Prediction of Adult Income

Milestone: Draft of Project Report

Group 44

Student 1 - Krishna Barfiwala

Student 2 - Tanishka Adhlakha

barfiwala.k@northeastern.edu

adhlakha.t@northeastern.edu

Percentage of Effort Contributed by Student 1: 50%

Percentage of Effort Contributed by Student 2: 50%

Signature of Student 1: Krishna Barfiwala

Signature of Student 2: Tanishka Adhlakha

Submission Date: 21st April 2023

Table of Contents

Problem Setting:

Low income of people has caused problems and concern in recent years. In this project we will aim to conduct comprehensive analysis and highlight the main factors that are essential in improving unequal income of an individual. Classification is performed to predict whether a person's yearly income is greater than or less than 50k based on various sets of attributes and parameters.

Problem Definition:

This data is extracted from the Census bureau database by Barry Becker and Ronny Kohavi (Data Mining and Visualization, Silicon Graphics). The aim of the project is to predict if the adult income is greater than 50K or less than and equal to 50K. There are certain factors influencing the setting up a business in a city, which rely on the average income of people living in that city; Some other factors which affect the income of people are age, education, occupation, capital gain or loss, etc. This helps us determine various things such as the scope of profit-making business, preferences of real estate and bank loans eligibility for every individual, the type of people who would like to visit any tourist place and whether the people living there would put their students in private or public colleges in future. The goal of the project is to predict and extract as much information as possible from the data by using machine learning algorithms and finding appropriate patterns in the dataset using Association rules.

Data Sources:

The prediction of Adult Income Based on Census Data has been taken from kaggle, an open source, secure online repository-
https://www.kaggle.com/datasets/wenruliu/adult-income-dataset

About the dataset:

The dataset consists of basic information about the individual including the age, sex, education level, marital status, that might be affecting the income level of an individual. These details are the factors affecting the income of a person. There are 32561 instances (rows) and 15 attributes as shown in the snippet below.

Categorical variables:

1. Education
2. Education.num
3. Marital.status
4. Occupation
5. Relationship
6. Race
7. Sex
8. Native.country
9. Income

Numerical variables:

1. Age
2. Fnlwgt
3. Capital.gain
4. Capital.loss
5. Hours.per.week

Dataset and Data Cleaning:

We cleaned our data by replacing "?" with NaN values and then dropping the null values to ensure data quality, consistency, and accuracy which is an essential part of analysis and training models. Machine learning models trained on datasets with null values may not perform well in the real world because they lack the necessary information to make accurate predictions. By cleaning the data and removing null values, you can improve the model's performance and enhance the accuracy of predictions.

The original dataset without any data cleaning looks like this in fig 1 below:

	age	workclass	fnlwgt	education	education.num	marital.status	occupation	relationship	race	sex	capital.gain	capital.loss	hours.per.week	native.country	income
0	90	?	77053	HS-grad	9	Widowed	?	Not-in-family	White	Female	0	4356	40	United-States	<=50K
1	82	Private	132870	HS-grad	9	Widowed	Exec-managerial	Not-in-family	White	Female	0	4356	18	United-States	<=50K
2	66	?	186061	Some-college	10	Widowed	?	Unmarried	Black	Female	0	4356	40	United-States	<=50K
3	54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct	Unmarried	White	Female	0	3900	40	United-States	<=50K
4	41	Private	264663	Some-college	10	Separated	Prof-specialty	Own-child	White	Female	0	3900	40	United-States	<=50K

Fig 1: Original Dataset before Cleaning

The dataset in shown after data cleaning looks like this in fig 2 below:

	age	workclass	fnlwgt	education	education.num	marital.status	occupation	relationship	race	sex	capital.gain	capital.loss	hours.per.week	native.country	income
0	82	Private	132870	HS-grad	9	Widowed	Exec-managerial	Not-in-family	White	Female	0	4356	18	United-States	<=50K
1	54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct	Unmarried	White	Female	0	3900	40	United-States	<=50K
2	41	Private	264663	Some-college	10	Separated	Prof-specialty	Own-child	White	Female	0	3900	40	United-States	<=50K
3	34	Private	216864	HS-grad	9	Divorced	Other-service	Unmarried	White	Female	0	3770	45	United-States	<=50K
4	38	Private	150601	10th	6	Separated	Adm-clerical	Unmarried	White	Male	0	3770	40	United-States	<=50K

Fig 2: Dataset after cleaning

Data Exploration:

Statistical and visualization methods used to explore data. We are performing investigation and analyzing data to uncover patterns, relationships, and insights from our dataset. The goal of data exploration is to gain a deeper understanding of the data and to identify potential relationships or patterns that can inform decision-making or further analysis. Below are couple of data exploration and visualization performed -

1. Count of the number of people categorized as Males and Females:

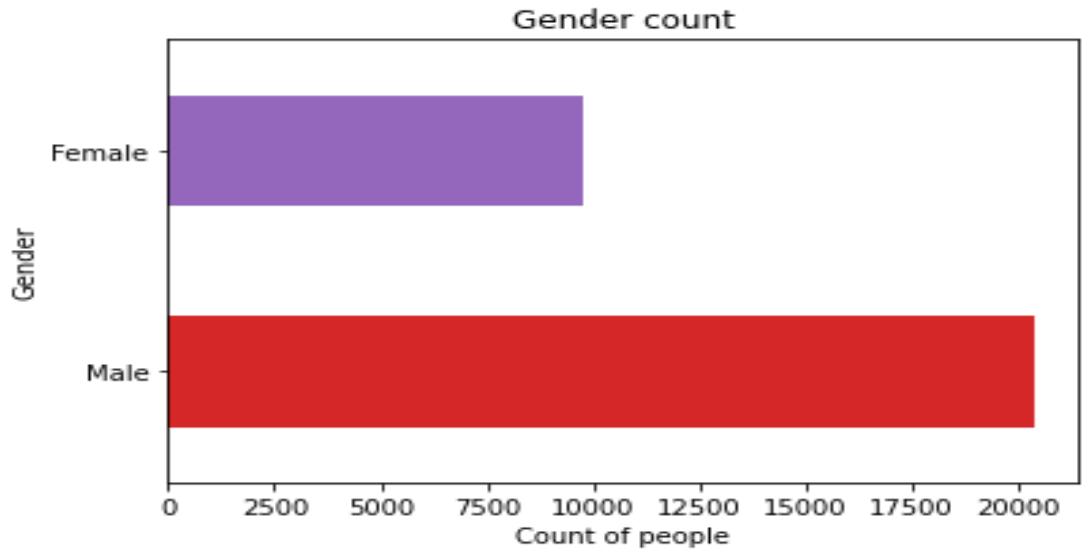

Fig 3: Gender Distribution

Interpretation: Above plot fig 3 interprets that there are a greater number of male counts (approx 22000) as compared to female counts (approx 8000) in our data set. We can further use this information to analyze other columns and perform observations.

2. Difference in the Work class and age with respective to Gender:

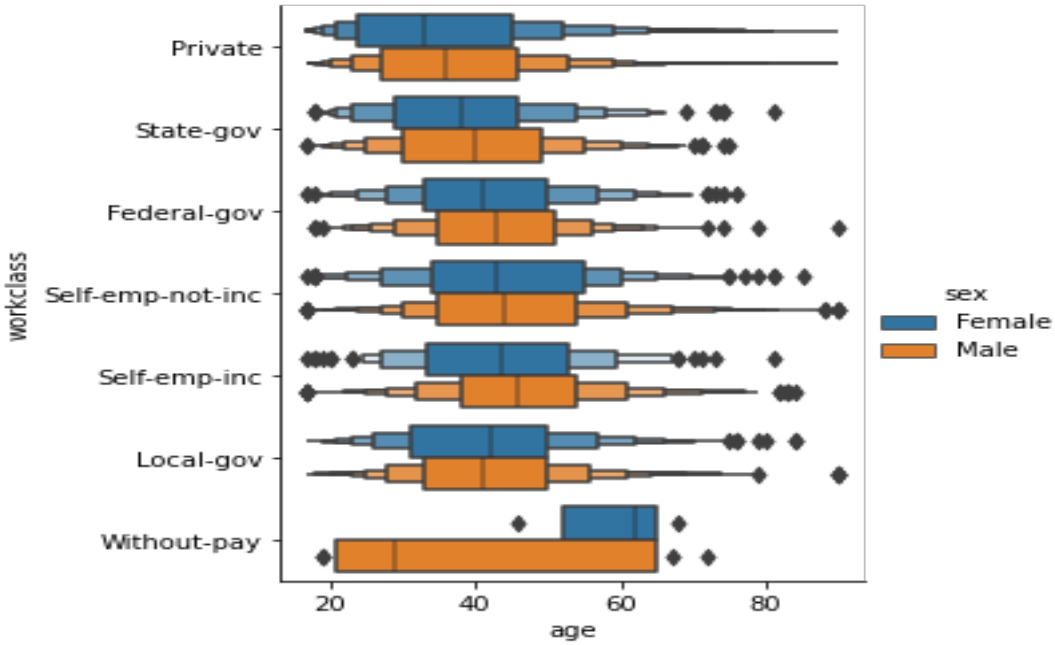

Fig 4: Work-class and Age Difference

Interpretation: The females and males have almost equal mean, differing by a small value, for almost all the work classes except 'Without-pay'. For the workclass 'Without-pay', the mean for females is relatively much higher than the males. The range for this work class is larger for males than females.

In the above graph we can also see that there are some outliers for workclass and their respective age. We can see that work class with age more than 80 have more outliers and very few number of work class people are falling under that age category.

3. Proportion of Work class, Age classified in income:

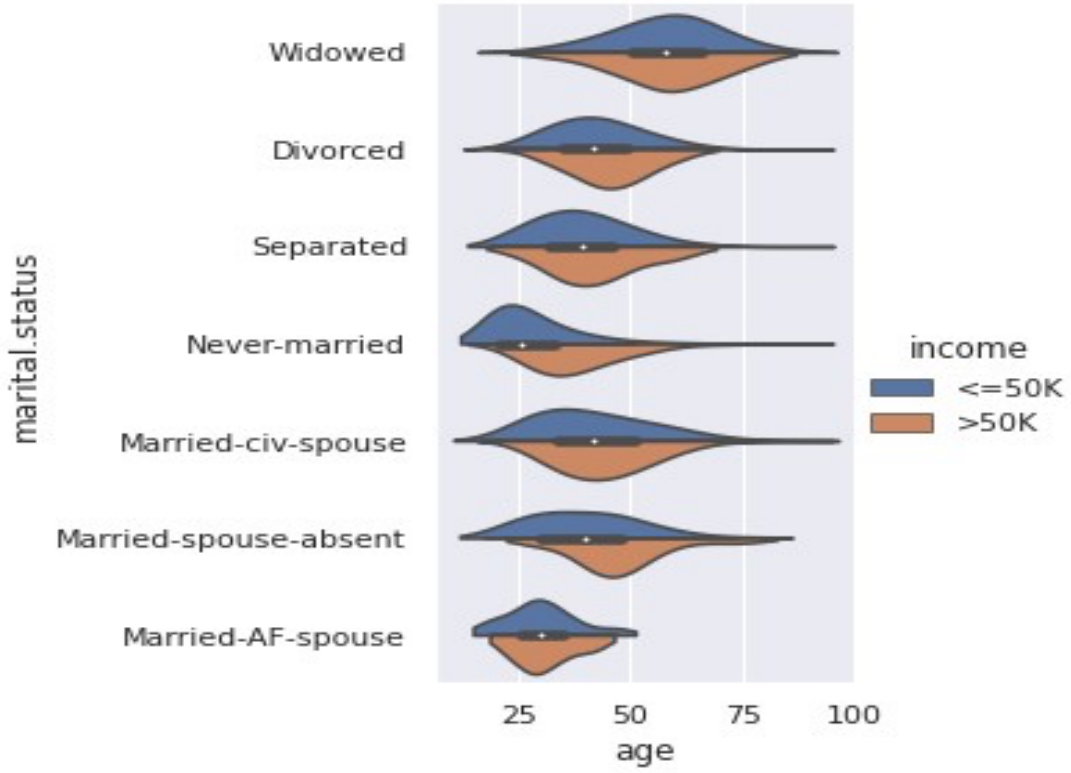

Fig 5: Proportion of work-class and age classified in income

Interpretation: The X-axis in this plot shows age and Y-axis shows the marital status in fig 5 above. The shape of the graph indicates the spread of individuals in that category. Widowed people are highly spread over the age range. Married-AF spouses are least spread. The graph also shows the spread of people with income greater than, less than or equal to 50k.

4. Distribution of age:

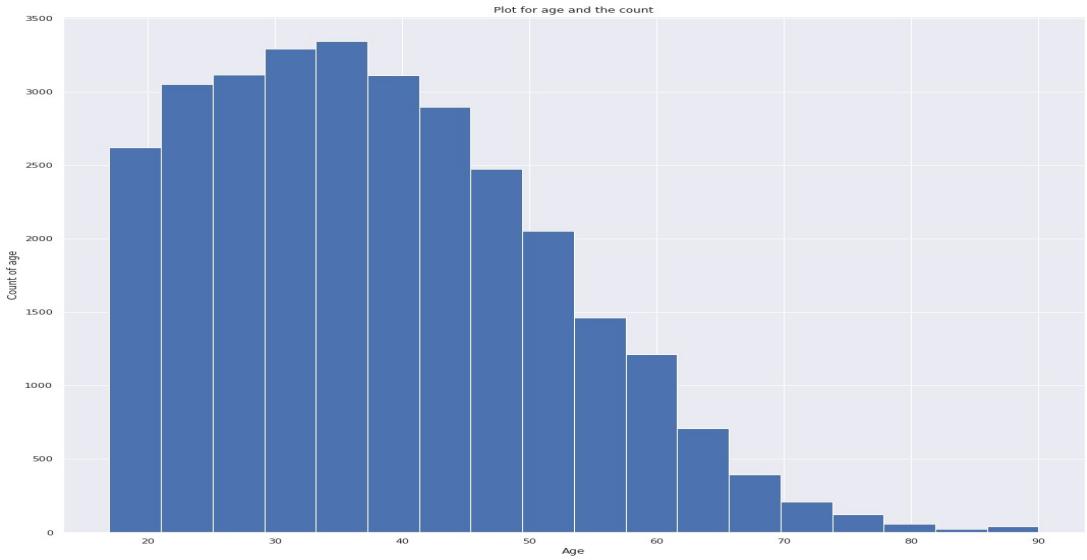

Fig 6: Age Distribution

Interpretation: This plot shows a distribution of age and its count. It is seen the age distribution is right skewed and most of the population in the data are from the 20 to 50 age group.

5. Age Group and occupation wise distribution:

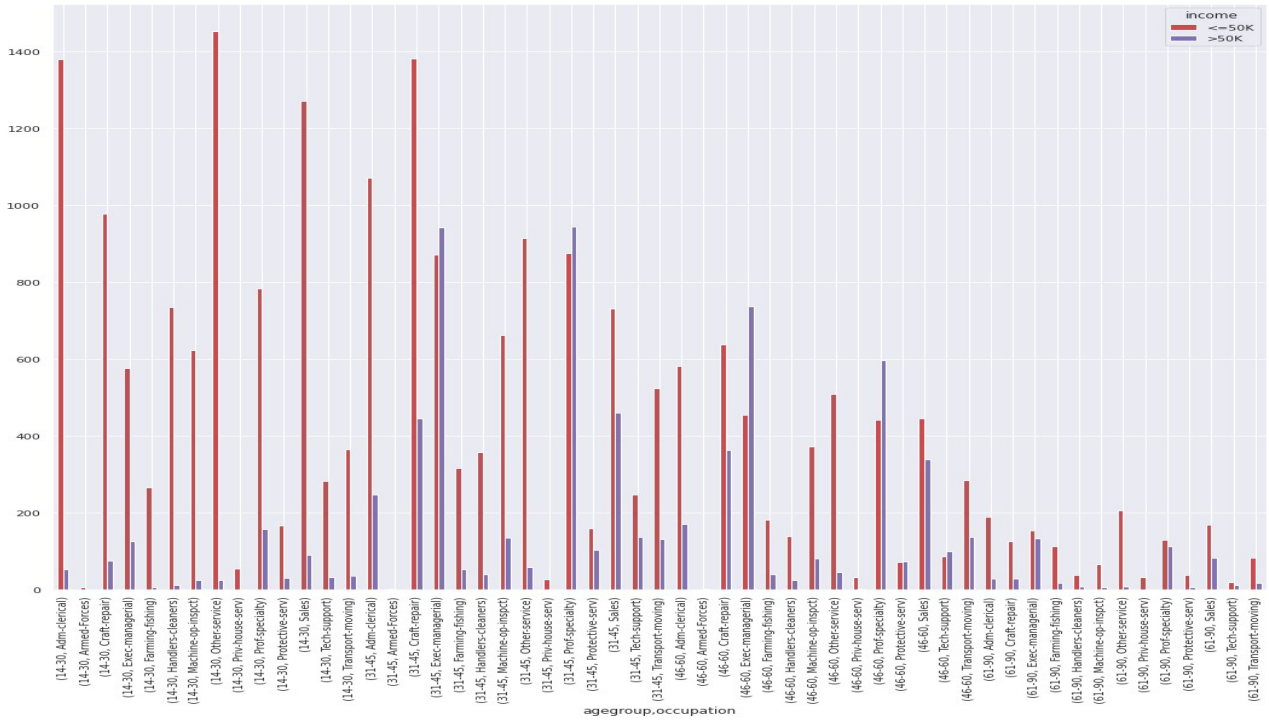

Fig 7: Age and Occupation Distribution

Interpretation: The line in the graph represents the count of the people in that particular age group and occupation. The X-axis represents the age, occupation and the Y-axis represents the count of individuals in that group. It is seen that in the age group 14-30, people working at other services are the maximum with income equal to or less than to 50k. For the age group 31-45, most people are working as Craft repair. Overall, for finding people with income greater than 50k in the range 31-45, people with Prof-speciality and Exec-managerial are the highest. For people in the age range 46-60, people working as Exec-managerial are earning salaries greater than 50k.

6. To check the Correlation between the variables:

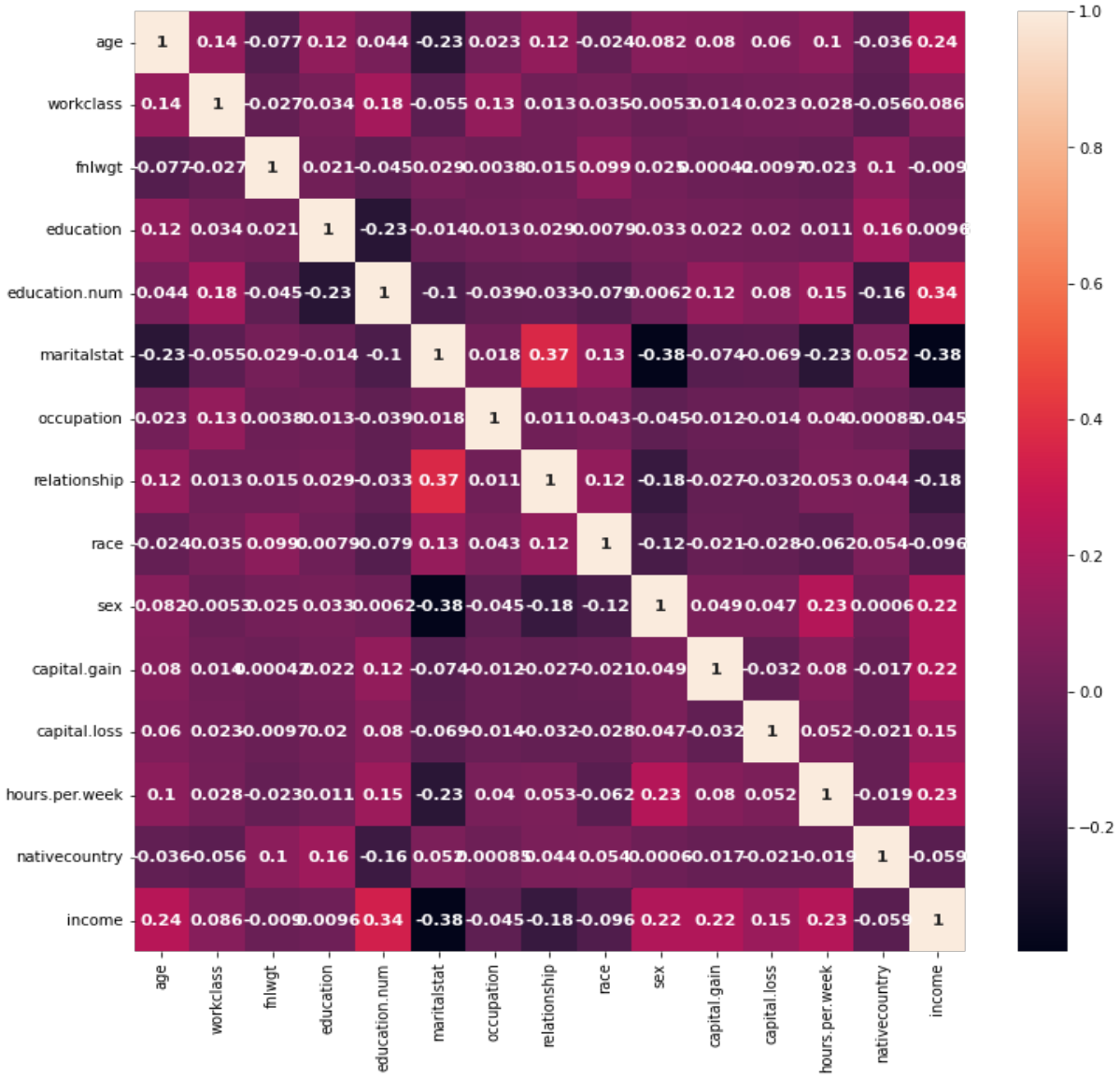

Fig 8: Correlation among variables

Interpretation: This correlation map shows the correlation between the variables shown in above fig 8. The darker shades indicate the variables are negatively correlated and lighter shade includes it is positively correlated. There is no strong relationship between any two variables.

7. Plot of Tree map to show the capital gain country specific and their work-class:

Fig 9: Native Countries hierarchical distribution

Interpretation - Tree map shows hierarchical distribution of data based on native countries for top 5 and their workclass with highest values of capital gain of their respective countries in fig 9. Taiwan has the highest capital gain with more number of private employees and Vietnam has the lowest capital gain with more number of private employees

8. Plot a tree map to show the capital loss country specific and their work-class

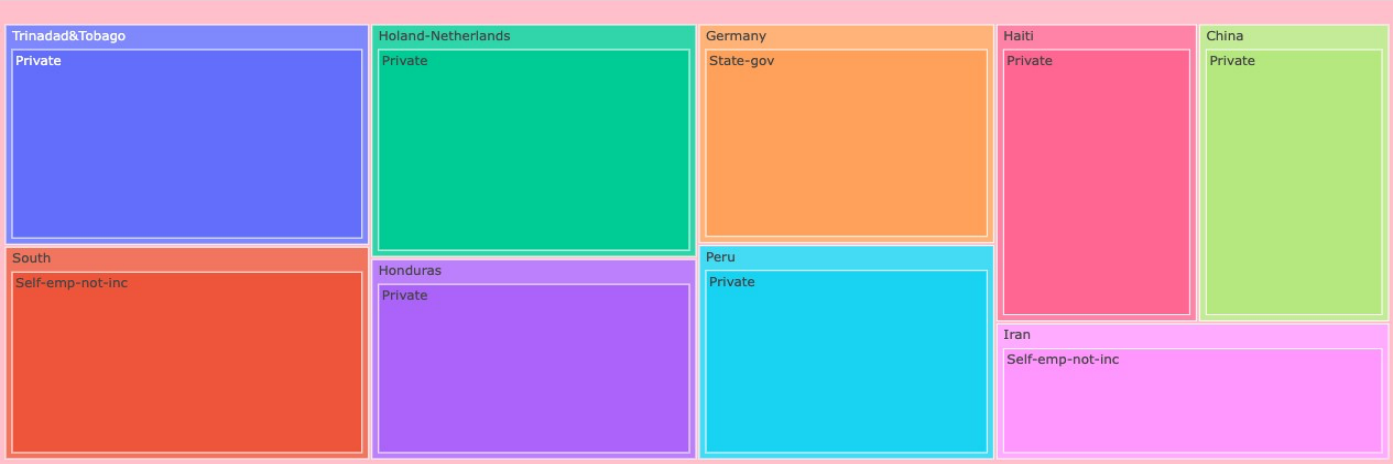

Fig 10: Capital loss country specific and their work class

Interpretation - Below tree map shows hierarchical distribution of data based on native countries for top 5 and their work-class with highest values of capital loss of their respective countries in fig 10. Trinadad & Tobago has the highest capital loss with a greater number of private employees and Iran has the lowest capital loss with a greater number of private employees

Data Mining Tasks:

Data mining is performed in our project to extract useful insights and knowledge from large datasets. We have identified patterns, relationships, and trends that are not immediately obvious from the raw data, and to use this information to make better decisions, predictions, or recommendations. Below were the important data mining tasks performed:

Analysis of Missing Values -

Analysis of missing values is performed using isnull() function to find if there are any null values present in our data set. The analysis of missing data is important because it can affect the accuracy and validity of statistical analyses and machine learning models. Statistical description and information is obtained using the describe function. Below function is used to find null values if present -

```
df_adult.isnull().values.any() # Check if there are any null values using isnull() function
```

Data Encoding –

We converted categorical variables to numbers such that the model is able to understand and extract the information. Most machine learning models only accept numeric variables, preprocessing the categorical variables becomes a necessary step. Below is the encoded data we have used to perform PCA and statistical analysis of data -

	age	workclass	fnlwgt	education	education.num	maritalstat	occupation	relationship	race	sex	capital.gain	capital.loss	hours.per.week	nativecountry
0	82	1	132870	4	9	5	5	4	1	1	0	4356	18	1
1	54	1	140359	9	4	2	8	6	1	1	0	3900	40	1
2	41	1	264663	2	10	4	6	2	1	1	0	3900	40	1
3	34	1	216864	4	9	2	3	6	1	1	0	3770	45	1
4	38	1	150601	13	6	4	9	6	1	2	0	3770	40	1

Fig 11: Data Encoding of Categorical Variables

Correlation Analysis

We have used Pearson correlation coefficient to evaluate the relationship between two and find the relationship between two or more variables. This can be done by calculating a matrix of the relationships between each pair of variables in the dataset. Correlation takes values between -1 and +1. A positive value for r indicates a positive association, and a negative value for r indicates a negative association. When r is 0, then there is no linear association between the variables.

Standardizing the Data

Data standardization is the process of rescaling the attributes so that they have mean as 0 and variance as 1. Standardization will bring the data to a common scale so that it allows the system to share and efficiently use data. Moreover, to find the Principal Component Analysis it is essential to standardize the data and it is normally distributed.

Principal Component Analysis –

PCA will simplify the complexity in high dimension data while retaining the trends and patterns. Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance.

Statistical Analysis:

Chi-Square test:

A Pearson's chi-square test is a statistical test for categorical data.

H0: The Gender and Income variables have no correlation between them

H1: The Gender and Income variables have a correlation between them

Dataset:

sex	Female	Male
income		
<=50K	0.388026	0.611974
>50K	0.150363	0.849637

Fig 12: Chi-square Test

Interpretation: Running the chi-square test, the value of p= 1 at 0.05% level of significance. Rejecting H0 if p< 0.05. Since p=1, we reject H1 and conclude that the gender and income variable have no correlation between them.

Conclusion On Statistical Analysis and Visualizations

We performed data analysis and visualizations to uncover patterns, relationships, and insights from our dataset. We used bar plots to study data regarding the male and female count present in the country and found that male count is more than the female count. We created a box plot to find the difference in the Workclass and age with respect to Gender and found that the mean for females is relatively much higher than the males. We created a histogram to find the distribution of age and found that people between the ages of 30-40 are more likely to be working. We created treemap, heatmap, corrplot to perform analysis and visualizations on the data. We used statistical methods like correlation, standardization, statistics of the variable, conducting dimension reduction etc to statistically analyze and visualize the data.

Model Performance Evaluation and Interpretation

Model performance evaluation and interpretation is required to assess the effectiveness of machine learning model, compare different models, identify model limitations, improve model accuracy and generalization, and explain the model's behavior and predictions to stakeholders. Without proper evaluation and interpretation, it is difficult to trust the results produced by a machine learning model and make informed decisions based on them. Therefore, it is an essential step in the machine learning and data science process.

Classification is a type of supervised machine learning task that involves assigning input data to one of several predefined categories or classes based on a set of features. Classification models are designed to learn patterns in labeled data and use them to predict the class of new, unseen data points.

There are many different classification algorithms and models that can be used to solve classification problems like:

10. **Logistic Regression:**
11. **Naives Bayes:**
12. **K-NN:**
13. **Decision tree:**
14. **Random Forest:**
15. **Support Vector machine:**
16. **Neural Networks**

In this project, we will discuss about a few classification models:

1. Logistic Regression:

In Logistic Regression, the main idea is to predict a dependent variable based on one or more independent variables. The Logistic regression uses a Sigmoid function which gives the probability of the outcome.

Advantages:

1. It is simple to understand and perform and does not require complex mathematical techniques.
2. Efficiency: Can handle large datasets with many independent variables.

Disadvantages:

1. Linear assumption: Logistic Regression assumes linearity between the independent variables, which might not hold true in all cases.
2. Non-probability predictions: Logistic regression does not always give the category the dependent variable belongs to, it gives the probabilistic predictions.
3. Large sample size: It requires large sample size to produce a reliable prediction.

Model Performance using Logistics Regression:

	precision	recall	f1-score	support
0	0.87	0.93	0.90	6761
1	0.75	0.59	0.66	2288
accuracy			0.85	9049
macro avg	0.81	0.76	0.78	9049
weighted avg	0.84	0.85	0.84	9049

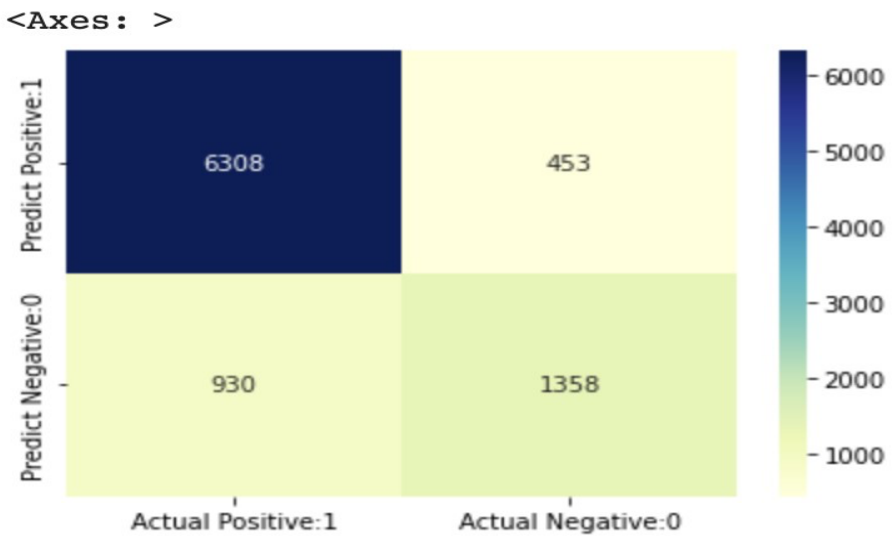

Fig 13: Confusion Matrix for Logistics Regression

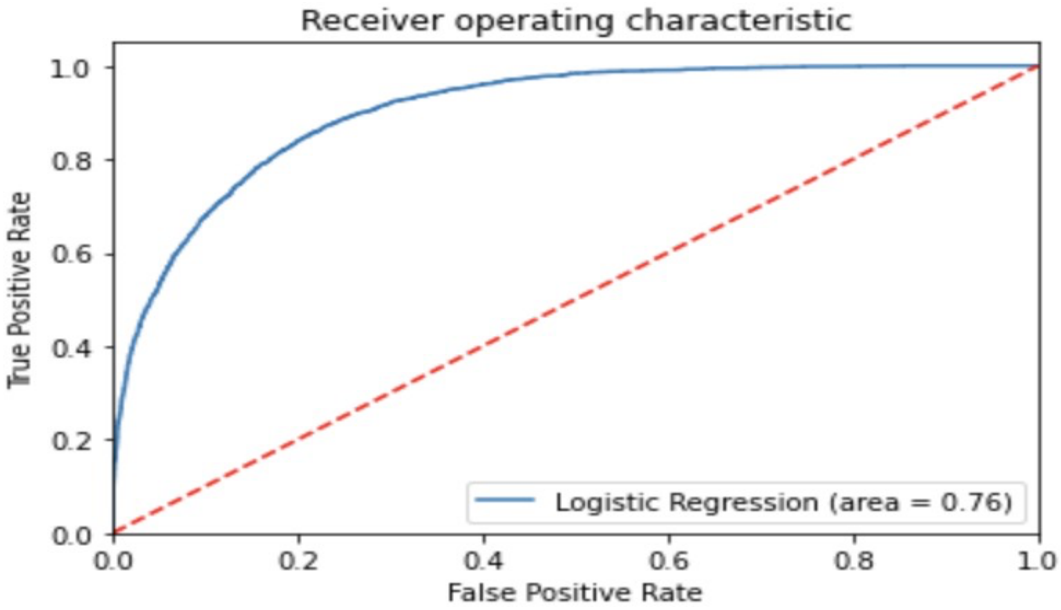

Fig 14: ROC Curve for Logistic Regression

Interpretation: It can be seen the Logistic Regression model has an accuracy of 85%. The confusion matrix can be seen with a good number of True Positives and True Negatives. There are comparatively a smaller number of False Positives and False Negatives.

The ROC curve shows the graph of TPR vs the FPR at various thresholds. The area under the curve gives the classification model's performance. As the ROC curve moves towards 1, it is considered as a perfect classifier. In this model, the ROC area curve is 0.76 indicating it is a good classifier. The Logistic regression classifier is a good model rather than random guessing.

2.Naives Bayes:

Naives Bayes is used to predict the certainty of an event occurring based on prior assumptions. This technique is applied by Naive Bayes to calculate the likelihood of each class given a set of features, and the most likely class is then chosen as the final classification.

Advantages:

1. It can be performed on a small set of training data.
2. It performs relatively faster, than other models, for a large set of data.
3. It performs very well with imbalanced data as it finds the probability and handles missing and irrelevant data

Disadvantages:

1. It assumes the features are independent of each other, even when they are not.
2. It may not perform well when there are outliers.

Model performance using Naives Bayes:

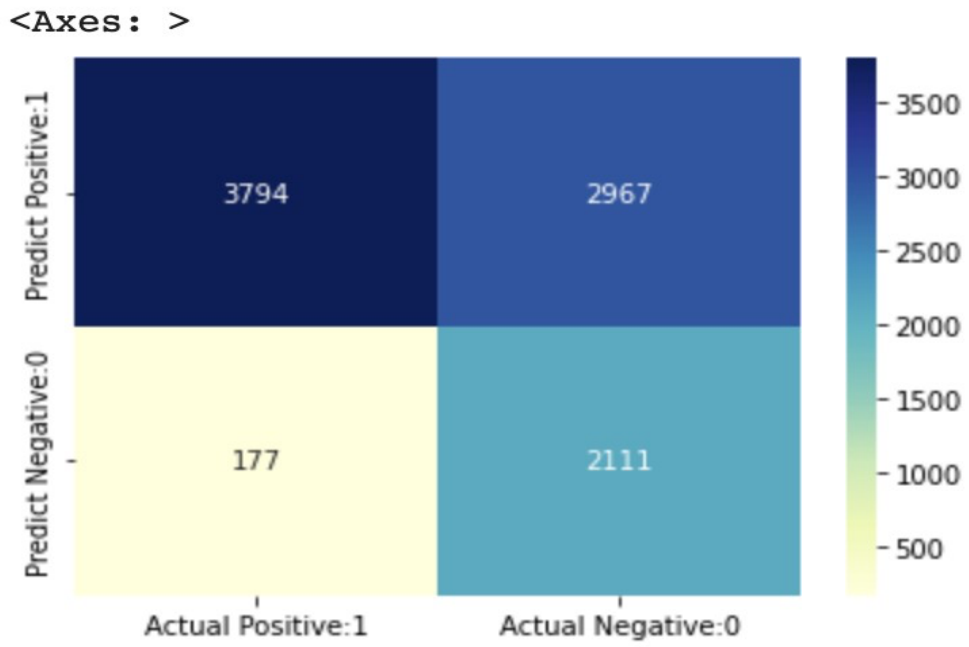

Fig 14: Confusion matrix for Naives Bayes

	precision	recall	f1-score	support
0	0.96	0.56	0.71	6761
1	0.42	0.92	0.57	2288
accuracy			0.65	9049
macro avg	0.69	0.74	0.64	9049
weighted avg	0.82	0.65	0.67	9049

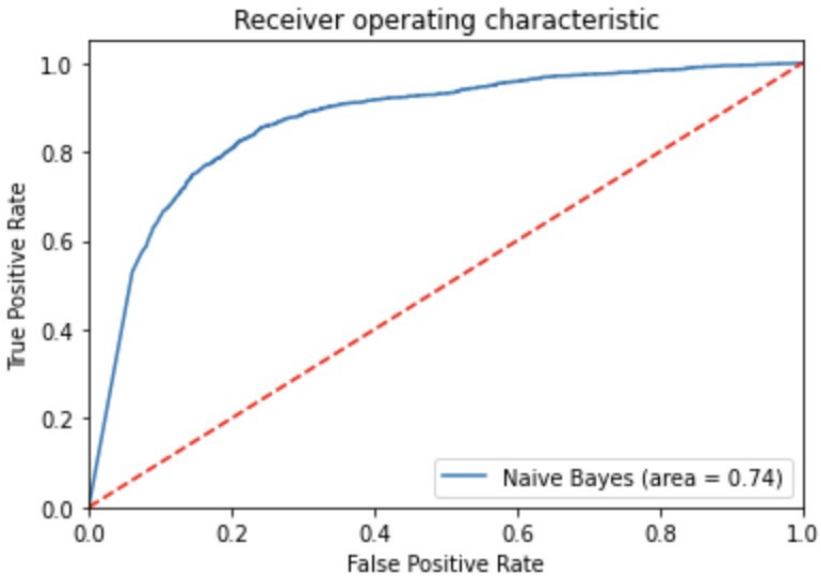

Fig 15: ROC curve of Naives Bayes

Interpretation: It can be seen the Naives Bayes model has an accuracy of 65%. The confusion matrix can be seen with a high number of False Positives, which is not considered to be a good characteristic of the model.

The ROC curve shows the graph of TPR vs the FPR at various thresholds. The area under the curve gives the classification model's performance. As the ROC curve moves towards 1, it is considered as a perfect classifier. In this model, the ROC area curve is 0.74 indicating it is a good classifier. The Naives Bayes classifier is a good model if we consider the ROC curve individually. But since the accuracy and the confusion matrix numbers are not good, we cannot consider Naives Bayes as a good classifier for this dataset.

3. KNN Classification –

KNN Classification is a widely used method for evaluating machine learning models. In KNN classification, the algorithm assigns the class label to a new data point based on its distance to the K nearest neighbors in the training set. The K value is a user-defined parameter that specifies the number of neighbors to consider. The class label assigned to the new data point is determined by the majority class label of the K nearest neighbors.

Below are some of the advantages and disadvantages of using k-fold cross-validation-

Advantages:

1. Robust to noisy data and outliers in the training set
2. No training phase is required, so it can be applied directly to new data
3. Can work well for small to medium-sized datasets with a low number of features

Disadvantages:

1. For large datasets or high-dimensional data, KNN classifier is not suitable and can be expensive
2. Not suitable for imbalanced datasets, where one class has a much larger number of samples than the other classes
3. Requires a complete and labeled training set, which can be a challenge for some applications

Implementation:

The KNN Classification procedure provides a good general estimate of model performance that is not too optimistically biased, at least compared to a single train-test split. We will use n=3, meaning we will have 3 neighbors.

We then create a KNN classifier with n_neighbors=3 and train the model on the training set using the fit method. Finally, we make predictions on the testing set using the predict method and evaluate the accuracy of the model using the score method

```
KNeighborsClassifier
KNeighborsClassifier(n_neighbors=3)
```

Model Performance using KNN Classification

	precision	recall	f1-score	support
0	0.88	0.90	0.89	6830
1	0.65	0.61	0.63	2219
accuracy			0.83	9049
macro avg	0.77	0.75	0.76	9049
weighted avg	0.82	0.83	0.82	9049

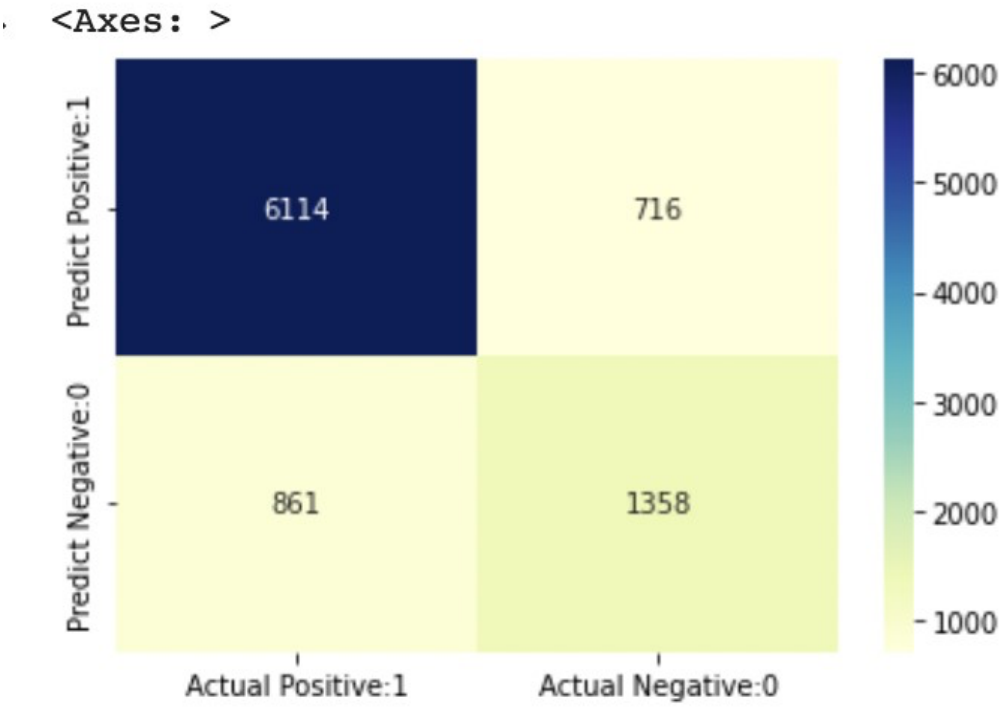

Fig 16: Confusion Matrix of KNN Classification

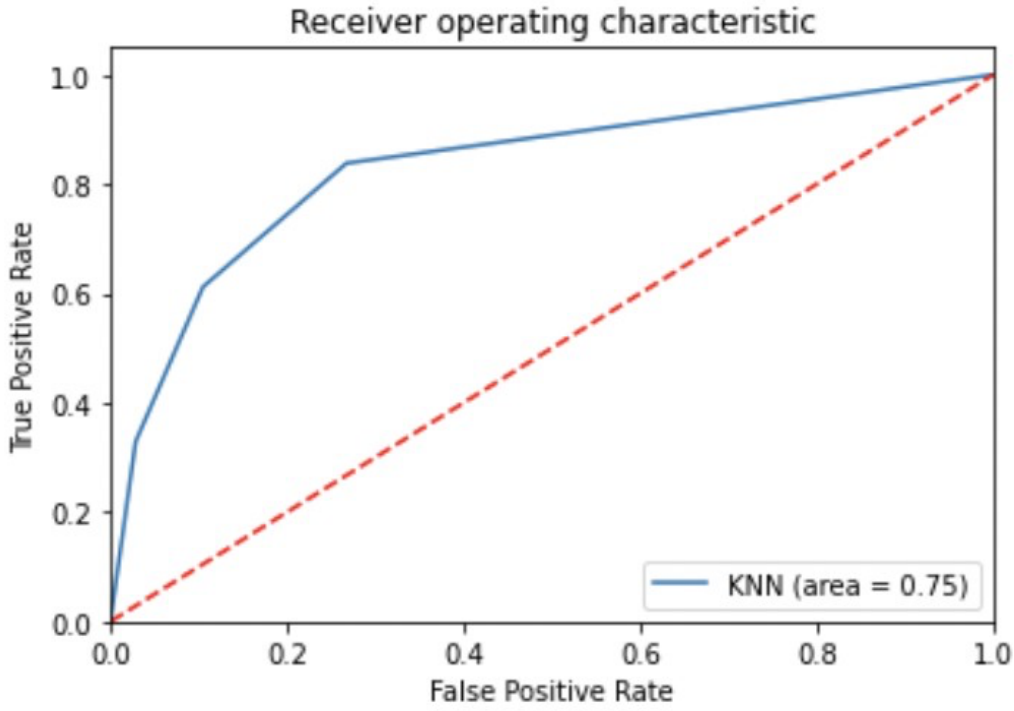

Fig 17: ROC curve of KNN

Interpretation - We calculated the accuracy, precision, recall and F1 score to find the performance of the model. The accuracy of the model is 83% and tells that it was able to classify 83% of the instances in the test dataset correctly. Accuracy and F1 score of KNN is less than SVC and Logistic Regression therefore it is not advisable to implement in the model.

The above ROC curve also has an area of 75% using KNN Classifier which is again lesser than SVC and Logistics model, indicating that the classifier with the 77% AUC has a better overall performance than the one with the 75% AUC.

4. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a type of supervised machine learning algorithm that is commonly used for classification and regression analysis. The basic working principle of SVM is to identify the optimal hyperplane that maximizes the margin between the classes. The margin is the distance between the hyperplane and the closest data points from each

class. SVM tries to find the hyperplane that maximizes the margin because it is thought to provide better generalization performance and be less prone to overfitting.

Advantages:

1. Robust to outliers: SVM is relatively robust to outliers in the data because it tries to maximize the margin, which can help avoid overfitting.
2. Effective in high-dimensional spaces: SVM is effective in high-dimensional spaces, which means it can work well even when there are many features to consider.
3. Good generalization performance: SVM is known for its good generalization performance, which means that it can perform well on new, unseen data.

Disadvantages:

1. Computationally intensive: SVM can be computationally intensive, especially when dealing with large datasets or complex models.
2. Limited scalability: SVM can be limited in terms of scalability because it requires the entire dataset to be stored in memory to train the model

SVM with a linear kernel is called SVM linear and uses linear decision boundaries to separate the classes. This means that the decision boundary is a straight line in the feature space. SVM linear is particularly useful when the input features are linearly separable, i.e., when the classes can be separated by a straight line in the feature space. In our dataset it is advisable to use a linear kernel since the data is linearly separable.

```
SVC
SVC(kernel='linear', probability=True)
```

Model Performance using SVM:

	precision	recall	f1-score	support
-1	0.88	0.93	0.90	6893
1	0.73	0.58	0.64	2156
accuracy			0.85	9049
macro avg	0.80	0.76	0.77	9049
weighted avg	0.84	0.85	0.84	9049

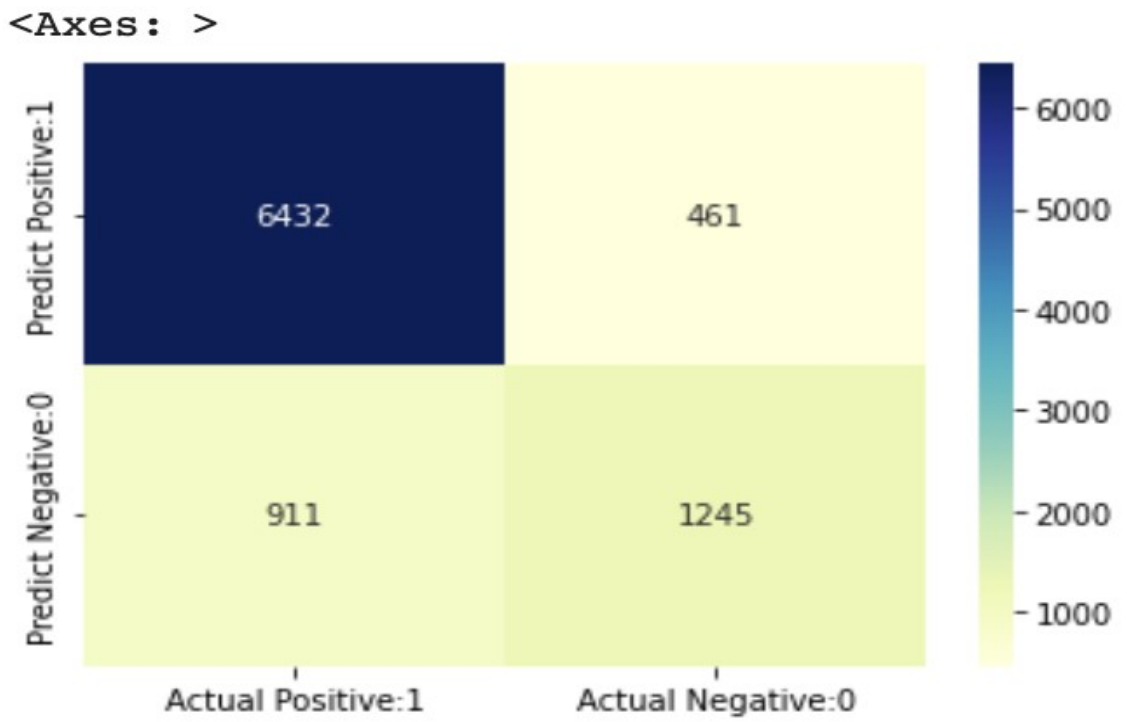

Fig 18: Confusion Matrix for SVM

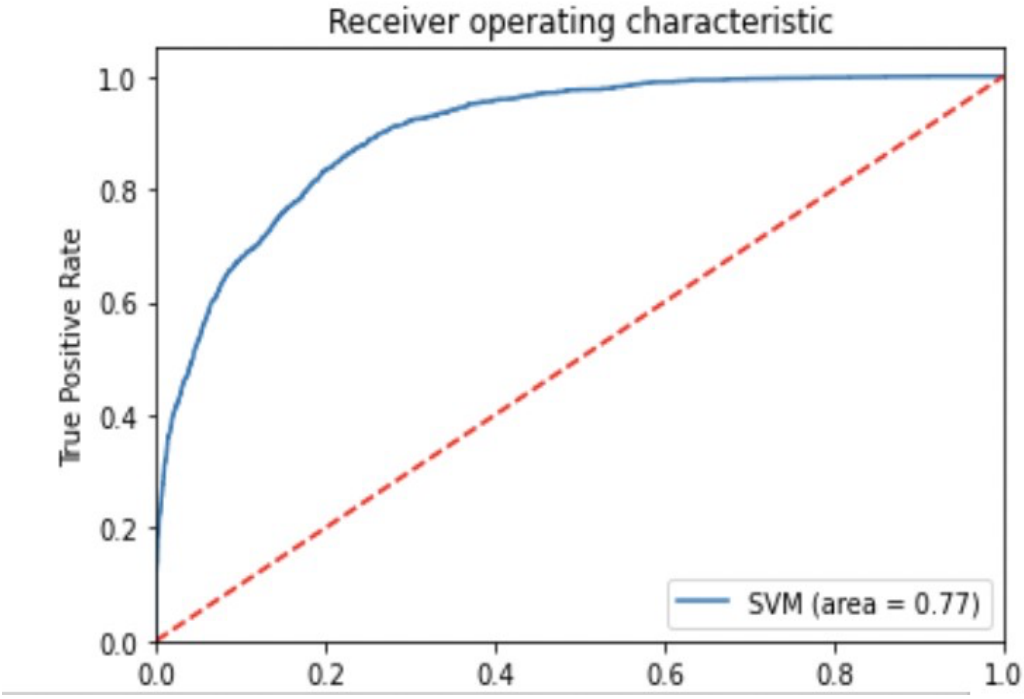

Fig 19: ROC of SVM

Interpretation - We calculated the accuracy, precision, recall and F1 score to find the performance of the model. The accuracy of the model is 85% and tells that it was able to classify 85% of the instances in the test dataset correctly. Accuracy and F1 score of SVC is greater than KNN and Naives Bayes therefore it is not advisable to implement in the model.

The above ROC curve also has an area of 77% using SVM which is again greater than KNN, Logistic and Naives Bayes model, indicating that the classifier with the 77% AUC has a better overall performance than the one with the 75% AUC.

Conclusion

Based on the above model predictions and comparison using various methods, we found that SVM has the highest accuracy with more percentage of AUC which tells that SVM is the best model selection to implement for this dataset. SVM is a powerful and effective

algorithm for classification tasks and is often the preferred choice for many applications. SVM is computationally efficient because it only considers a subset of the data points, called support vectors, which lie closest to the decision boundary. This reduces the computational complexity of the algorithm and makes it scalable to large data sets.