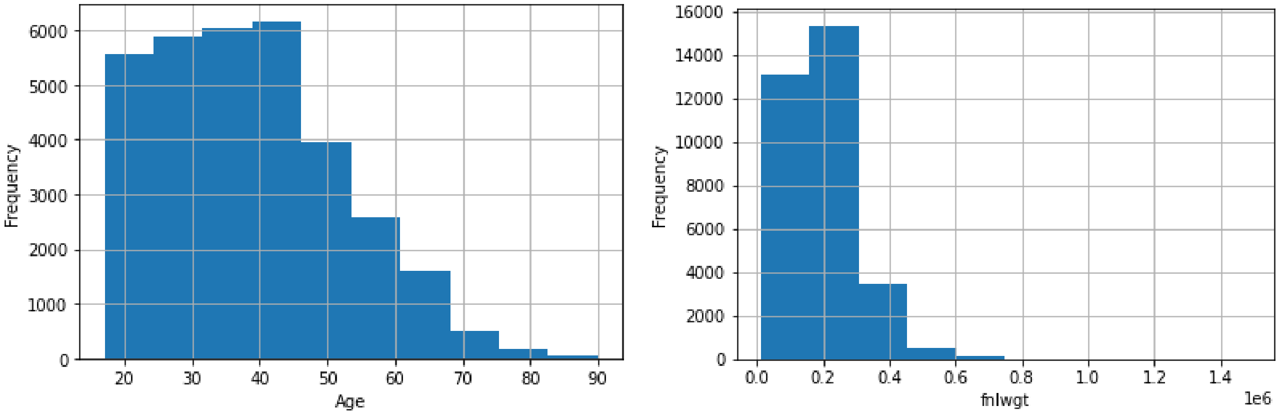
- Continuous attributes include
    1. age: age of the individual
    2. fnlwgt: represents how many people have the same list of features
    3. education-num: the highest level of education achieved in numerical form.
    4. capital-gain: capital gains for an individual
    5. capital-loss: capital loss for an individual
    6. hours-per-week: the hours an individual has reported to work per week
- Categorical variables include
    1. workclass: a general term to represent the employment status of an individual.
       [Domain - Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked]
    2. education: the highest level of education achieved by an individual.
       [Domain - Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool]

3. marital-status: marital status of an individual. Married--civ--spouse corresponds to a civilian spouse while Married--AF--spouse is a spouse in the Armed Forces.
   [Domain - Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse]
4. occupation: the general type of occupation of an individual.
   [Domain – Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces]
5. relationship: represents how this individual is related to others.
   [Domain -Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried]
6. race: [Domain - White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black]
7. sex: [Domain - Female, Male]
8. native-country: [Domain - United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad &Tobago, Peru, Hong, Holand-Netherlands]

Exploratory Data Analysis

A. Univariate Analysis of Numeric Variables

From the below figures in Fig.1, it is evident that all the numeric variables of this dataset are not normally distributed or are skewed in nature, which is also the case with most real-world data. The Age attribute shows that the highest frequency of people receiving income belong to the Age group of 18 to 45 years and mostly work full-time between 30 to 40 hours per week .



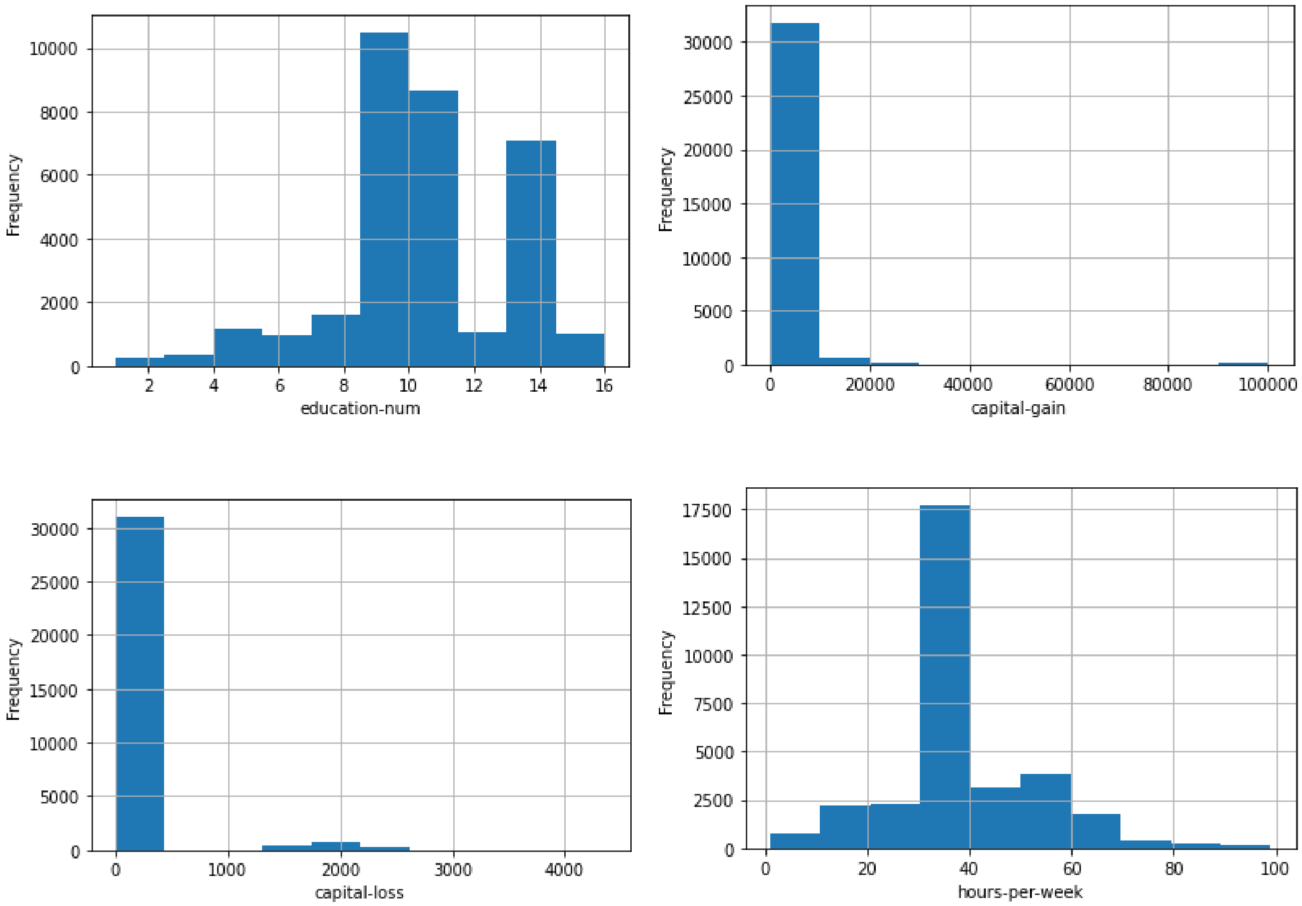
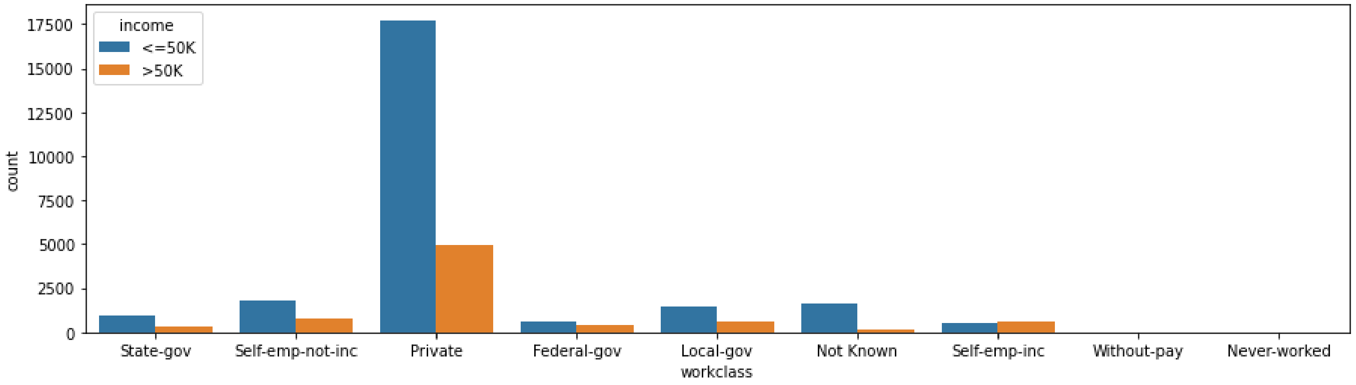
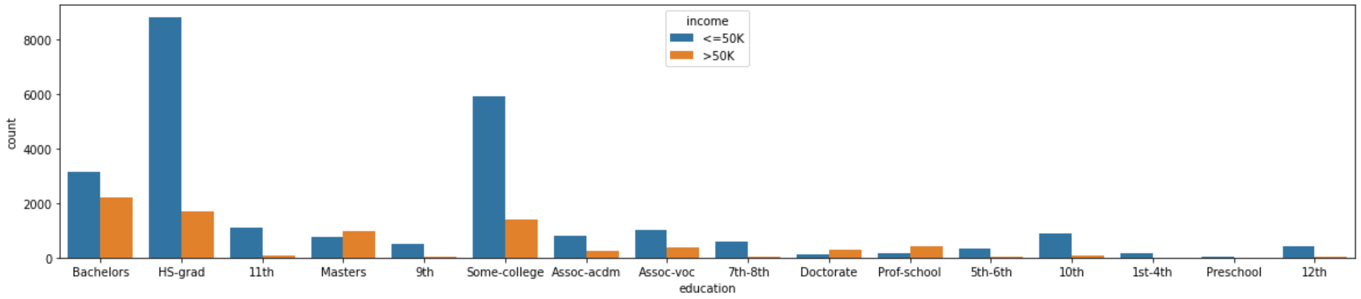


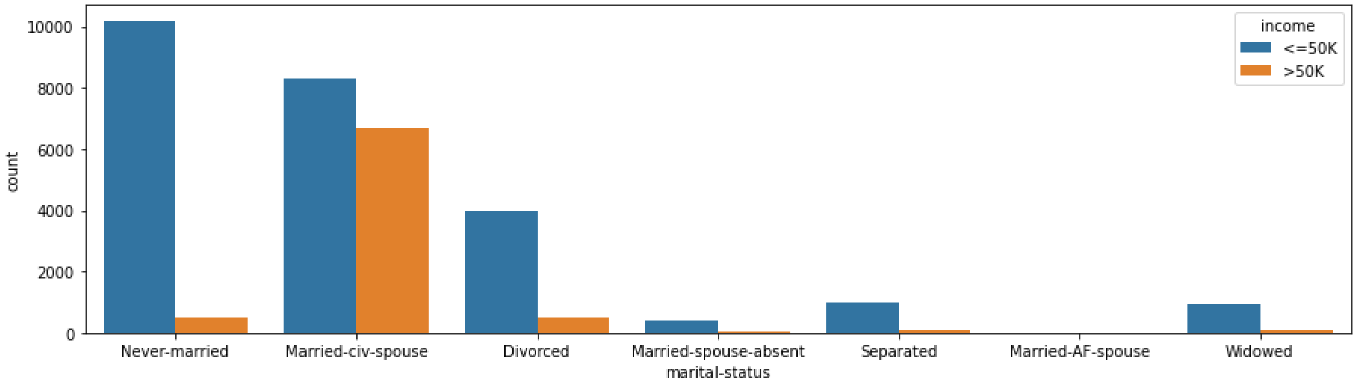
Fig.1. Univariate Analysis of the numerical variables

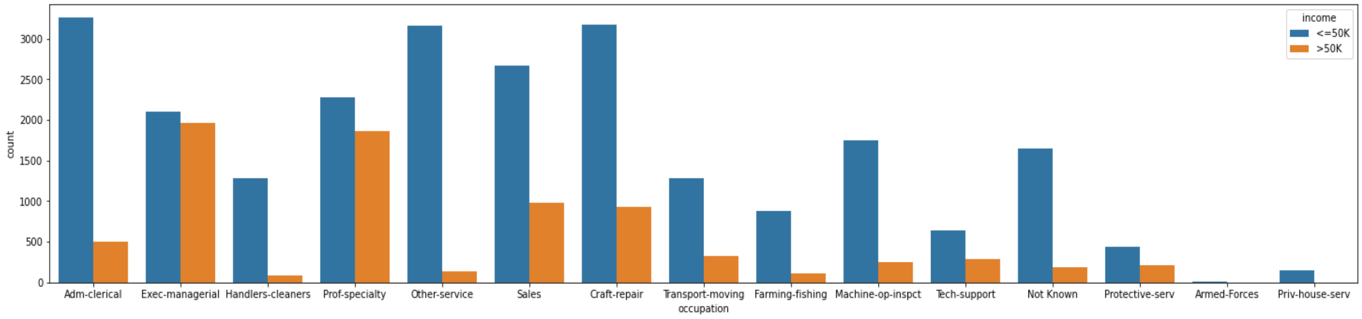
B. Univariate Analysis of the Categorical Variables

As seen in the below barplots in Fig.2, the Private workclass represents the highest records where income<=$50K and highest records for Income>$50K as compared to the other workclass categories. HS-grad Education type has highest records for income<=$50K, and highest records for Income>$50K, followed by those of Some-College. Never-Married marital status type has highest records for income<=$50K and highest records for Income>$50K, followed by Married-civ-spouse and Divorced. Admin-clerical staff of occupation type have the highest records for income<=$50K and highest records for Income>$50K, followed by Craft-repair and other-services. Most people not in a family tend to have a higher income of <=$50K than husbands and wives, and a high number of husbands have an income >$50K. The White population has the highest income compared to other race types, and most frequently were observed to have an income <=$50K, followed by the Black population. As compared to other countries, it was observed that the people of the United States had the highest income, mostly <=$50K, followed by Mexico.


income
<=50K
>50K
count
17500
15000
12500
10000
7500
5000
2500
0
State-gov
Self-emp-not-inc
Private
Federal-gov
Local-gov
Not Known
Self-emp-inc
Without-pay
Never-worked
workclass


income
<=50K
>50K
count
8000
6000
4000
2000
0
Bachelors
HS-grad
11th
Masters
9th
Some-college
Assoc-acdm
Assoc-voc
7th-8th
Doctorate
Prof-school
5th-6th
10th
1st-4th
Preschool
12th
education


income
<=50K
>50K
count
10000
8000
6000
4000
2000
0
Never-married
Married-civ-spouse
Divorced
Married-spouse-absent
Separated
Married-AF-spouse
Widowed
marital-status


income
<=50K
>50K
count
3000
2500
2000
1500
1000
500
0
Adm-clerical
Exec-managerial
Handlers-cleaners
Prof-specialty
Other-service
Sales
Craft-repair
Transport-moving
Farming-fishing
Machine-op-inspct
Tech-support
Not Known
Protective-serv
Armed-Forces
Priv-house-serv
occupation

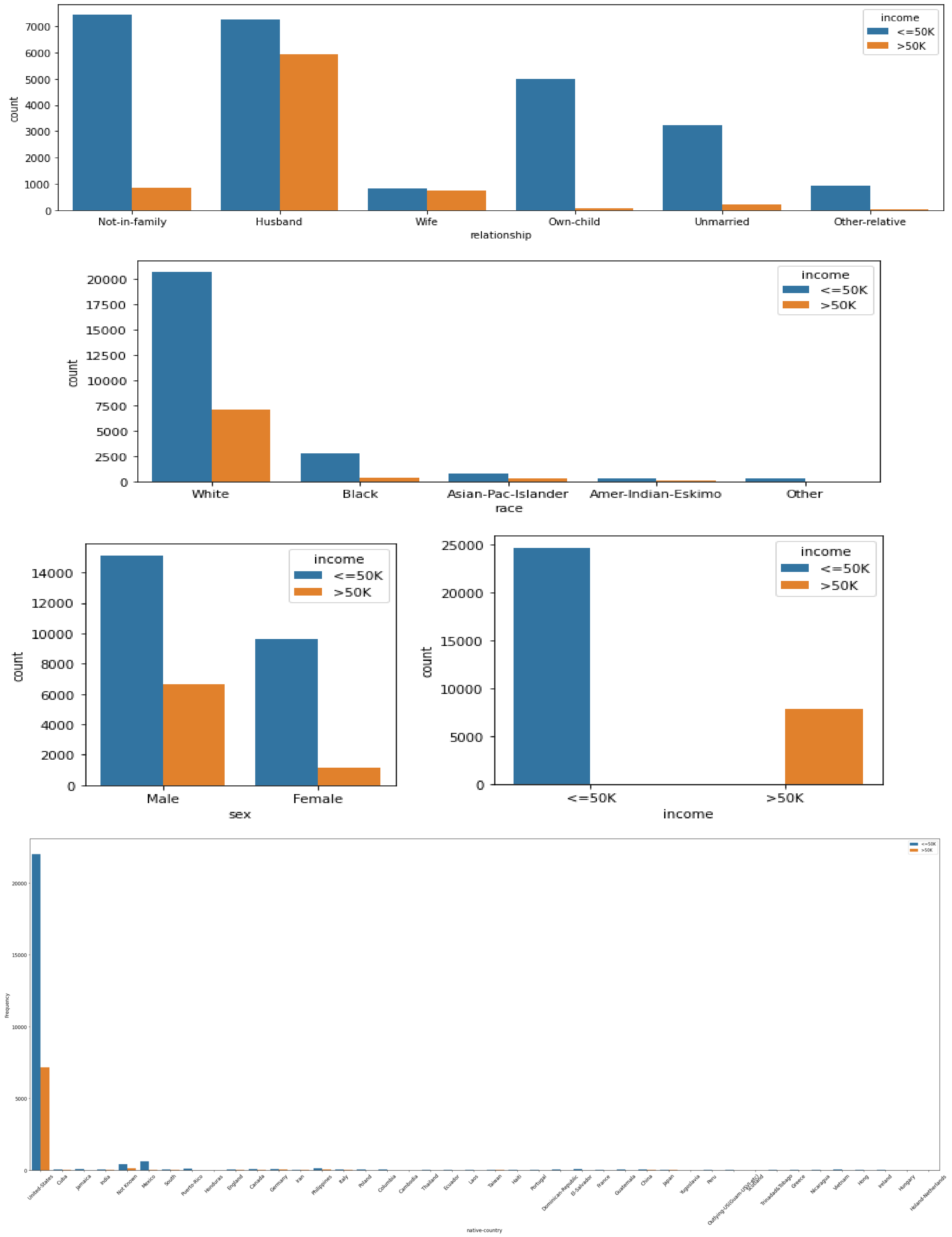


Fig.2. Univariate Analysis of the categorical variables

Data Pre-Processing

Correlation Analysis of Numeric Variables using Pearson's Correlation

From the heatmap represented in Fig.3., it can be inferred that there is no strong correlation amongst any of the numeric variables. Therefore, they do not exhibit Multicollinearity and thus, do not need to be removed.

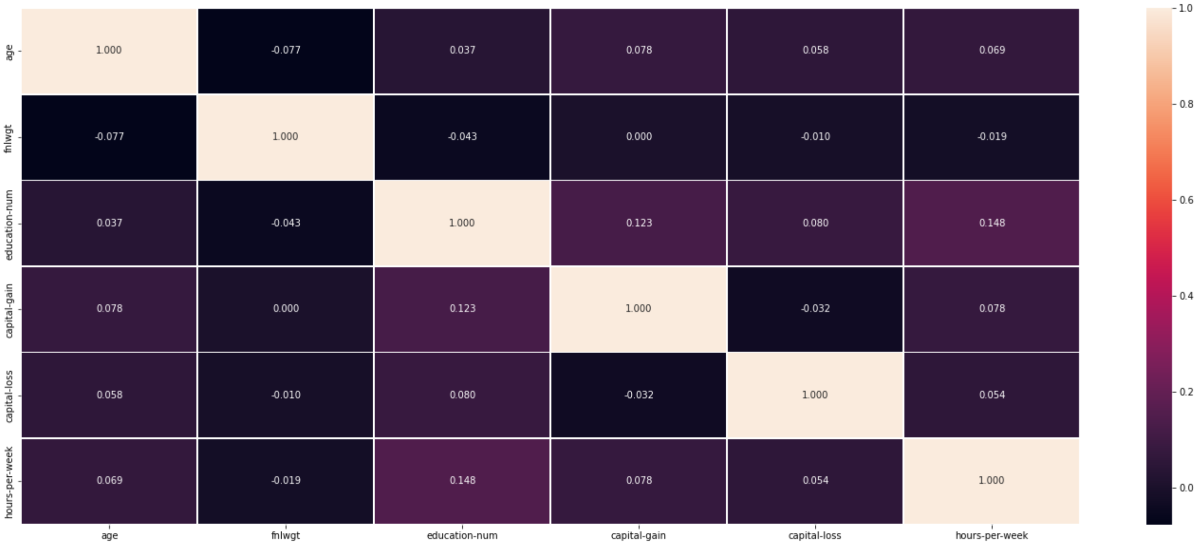


Fig.3.Results from the Pearson's Correlation Analysis

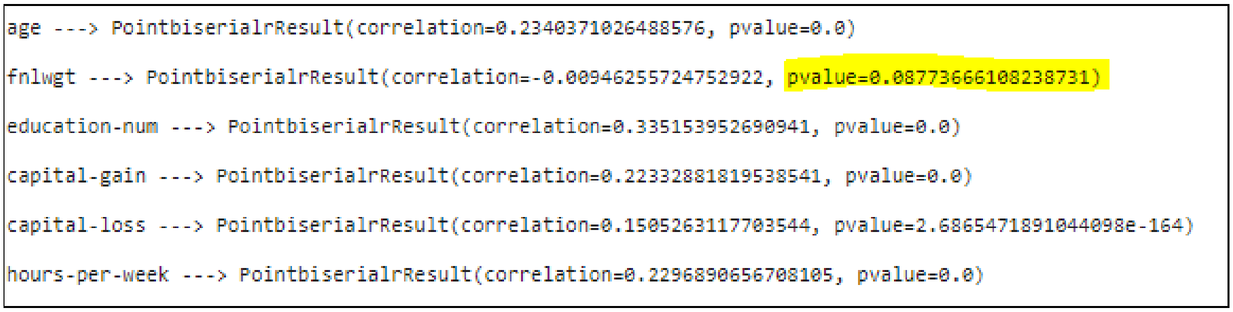
Point Biserial Test for Correlation Analysis of Numeric and Categorical Target Variable

For measuring the relationship between the binary target variable and the continuous variables of this dataset, we make use of the PointBiserial Correlation Result by defining a Null Hypothesis and an Alternate Hypothesis regarding the correlation amongst the features under consideration.

1. Null hypothesis H0 : the continuous and target variables have no correlation between them.
2. Alternate Hypothesis H1: the continuous and target variables have a correlation between them.

Conclusion from the PointBiserial Correlation Test:

As seen in Fig.3., **fnlwgt** has a p-value greater than 0.05, so we fail to reject the Null Hypothesis and assume that this numeric variable and binary categorical variable **income** are not correlated, as also explained by the correlation values in the result. For all the other numeric variables, since the p-value is less than 0.05, we reject the Null Hypothesis and assume that these numeric variables and binary categorical variable income are correlated.



```
age ---> PointbiserialrResult(correlation=0.2340371026488576, pvalue=0.0)

fnlwgt ---> PointbiserialrResult(correlation=-0.00946255724752922, pvalue=0.08773666108238731)

education-num ---> PointbiserialrResult(correlation=0.335153952690941, pvalue=0.0)

capital-gain ---> PointbiserialrResult(correlation=0.22332881819538541, pvalue=0.0)

capital-loss ---> PointbiserialrResult(correlation=0.1505263117703544, pvalue=2.6865471891044098e-164)

hours-per-week ---> PointbiserialrResult(correlation=0.2296890656708105, pvalue=0.0)
```

Fig.4.Results from the Point Biserial Correlation Test

As observed above in Fig.4, **Fnlwgt** is not correlated with the target variable **income** and the **Education** is redundant with **Education-num**, which has data as per weightage. Therefore, we can drop these 2 variables.

Feature Engineering

For the train and test sets, the dummy variables were generated for the categorical variables to ensure data is encoded in a binary format for ease of training the different classification model during the model building phase. As the data was missing for the native country Netherlands in the test set, a separate column consisting of all 0's was added as a substitute in the test set. The class representing income of >$50K is represented by class 1 and can also be referred to as the class of interest. Initially in the train set, there were 24720 records labelled as class 0 and 7841 records labelled as class 1. Thereafter, the SMOTE technique was used for oversampling the minority class of the training dataset, which was class 1. Therefore, before the model building phase, the dimensions of the training dataset were 49440 instances and 92 attributes.

Methods

1. Logistic Regression

It is a parametric classification model that relies on a specific model relating the predictor variables with the target variable, such that the outcome variable is categorical in nature. The output is the estimates of the probabilities of belonging to each class, then using a threshold cutoff on the probabilities for classifying into either of the classes. The outcome variable can be modeled as a linear function of the predictors.

Advantages:

- Easy to interpret and offers an intuitive explanation of predictors.
- Computationally fast and cheap to classify large samples of new data.

Disadvantages:

- Cannot be used to solve nonlinear problems.
- Sensitive to outliers and cannot handle missing values.

2. SVM

SVM uses the idea of identifying a hyperplane that is best to separate the features into different domains. The farther the Support Vector points are from the hyperplane, the higher is the probability of correctly classifying the points in their respective regions or classes. The support vector points are critical in determining the hyperplane because if the position of the vectors changes the hyperplane's position is also altered.

Advantages:

- Works well with a clear margin of separation
- Effective in high dimensional spaces

Disadvantages:

- Soft margin SVM had a high computation time and did not work well with our dataset

3. K-Nearest Neighbors

The key idea of the k-NN Classification algorithm is the use of similar records in the training data to classify a new record from the test data. The k-nearest neighbors are used to determine the similar records, and then a majority decision rule is used to classify the new record as a member of the majority class of the k-neighbors.

Advantages:

- Simple method, with no parametric assumptions.
- Performs well on large training datasets.

Disadvantages:

- Sensitive to large scale differences and outliers.
- Time-consuming computation involved in finding the distances.
- The number of records required in the training set to qualify as large increases exponentially with the number of predictors p.

4. Neural Networks

Neural networks are models which may be used for a classification or prediction problem. Some complex neural network models can also be used for deep neural networks or for feature extraction. The neural network model mimics the neurons present in human brains, as they have properties that resemble learning and memory based on experience.

Advantages:

- Results in a good predictive performance
- Can tolerate noisy data
- Capable of capturing highly complicated relationships between predictor variables and the target variable.

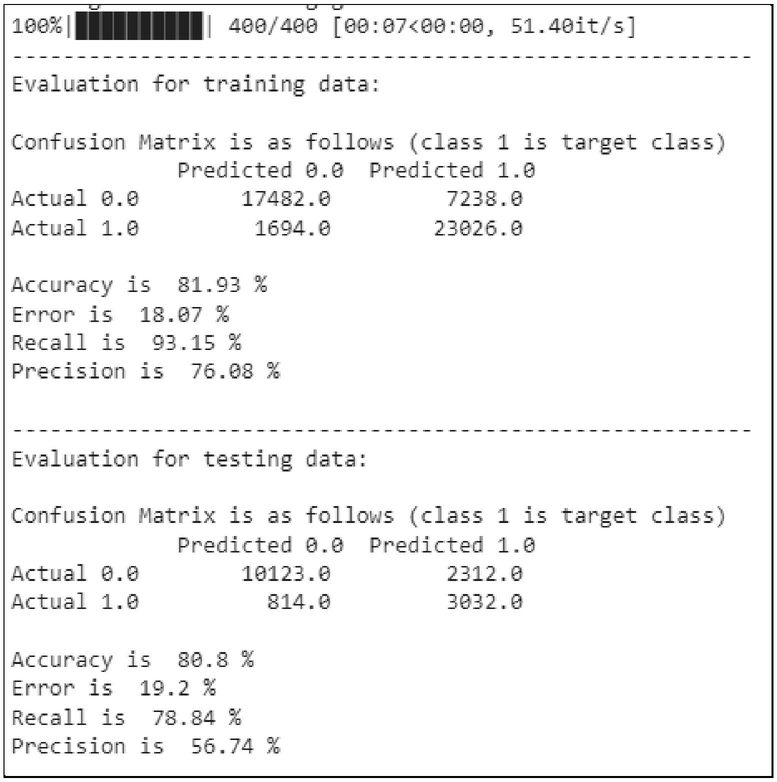
Disadvantages:

- Computationally expensive as they require longer runtime for training, which increases as the number of predictors increase
- The structure of the model lacks interpretability and is prone to overfitting.

Results

1. Logistic Regression

Initially with tolerance = 0.8 and learning rate = 0.001, and maxIterations = 100, we were getting a training error of 12.62%, while the test error was 18.72%. From this we could observe that our model was overfitting, therefore this indicated that there was low bias and high variance. To tackle the problem of overfitting, we manipulated the learning parameters, and the best parameters that resulted in reducing overfitting were for tolerance = 0.65, learning rate=0.00001, max iterations=400. Therefore, the logistic regression model resulted in 81.93% training accuracy and 80.81% test accuracy. And we can see from the confusion matrix in Fig.5, that even though a few data points were misclassified, the model in general fits most of the data points correctly.



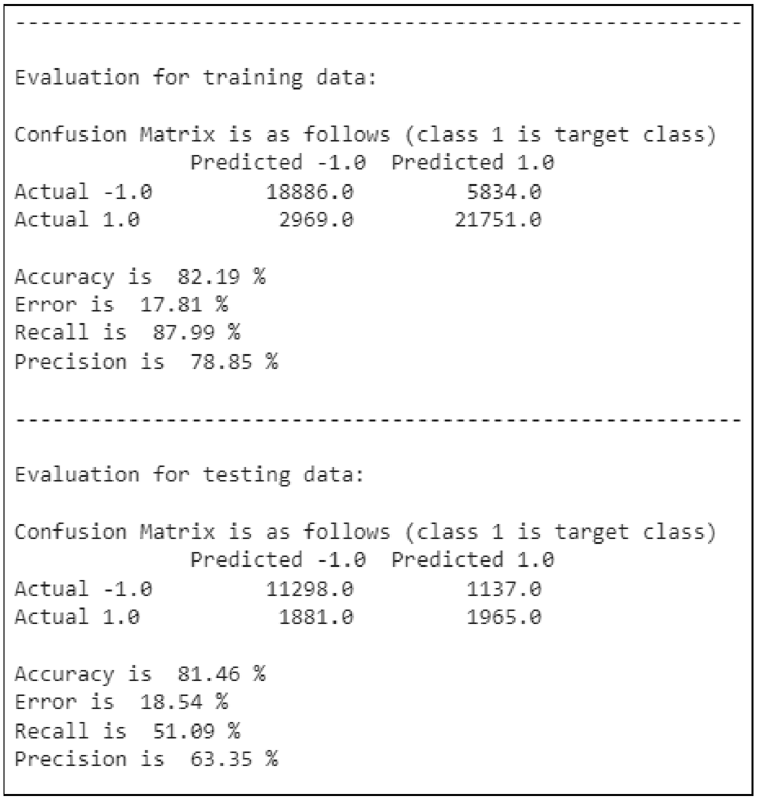
```
100%|██████████| 400/400 [00:07<00:00, 51.40it/s]
----------------------------------------------------------
Evaluation for training data:

Confusion Matrix is as follows (class 1 is target class)
             Predicted 0.0  Predicted 1.0
Actual 0.0        17482.0         7238.0
Actual 1.0         1694.0        23026.0

Accuracy is  81.93 %
Error is  18.07 %
Recall is  93.15 %
Precision is  76.08 %

----------------------------------------------------------
Evaluation for testing data:

Confusion Matrix is as follows (class 1 is target class)
             Predicted 0.0  Predicted 1.0
Actual 0.0        10123.0         2312.0
Actual 1.0          814.0         3032.0

Accuracy is  80.8 %
Error is  19.2 %
Recall is  78.84 %
Precision is  56.74 %
```

Fig.5.Performance Evaluation of Logistic Regression classifier

2. Hard Margin SVM

Initially with learning_rate = 0.000001, lamda = 0.01, n_iters = 200, we were getting a training error of 19.67%, while the test error was 18.01%. From this we could observe that our model was underfitting, therefore this indicated that there was high bias and low variance, i.e., the model does not generalize the data well. To tackle the problem of underfitting, we manipulated the learning parameters, and the best parameters that resulted in achieving the bias-variance tradeoff, were for learning_rate = 0.001, lamda = 0.01, n_iters = 200. Therefore, the SVM model resulted in 22.56% training error and 22.7% test error.

And we can see from the confusion matrix in Fig.6., that even though a few data points were misclassified, the model in general fits most of the data points correctly.

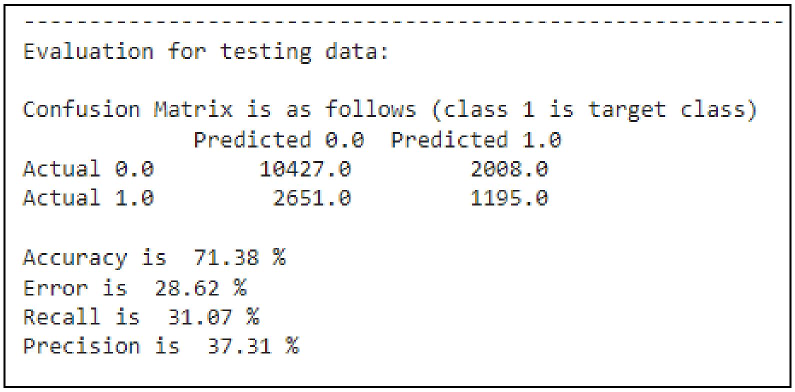


```
---------------------------------------------------------

Evaluation for training data:

Confusion Matrix is as follows (class 1 is target class)
              Predicted -1.0  Predicted 1.0
Actual -1.0        18886.0         5834.0
Actual 1.0          2969.0        21751.0

Accuracy is  82.19 %
Error is  17.81 %
Recall is  87.99 %
Precision is  78.85 %

---------------------------------------------------------

Evaluation for testing data:

Confusion Matrix is as follows (class 1 is target class)
              Predicted -1.0  Predicted 1.0
Actual -1.0        11298.0         1137.0
Actual 1.0          1881.0         1965.0

Accuracy is  81.46 %
Error is  18.54 %
Recall is  51.09 %
Precision is  63.35 %
```

Fig.6. Performance Evaluation of the Hard-Margin SVM classifier

3. K Nearest Neighbours

As k-NN is a non-paramteric model, no hyperparameter tuning could be done to improve its performance. The best test accuracy achieved was 71.38% for the value of k=3, as depicted in Fig.7.

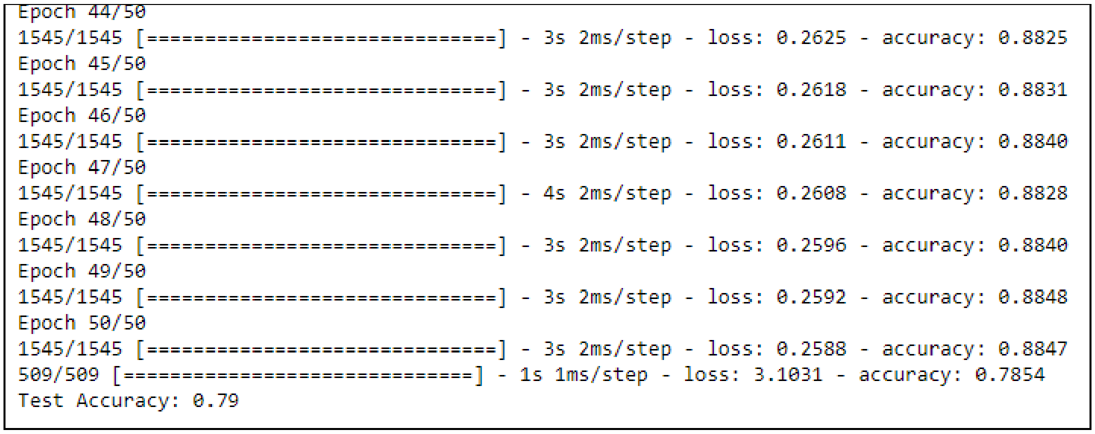


```
---------------------------------------------------------
Evaluation for testing data:

Confusion Matrix is as follows (class 1 is target class)
             Predicted 0.0  Predicted 1.0
Actual 0.0       10427.0         2008.0
Actual 1.0        2651.0         1195.0

Accuracy is  71.38 %
Error is  28.62 %
Recall is  31.07 %
Precision is  37.31 %
```

Fig.7. Performance Evaluation of the K-Nearest Neighbors classifier

4. Neural Networks

We executed the Sequential model of Keras for 50 epochs, with 1 input layer, 1 hidden layer and 1 output layer. The 'relu' and 'sigmoid' activation functions were used. We used the 'adam' optimizer and 'binary_crossentropy' loss function, which resulted in 79% test accuracy.



```
Epoch 44/50
1545/1545 [==============================] - 3s 2ms/step - loss: 0.2625 - accuracy: 0.8825
Epoch 45/50
1545/1545 [==============================] - 3s 2ms/step - loss: 0.2618 - accuracy: 0.8831
Epoch 46/50
1545/1545 [==============================] - 3s 2ms/step - loss: 0.2611 - accuracy: 0.8840
Epoch 47/50
1545/1545 [==============================] - 4s 2ms/step - loss: 0.2608 - accuracy: 0.8828
Epoch 48/50
1545/1545 [==============================] - 3s 2ms/step - loss: 0.2596 - accuracy: 0.8840
Epoch 49/50
1545/1545 [==============================] - 3s 2ms/step - loss: 0.2592 - accuracy: 0.8848
Epoch 50/50
1545/1545 [==============================] - 3s 2ms/step - loss: 0.2588 - accuracy: 0.8847
509/509 [==============================] - 1s 1ms/step - loss: 3.1031 - accuracy: 0.7854
Test Accuracy: 0.79
```

Discussion

In this Project, the goal was to be able to accurately identify whether a person earns >$50K or <=$50K per year.This problem was to be solved using the real-world data, through which it was observed that the number of people with an income of <=$50K were more in number as compared to people with an income of >$50K.Therefore, the class of interest represented by label 1 was that of >$50K. We cleaned our dataset by handling missing values and used Exploratory Data Analysis by conducting univariate analysis of numerical and categorical predictors separately to understand the distribution of the data and draw meaningful insights. As part of data pre-processing, we ensured data was consistent to be used for Model Building. We rescaled the data using standardization and then applied SMOTE to handle the class-imbalance problem. Further, we trained the Logistic Regression, Hard-Margin SVM, K-Nearest Neighbors, and Neural Networks classification models on the train data and evaluated their performances using metrics such as Accuracy, Precision and Recall. We also handled the problem of overfitting and underfitting in a few models by hyperparameter tuning to identify the best learning parameters for the models. We ensured that our models were generalizable to fit unknown data accurately, which we were able to verify using the test set.

We were able to see that Hard-Margin SVM was the best performing Classification model with test accuracy of 81.46%, followed by Logistic Regression with 80.8%, Neural Networks with 79% and KNN with 71.38%.