Introduction

This project revolves around the analysis and prediction of individual's income levels based on a diverse set of attributes. Income is a crucial socioeconomic indicator that directly affects an individual's standard of living, opportunities, and access to resources. By analysing and predicting income levels, we can gain insights into the distribution of wealth and socioeconomic disparities within a population.

Predicting income levels can assist individuals, organizations, and policymakers in making informed decisions. To achieve this goal, the project employs a systematic methodology encompassing essential steps such as data pre-processing, feature selection, and model training. These steps ensure that the dataset is prepared and optimized for accurate predictions. By selecting relevant features, the project aims to focus on the attributes that have the most significant impact on an individual's income level. Subsequently, various machine learning algorithms are applied to the dataset to find predictive models.

The project goes beyond model analysis. It also emphasizes the practical application of the project's findings in predicting income classes. By discussing the implications of the project's results, it aims to shed light on how such predictive models can be effectively utilized in real-world scenarios, such as in assessing individual's earning potential or in targeted marketing strategies.

Overall, the UCI Adult income project is an exploration of the relationship between individual's attributes and their income levels. By leveraging machine learning techniques, the project seeks to develop an accurate and robust predictive model that can be valuable for decision-making processes in various domains.

1. Literature Review

We search for research papers related to classification, prediction using machine learning. We got five research papers which includes the different classification techniques such as SVM, Naïve Bayes, Decision Tree, Random Forest, Logistic Regression. We have the review of various research papers and some of them are briefly discussed in the next section.

1. "Supervised Learning for Binary Classification on US Adult Income" by Li-Pang Chen

2. “Higher Classification Accuracy of Income Class Using Decision Tree Algorithm over Naive Bayes Algorithm” by Mohamed Zaida and RajendranT

These two research papers studied for this project explore the use of supervised machine learning models for income prediction using the UCI Adult income dataset.

- The first paper compares the performance of different machine learning algorithms and investigates feature selection techniques.
- The second paper compares the performance of decision tree and naive Bayes algorithms and highlights the importance of selecting the appropriate algorithm for a specific task.

Both papers provide valuable insights into the potential applications of income prediction and inform future research and practical use cases.

2. Objectives

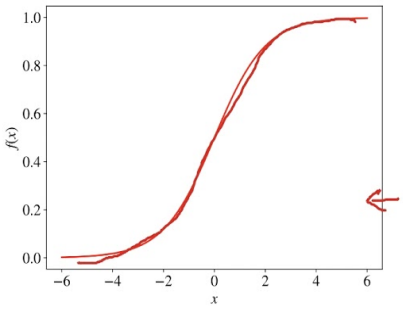
- To analyze the relationships between different Demographic variables in the dataset and income level.
- To fitting of machine learning model that can accurately predict income level based on individual characteristics.
- To compare the performance of different machine learning algorithms applied to the dataset.
- To predict an individual's income level based on their demographic, education, and employment-related features
- To investigate the potential applications of income prediction in fields such as finance, marketing, and social policy.

3. Methodologies

4.1 Logistic Regression:

- **Logistic regression** models a relationship between predictor variables and a categorical response variable.

Logistic Function



Sigmoid Function

$$y = \frac{1}{1 + e^{-(ax+b)}}$$

https://wv

- The multiple **binary logistic regression model** is the following:

$$\pi(X) = \frac{e^{(\beta_0+\beta_1X_1+\ldots+\beta_kX_k)}}{1+e^{(\beta_0+\beta_1X_1+\ldots+\beta_kX_k)}}$$

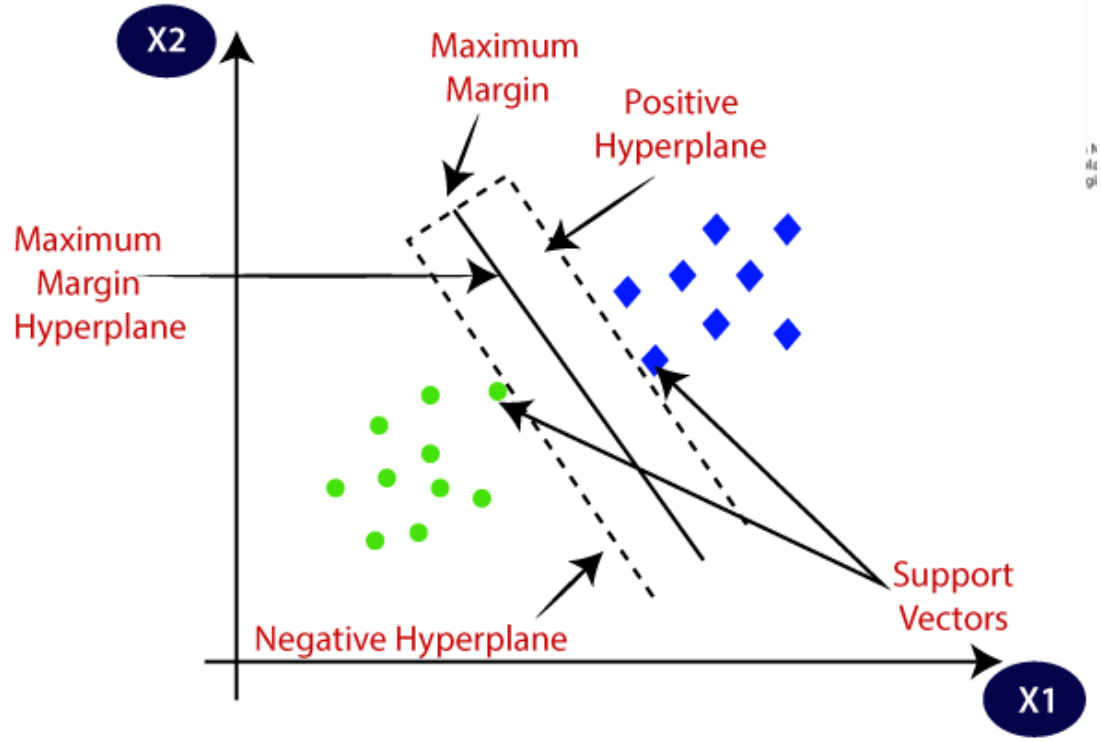
$$= \frac{e^{X\beta}}{1+e^{X\beta}}$$

$$= \frac{1}{1+e^{(-X\beta)^t}}$$

π is the probability that an observation is in a specified category of the binary Y variable, generally called the "success probability".

- It is trained to predict outcomes using a dataset with known outcomes, and it adjusts model parameters to maximize the likelihood of correct predictions.
- It models the probability of a binary outcome based on input features by fitting a logistic function to the data.

4.2 SVM (Support Vector Machine):



- ❖ SVM is a popular machine learning algorithm used for classification and regression problems.
- ❖ The goal of SVM is to find the best hyperplane that separates the data points into different classes with the largest margin.
- ❖ SVM can handle both linear and nonlinear data by using different types of kernels, such as linear, polynomial, and radial basis function (RBF) kernels.
- ❖ During training, the SVM algorithm calculates the optimal hyperplane that maximizes the margin and separates the data points into different classes.
- ❖ It constructs a hyperplane or a set of hyperplanes in a high-dimensional space to separate the classes. The Hyperplane is given by $\mathbf{f(x) = \beta_0 + x'\beta.}$
- ❖ SVM determines the class of a new observation using the decision function $\mathbf{G(x) = sign(f(x)) = sign(\beta_0 + x'\beta),}$ where sign() is the sign function.
  - G(x): The decision function used to classify the new observation.
  - $\boldsymbol{\beta_0}$: The intercept term in the model.
  - X: The predictor variables (features) of the new observation.

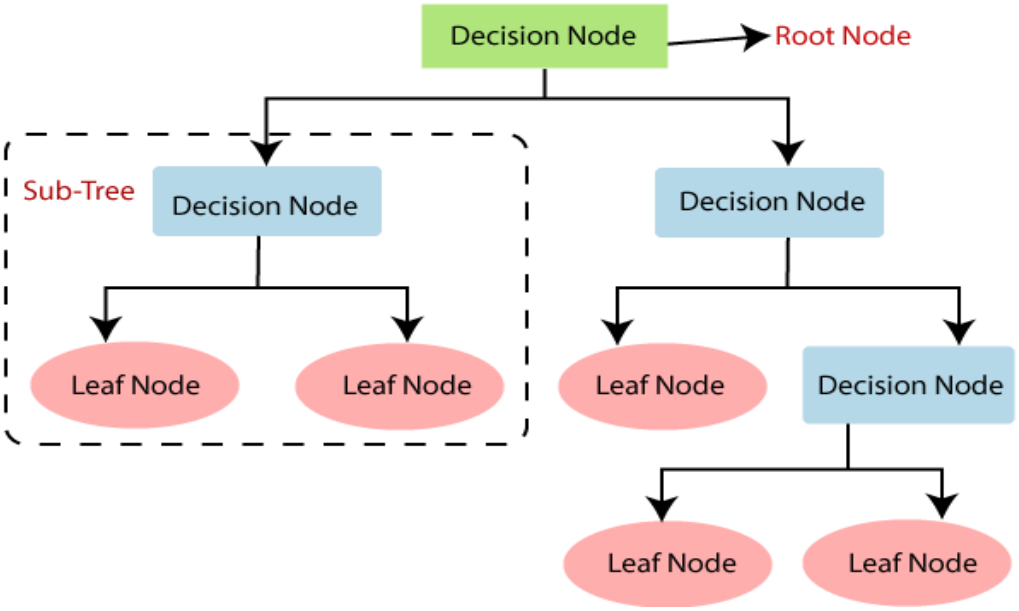
- **β**: The vector of coefficients (parameters) of the model.
- X': The transpose of the vector of predictor variables

❖ **Linear Kernel Function** It is defined as the dot product between the input features:

$$K(X_i, X_j) = X_i' X_j$$

where X_i and X_j are the input feature vectors for two data points i and j. The dot product captures the similarity between the input feature vectors in the original feature space.

4.3 Decision Tree:



❖ A decision tree classifier is a popular machine learning algorithm that uses a tree-like structure to model decisions and their possible consequences.

❖ The tree is built by recursively partitioning the input data based on the value of specific features, with each split based on maximizing the information gain.

❖ Once the tree is built, it can be used to classify new instances by traversing the tree from the root node to a leaf node that corresponds to a specific class label.

❖ For selection of best attribute in DTA the methods used are:

4.3.1. Gini Index:

$GINI(p) = 1 - p^2 - (1 - p)^2$

where p is the proportion of samples in the node that belong to the positive class (i.e., the class being predicted).

4.3.2. Information Gain:

Information Gain= Entropy(S)- [(Weighted Avg) *Entropy(each feature)]

4.3.3. Entropy: Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be calculated as:

Entropy(s)= -P(yes)log2 P(yes)- P(no) log2 P(no)

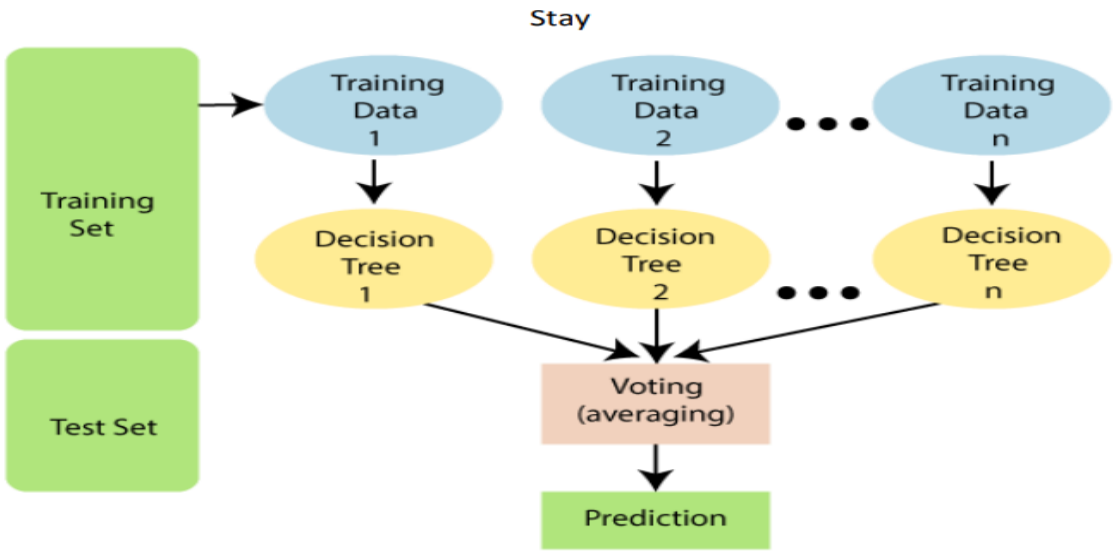
Where,

S= Total number of samples

P(yes)= probability of yes

P(no)= probability of n

4.4 Random Forest:



❖ Random forest is a popular machine learning algorithm that creates multiple decision trees by randomly selecting a subset of features and observations from the training data.

- Each tree is trained using a different subset of features and data, and the final prediction is made by aggregating the predictions of all the trees.
- Random Forest constructs many classification and regression trees (CARTs) randomly, and each CART is independent. When the forest is constructed, it can be used to make predictions based on new inputs.
- Here we use Gini index to select the most important feature from a feature subset to split the tree. Gini index represents the probability of a randomly selected feature being misclassified.
- $G(p) = \sum_{k=1}^{K}(1 - p_k)p_k$

where pk is the proportion of observations in category k, and K is the number of categories in a sample.

- $G(D) = 1 - \sum_{k=1}^{K}\left(\frac{C_k}{ND}\right)^2$

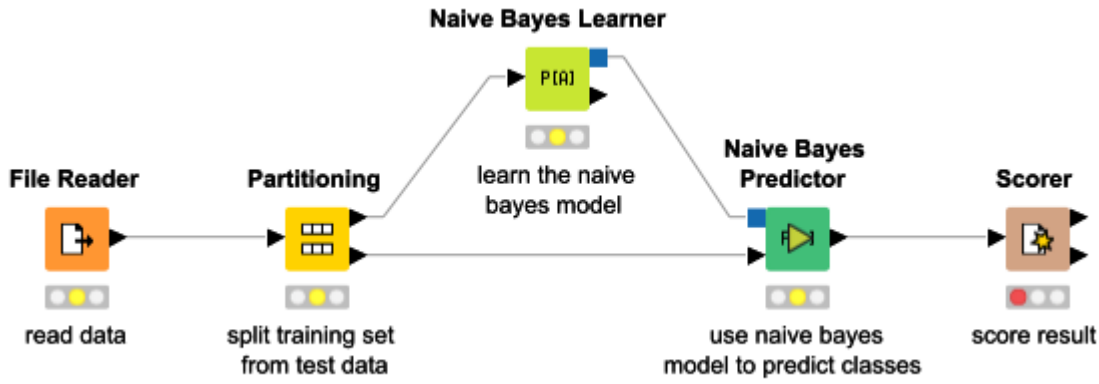
where Ck is the number of observations in category k, and ND is the sample size

- If a feature A separates the sample D into two samples D1 and D2, then the Gini index is:

$$G(D, A) = \frac{ND1}{ND} * G(D1) + \frac{ND2}{ND} * G(D2)$$

where ND1 and ND2 are the sample sizes of D1 and D2, respectively. A smaller Gini index indicates better results.

4.5 Naïve Bayes:



Naïve Bayes is a probabilistic algorithm used for classification problems in machine learning. It assumes that the features are independent of each other, which simplifies the calculations and makes it computationally efficient.
During training, Naïve Bayes calculates the conditional probability of each feature given each class, based on the training data.

- It contains the two probabilities Prior probability of a class label, conditional probabilities.
- For a new input record with features f1, f2, ..., fn, calculate the posterior probabilities for each class label and select the class label with the highest probability:
- P(class label | f1, f2, ..., fn) = P(class label) * P(f1 | class label) * P(f2 | class label) * ... * P(fn | class label)
- Select the class label with the highest posterior probability as the predicted class label.

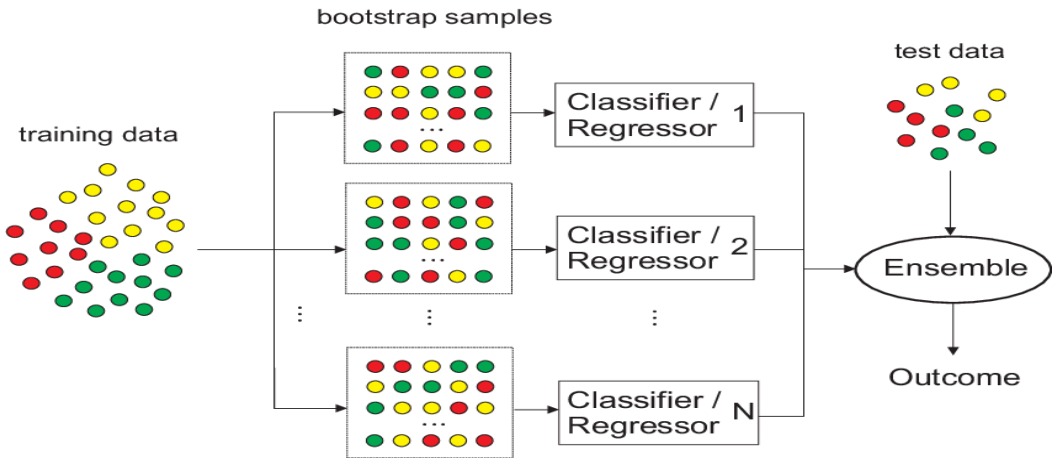
4.5.1. Gaussian Naïve Bayes (PCA)

Gaussian Naïve Bayes with PCA is a machine learning technique that involves applying PCA to reduce the dimensionality of a high-dimensional dataset, and then using Gaussian Naïve Bayes algorithm to classify the data. PCA reduces the noise and redundancy in the data, while Gaussian Naïve Bayes calculates the probability of each feature for each class based on the Gaussian distribution of the features.

4.5.2. MultinomialNB

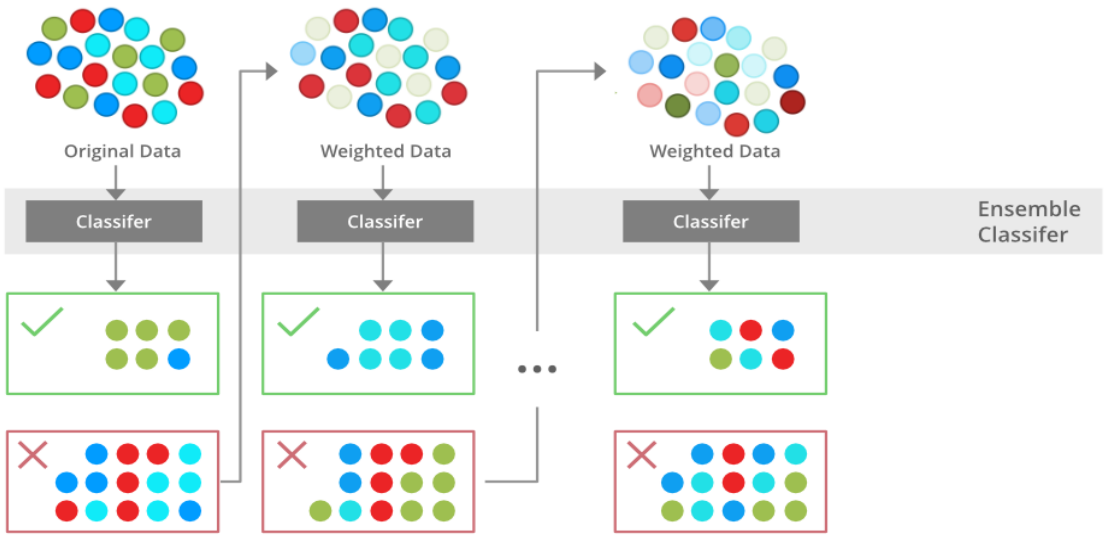
Multinomial NB is a variant of Naïve Bayes specifically designed for discrete data with a multinomial distribution, such as text classification. It assumes that the features are generated from a multinomial distribution and calculates the probabilities of each class based on the frequency of feature occurrences.

4.6 Bagging:



- Bagging (Bootstrap Aggregating) is a machine learning technique used for classification problems that involves training multiple classifiers on different subsets of the training data and aggregating their predictions to make a final prediction.
- During training, Bagging randomly selects subsets of the training data with replacement and trains a classifier on each subset. The predictions from each classifier are then combined using voting or averaging to make the final prediction.

4.7 Boosting:



- Boosting is a machine learning technique that combines several weak learners into a strong learner.
- The goal of boosting is to improve the overall accuracy of the model by combining the results of multiple weak learners.
- In boosting, each weak learner is trained on a subset of the data and the results are combined to create a more accurate model.

Types of Boosting:-

4.7.1. Adaboosting:

AdaBoost is a machine learning algorithm that combines multiple "weak" learners to create a strong learner.

It works by iteratively training a series of weak classifiers on the same dataset, giving more weight to misclassified examples.

The final output of AdaBoost is a weighted sum of the individual weak classifiers.

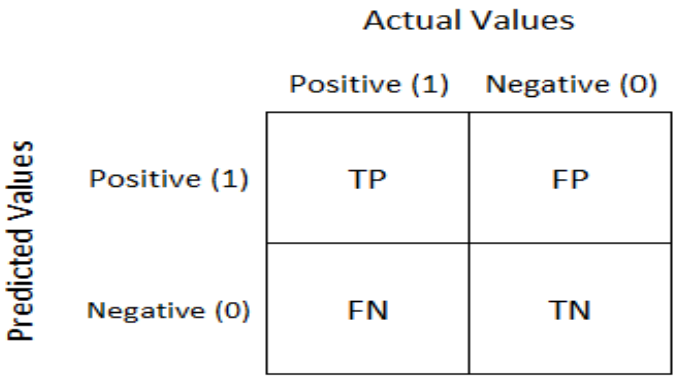
4.7.2. Gradient Boosting:

Gradient Boosting is a machine learning technique that combines multiple "weak" learners to create a strong learner, It works similar to Adaboost .

Gradient Boosting is known for its high predictive accuracy.

4.8 Confusion Matrix:

It is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values. It is useful for measuring True Positive Rate TPR and False Positive Rate. It is also used for plotting AUC-ROC curves.



TP (True Positive): The observations which are positive in real and predicted as positive are regarded as true positive values.

FP (False Positive): The observations which are negative in actual and are predicted as positive are regarded as false positive.

FN (False Negative): The observations which are positive in real and predicted as negative are regarded as false negative.

TN (True Negative): The observations which are negative in actual and predicted as negative are regarded as true negative.

4.9 Precision: The rate of values that were actually positive and predicted as positive.

It is given by: Precision = $\frac{TP}{TP+FP}$

4.10 Recall: The rate of predicted positive values that were positive actually.

It is given by: Recall = $\frac{TP}{TP+FN}$

4.11 F1-score: The F1-score is the weighted average of precision and recall.

It is given by: F1-score = $\frac{2*Precision*Recall}{Precision + Recall}$

4.12 AUC-ROC Curves:

Graphical representation of performance.

Dotted line is performance of random guessing model.

Best model passes through the left corner of the graph.

ROC Curve is the **plot of True Positive Rate vs False Positive Rate**. An Area under ROC Curve of

AUC (Area under the curve) is used to evaluate the performance of binary classification. An Area under ROC Curve of greater than 0.5 is acceptable.

4. Data

Data Structure:

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss	hours-per-week	native-country	income
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Data Description

The total number of rows in the dataset are 48842.

There are 15 columns in the dataset.

There is no null values present in dataset.

Imbalance Dataset

Imbalance in the UCI adult dataset means that there are unequal numbers of people earning above and below $50,000 annually. This imbalance can cause problems when training machine learning models, as they may favor the larger group. To address this, techniques like oversampling or undersampling can be used to balance the dataset and improve model accuracy for both income groups.

Oversampling

Oversampling technique is used to address the issue of imbalance dataset**.** By analysing the dataset we found that dataset is imbalance. So we decided to use oversampling technique to address this issue. We used SMOTE oversampling technique.

Variables Description:

The UCI Adult dataset contains several variables that provide information about individuals' attributes. Here is a description of the commonly included variables in the dataset:

X1. Age: Represents the age of the individual in years (numeric).

X2. Workclass: Describes the type of work the individual is engaged in. Categories may include Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked (categorical).

X3. fnlwgt: Final weight means total number of individuals of same demographic features.

X4. Education: Indicates the highest level of education attained by the individual. Examples of categories include Bachelors, Masters, Doctorate, etc. (categorical).

X5. Education-Num: Represents the numerical encoding of education levels for convenience in analysis (numeric).

X6. Marital Status: Describes the marital status of the individual, such as Married-civ-spouse, Divorced, Never-married, etc. (categorical).

X7. Occupation: Represents the type of occupation the individual is employed in. Examples may include Prof-specialty, Craft-repair, Exec-managerial, etc. (categorical).

X8. Relationship: Indicates the role of the individual in the family. Categories can include Husband, Wife, Own-child, Not-in-family, etc. (categorical).

X9. Race: Describes the individual's race or ethnic background, such as White, Black, Asian-Pac-Islander, etc. (categorical).

X10. Sex: Indicates the gender of the individual (categorical).

X11. Capital Gain: Represents the capital gains the individual has made from investments (numeric).

X12. Capital Loss: Indicates the capital losses the individual has incurred from investments (numeric).

X13. Hours per week: Represents the number of hours the individual works per week (numeric).

X14. Native Country: Describes the individual's country of origin or citizenship (categorical).

X15. Income Class: The target variable, indicating whether the individual's income exceeds $50,000 per year. This variable is often represented as '>50K' and '<=50K' (categorical).

These variables capture various aspects of an individual's demographic characteristics, education, occupation, and other relevant factors. They are used to predict the income class of the individual based on the given attributes.

6. Numerical Analysis

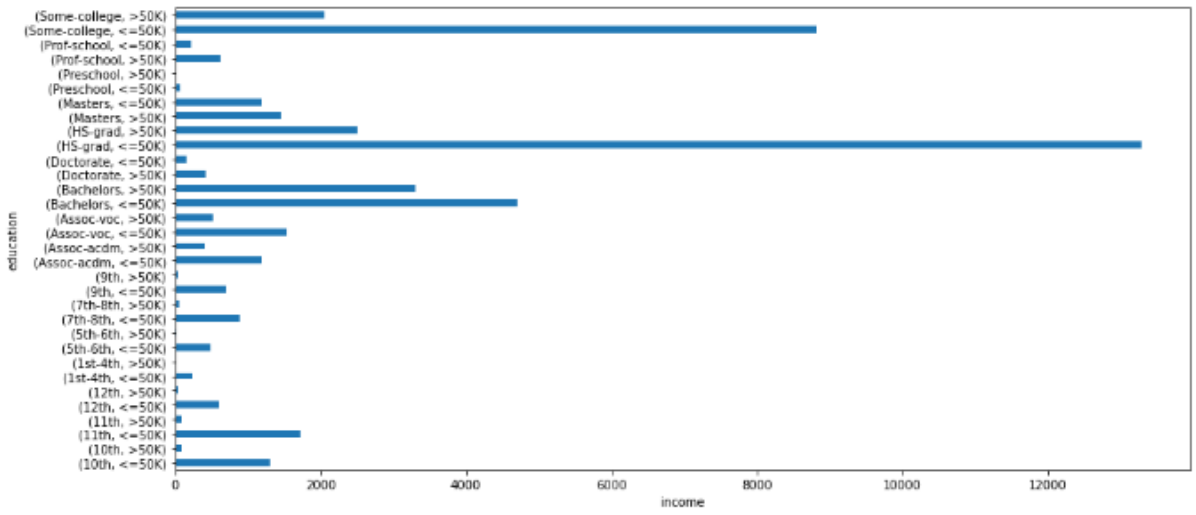


Fig.1 Education-Income vs Count of individuals

Maximum people who are earning less then 50k are in high school.

Maximum people who are earning more then 50k are in Bachelors.

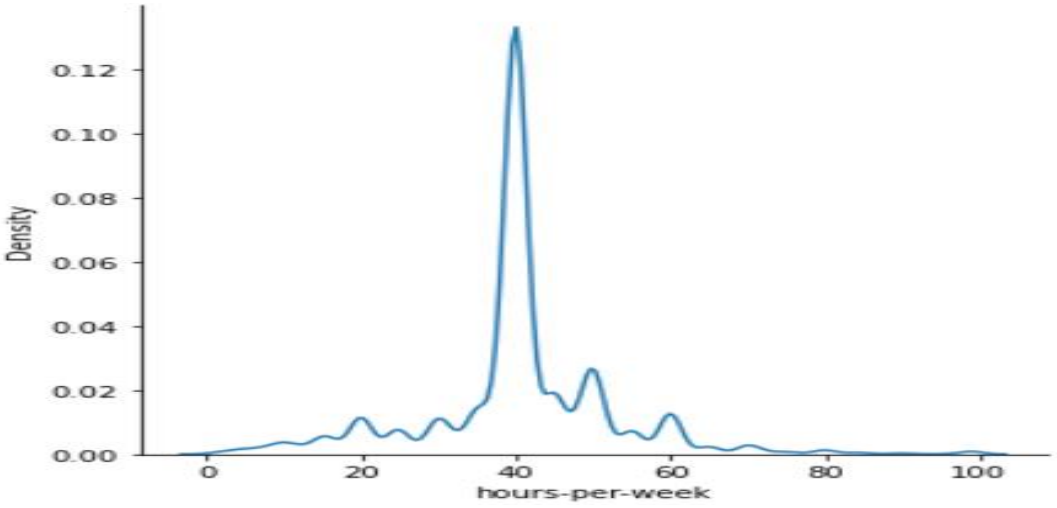


Fig.2 Density Plot

It is been observed that most of the people are working between 35 to 45 hours per week.

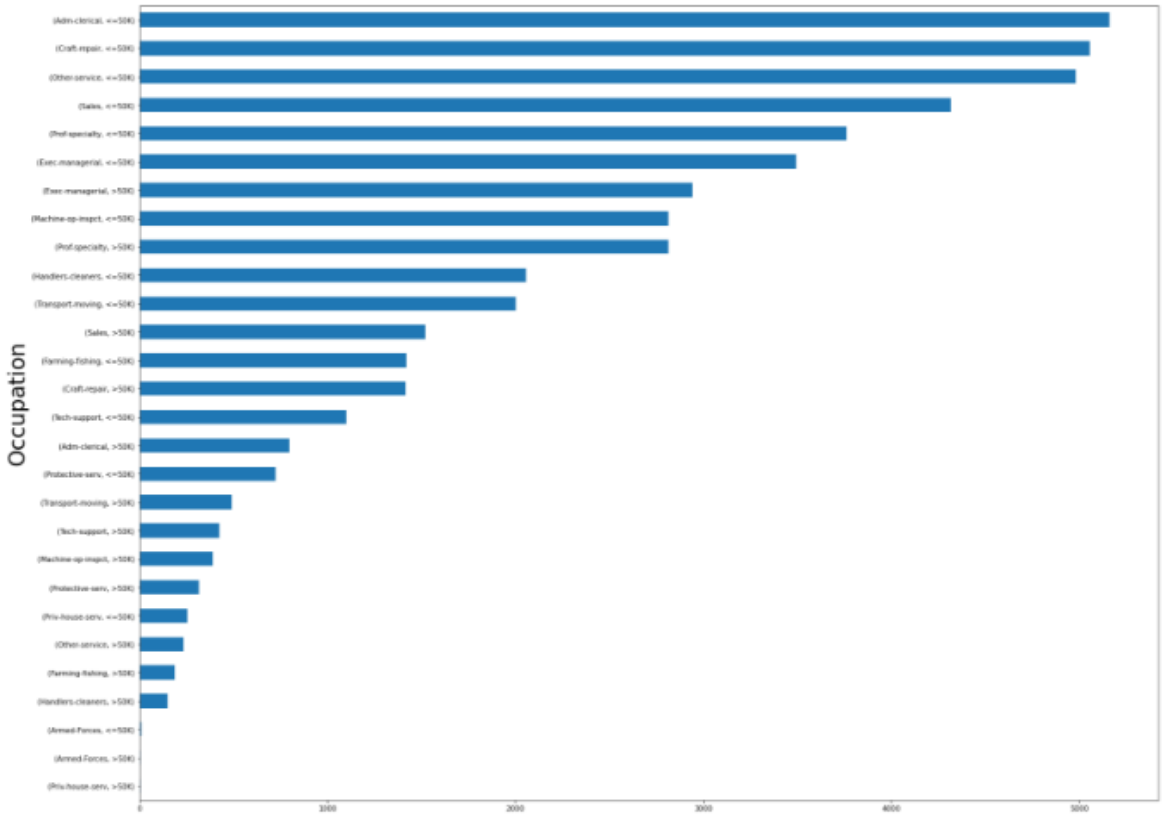


Fig.3 Occupation-Income vs Count of individuals

Occupations in which maximum people are earning >50K are

Exec - managerial and Prof. speciality

Occupations in which maximum people are earning <=50K are

Adm-clerical, Craft-repair and Other-service.

Occupations in which Minimum people are earning <=50K are

Armed-Forces.

Occupations in which Minimum people are earning >50K are

Priv-house-serv.

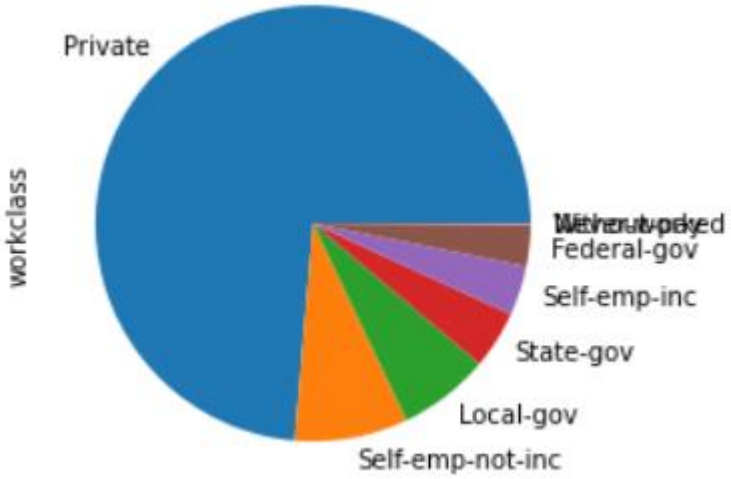


Fig.4 Pie-Chart

Most of the employee belongs to private jobs category.

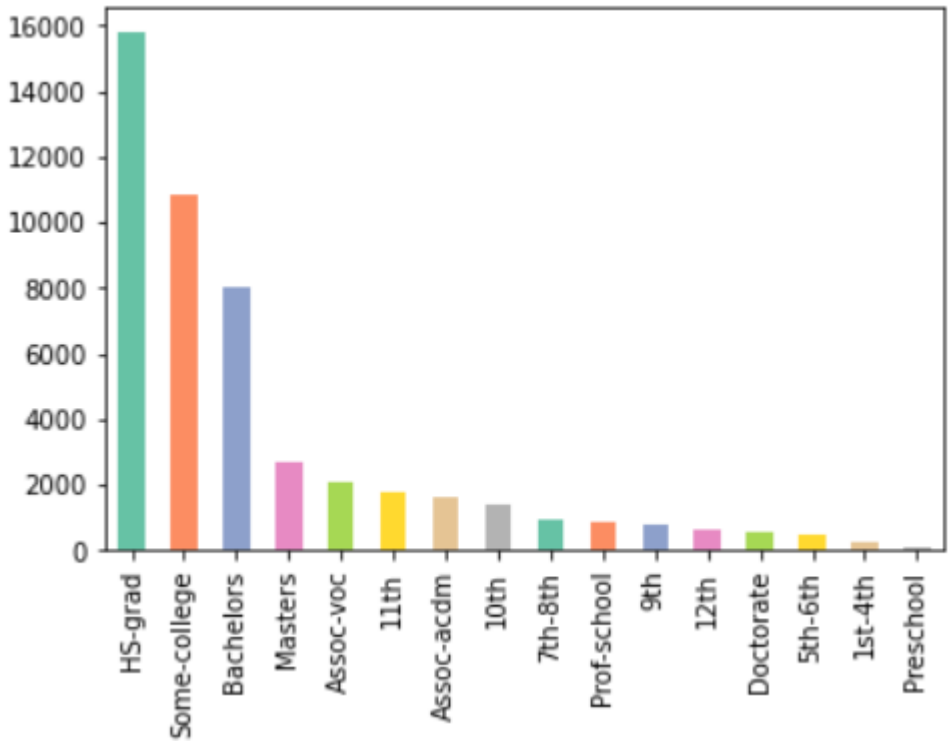


Fig.5 Number of Employees vs Education Bar-Plot

Most of the employees have completed HS-grad education

HS-grad	12291
Some-college	8150
Bachelors	5496
Assoc-voc	1548
11th	1537

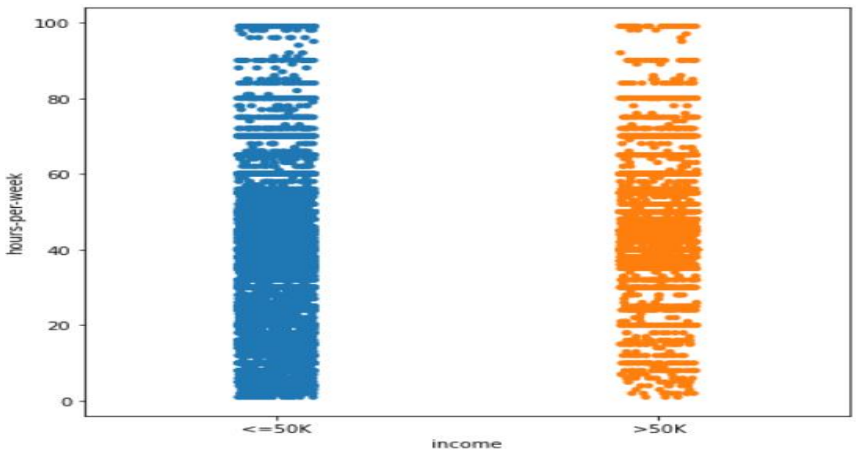


Fig.6 Hours-per-week vs Income Strip-plot

Even though the income vary a lot though the working hour for high and low income group is similar.

Bivariate Analysis

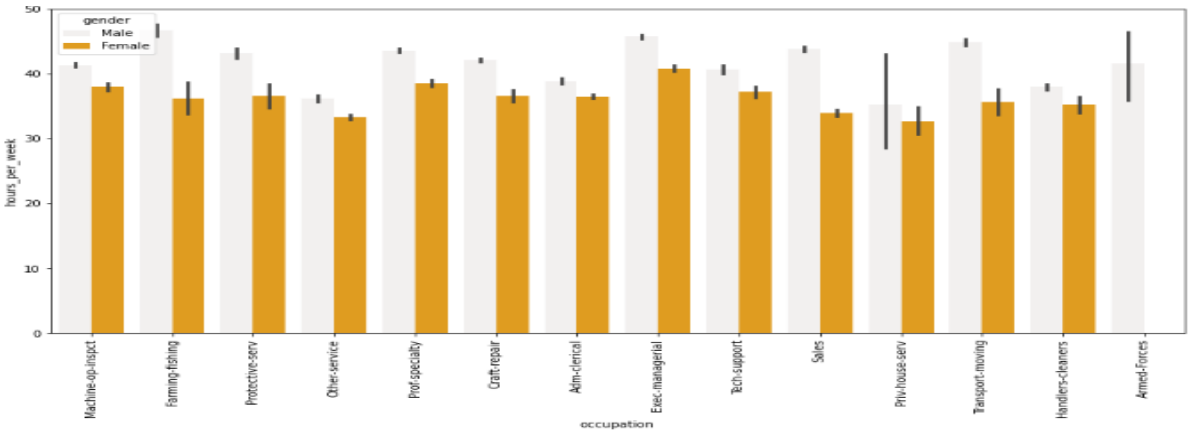


Fig.7 Hours-per-week vs Occupation Bar-Plot

Males with occupation of Farming- Fishing are working for longer period of time.

Womens with occupation of Exec-managerial are working for longer period of time.

There are no womens in Armed Forces.

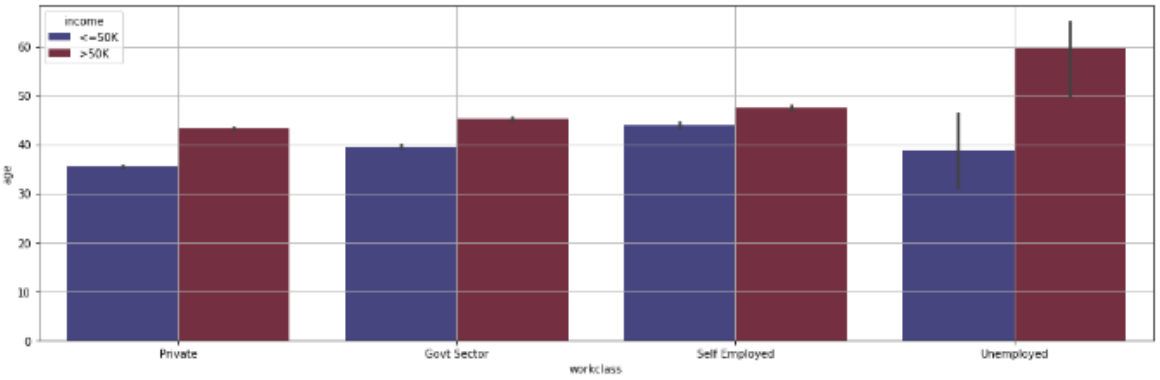


Fig.8 Age vs Income according to workclass Bar-Plot

People earning more then 50K are older in age compare to people earning less then 50K and Unemployed.

People working in govt sector seem to be having slighter more income compared to private sector.

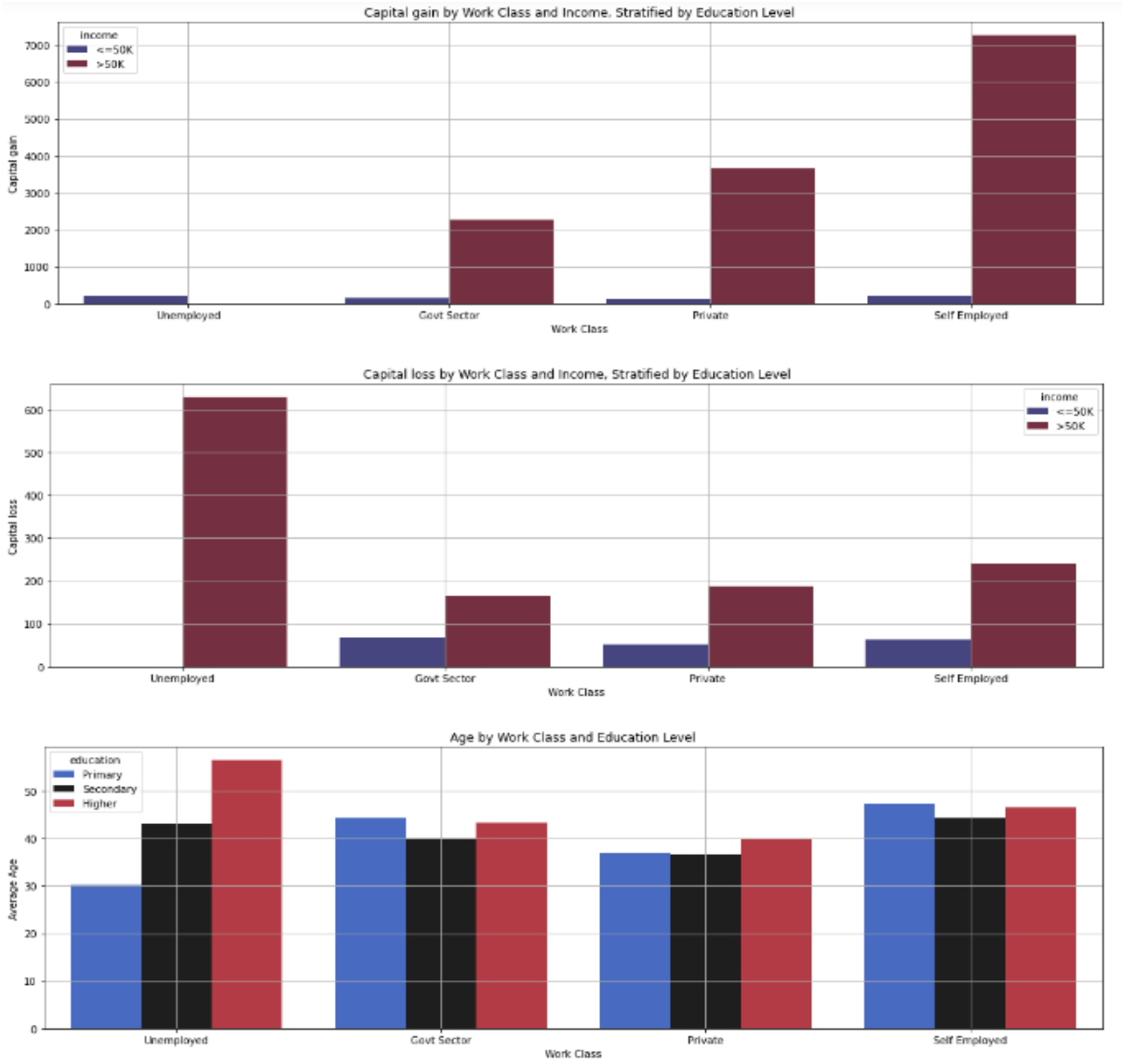


Fig.9 Capital gain and Capital loss

From first graph it is clear that from Self employed sector whatever persons having income greater than 50k are making huge profit.

From second graph it is clear that from unemployed section whatever peoples having income greater than 50k are making huge loss.

From third graph we observe that in Unemployed Section peoples are having higher education and are older there is a possibility they have done savings so that might be the reason they are having high income.

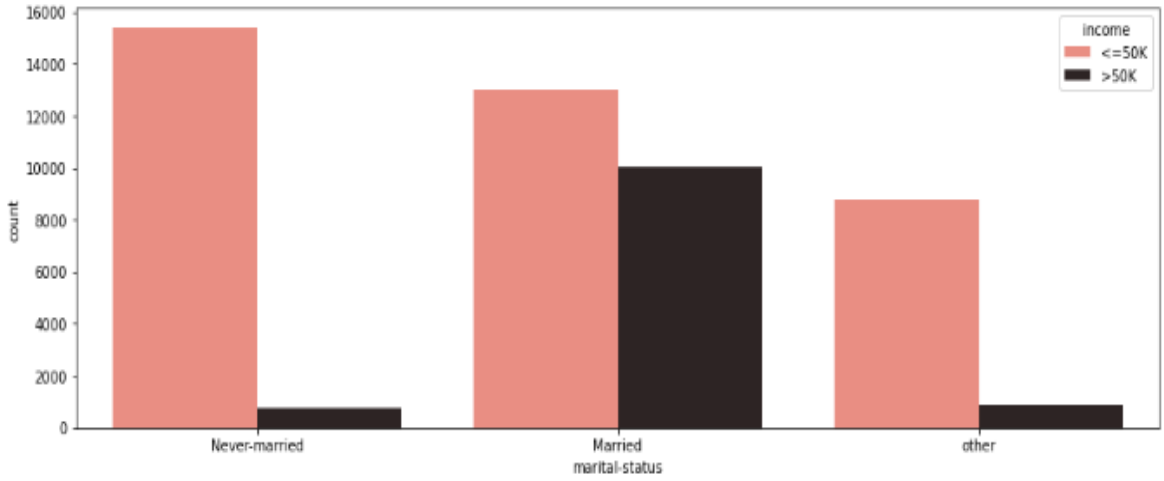


Fig.10 Number of Individuals vs Income According to Marital-status Bar-Plot

Above graph tell us that married person has high chance to earn more.

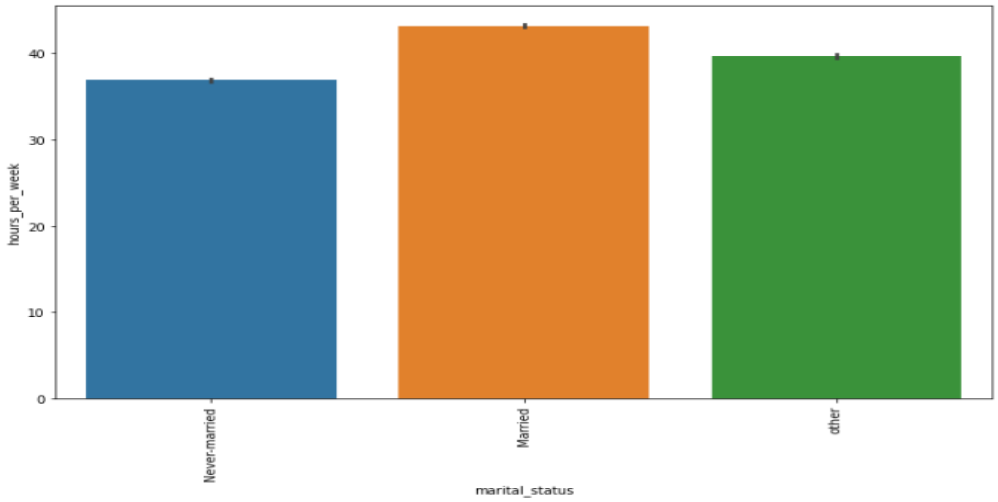


Fig.11 Hours-per-week vs Marital-status Bar-Plot

Married peoples work for long time as compared to unmarried and other peoples.

Using Chi-Square Testing

Relation between occupation and gender

Chi-Square Statistics = 8106.67
p-value = 0.0

Decision: If p-value < 0.05, then we can reject H0.

Here, p-value = 0.0 < 0.05

Conclusion: Therefore, There is some significance relation between the two Variables.

Relation between income and gender

Chi-Square Statistics = 2248.8
p-value = 0.0

Decision: If p-value < 0.05, then we can reject H0

Here, p-value = 0.0 < 0.05

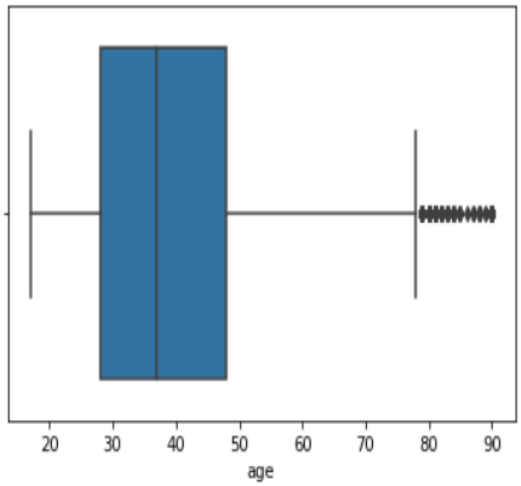
Conclusion: Therefore, There is some significance relation between the two Variables.

Detecting Outliers

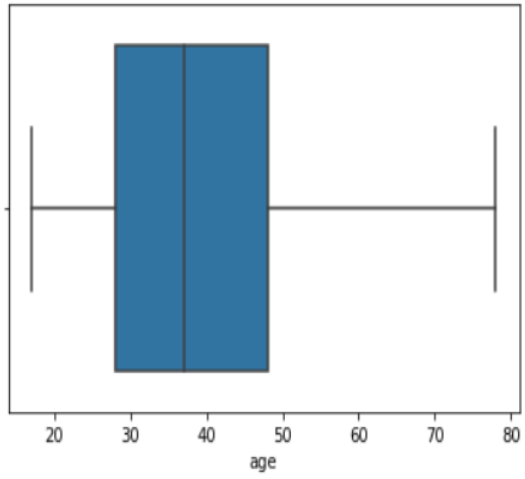
We used interquartile range to handle the outlier.

Outliers in age:

Before removing Outlier:

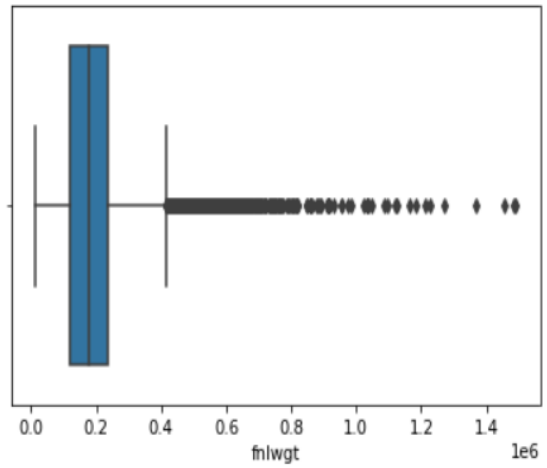


After removing Outlier:

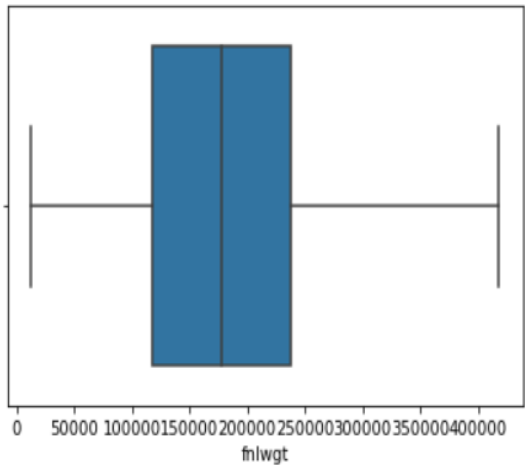


Outliers in fnlwgt i.e. (final count of individuals):

Before removing Outlier:

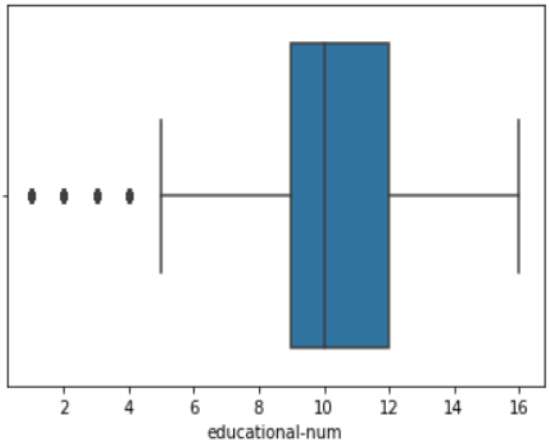


After removing Outlier:

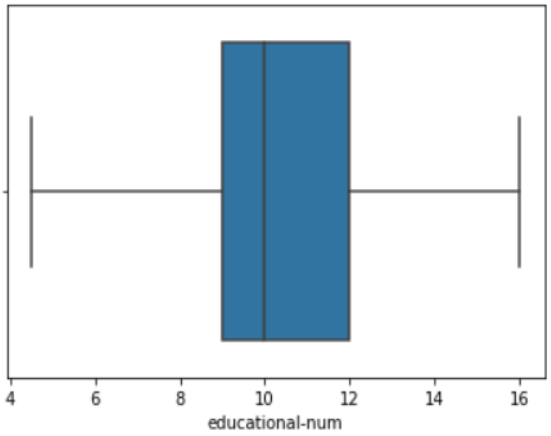


Outliers in educational-num:

Before removing Outlier:

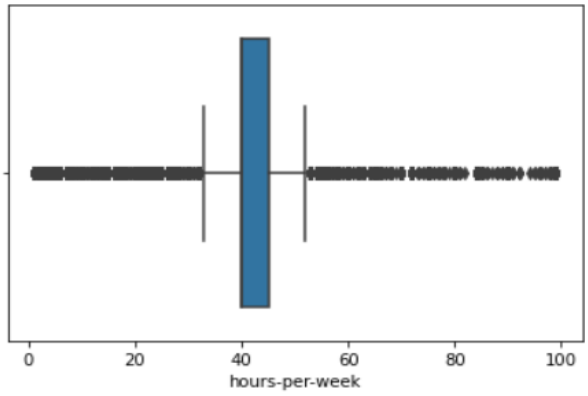


After removing Outlier:

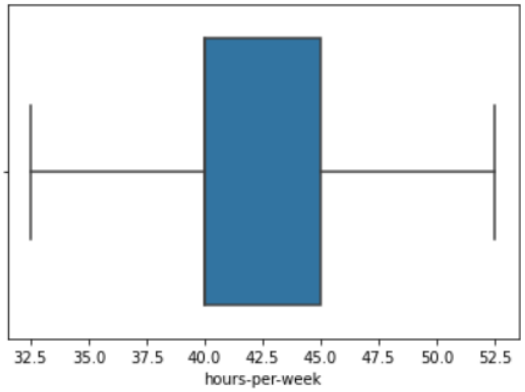


Outliers in hours-per-week:

Before removing Outlier:



After removing Outlier:



7. Detailing of Outputs

Models	Accuracy	Precision	Recall	F1-Score	Confusion Matrix
AdaBoosting	0.97	0.98	0.97	0.97	[[24155 523] [767 23911]]
Gradient Boosting	0.90	0.90	0.91	0.90	[[22108 2570] [2251 22427]]
Bagging	0.87	0.85	0.91	0.88	[[20587 4091] [2301 22377]]
Decision Tree	0.86	0.84	0.89	0.86	[[20451 4227] [2559 22119]]
Randome Forest	0.85	0.81	0.90	0.86	[[19485 5193] [2200 22478]]
SVM	0.83	0.79	0.89	0.84	[[18829 5849] [2707 21971]]
Logistic Regression	0.80	0.79	0.81	0.80	[[19281 5397] [4635 20043]]
Multinomial Naïve Bayes	0.74	0.71	0.82	0.76	[[16113 8565] [4304 20374]]
Gaussian Naïve bayes (PCA)	0.73	0.73	0.75	0.74	[[17725 6953] [6212 18466]]

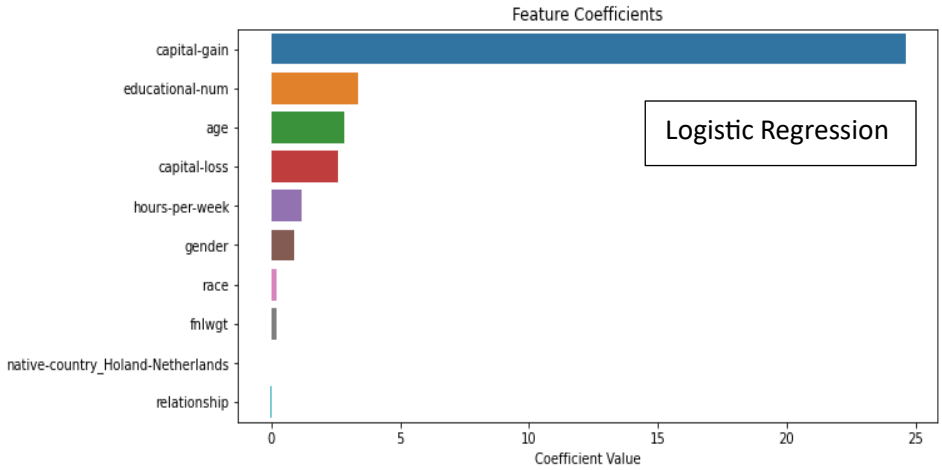
From the above results AdaBoosting have higher accuracy.

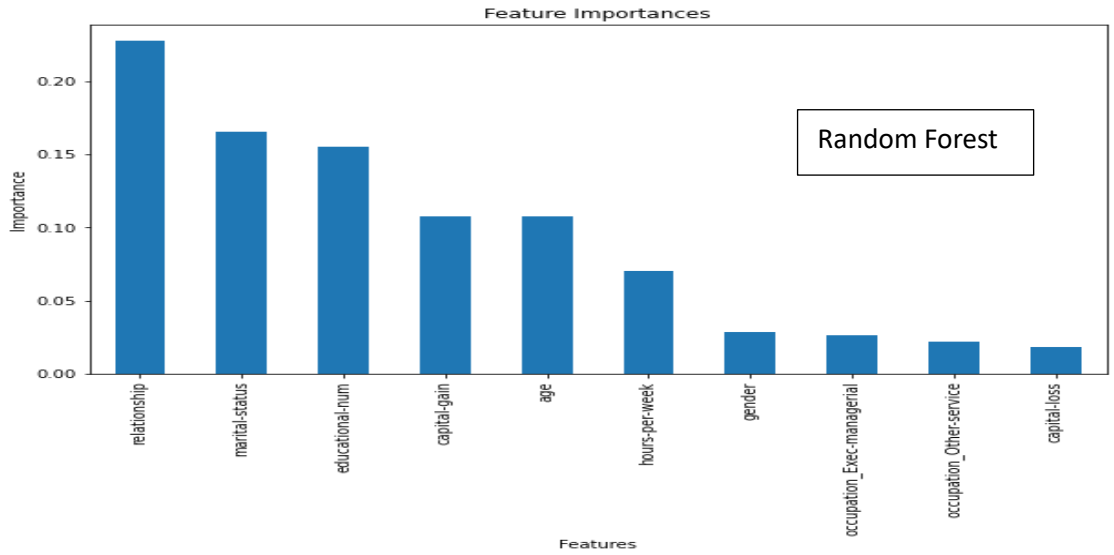
Actual	Logistic	SVM	PCA+GNB	MNB	rf	adboost	boost	bag	dt
0	1	0	0	0	0	0	0	0	1
0	1	1	1	1	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	1	0	1	1	1

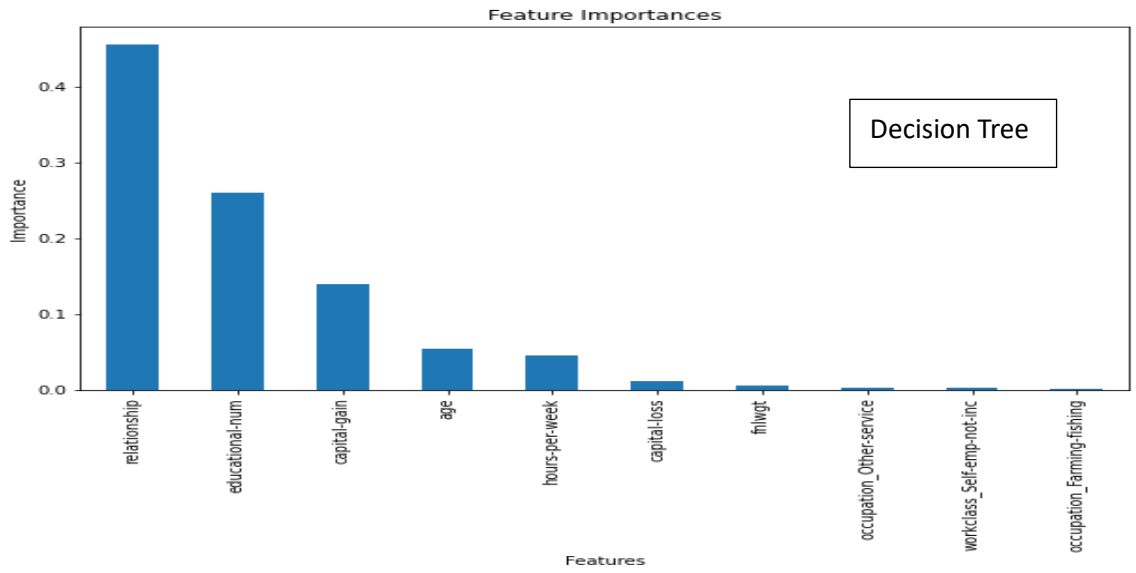
0 : <=50K & 1 : >50K

Interpretation: From table for the first record actual income is less than $ 50K, logistic regression predicts greater than $ 50K similarly for other records other models also given wrong prediction for some instances, while AdaBoost predicted accurately for all the records. So that we can say AdaBoost is performing well on our dataset.

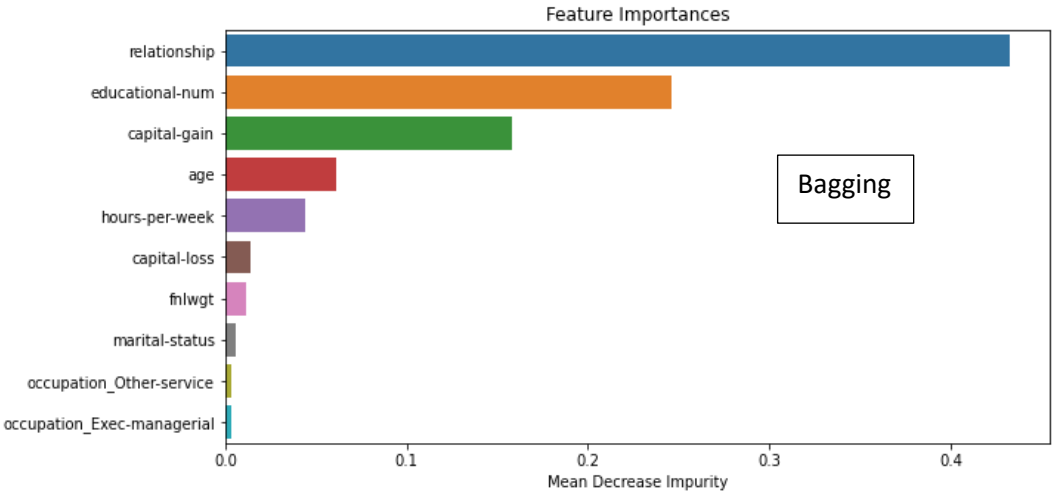
Feature Selection Graphs:

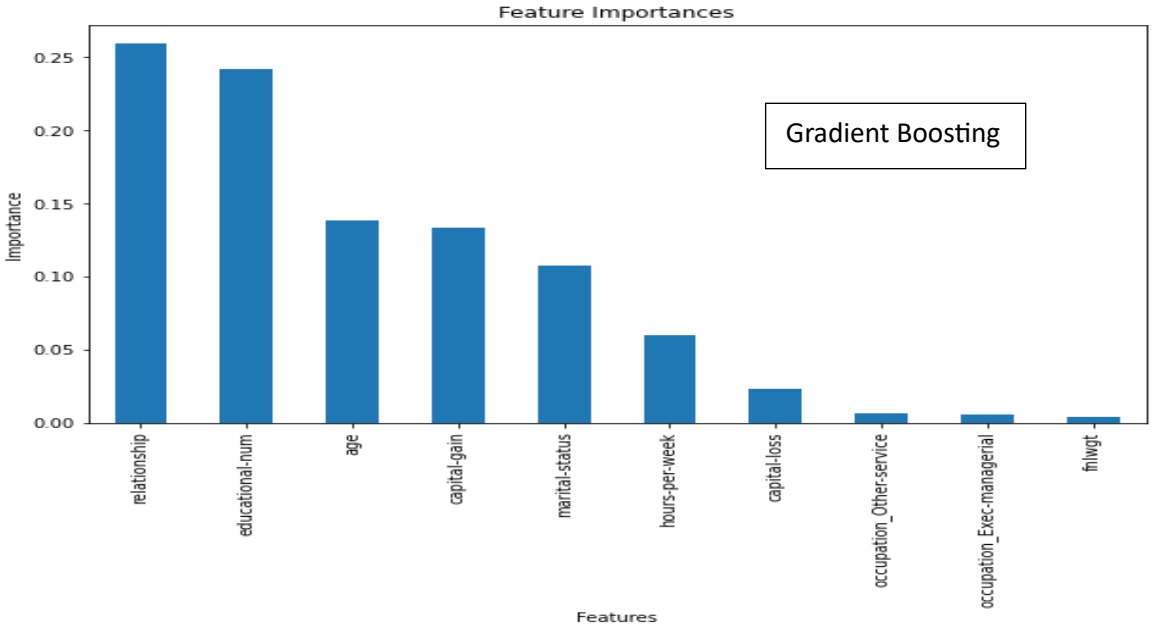


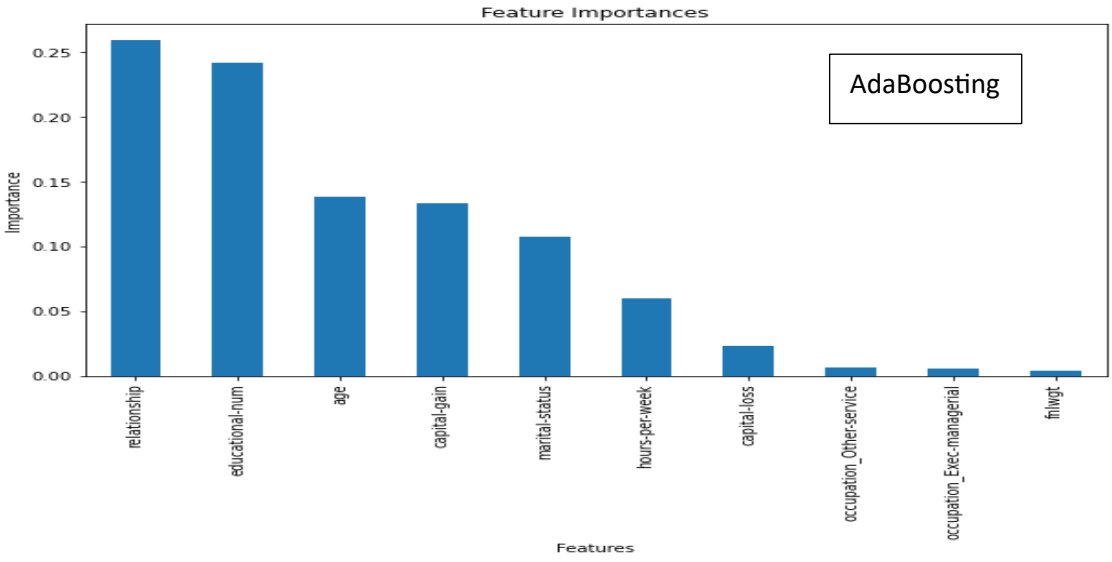




According to logistic regression capital-gain is important feature whereas according to random forest and decision tree relationship is important feature.







According to bagging, gradient boosting and AdaBoosting relationship is important feature.

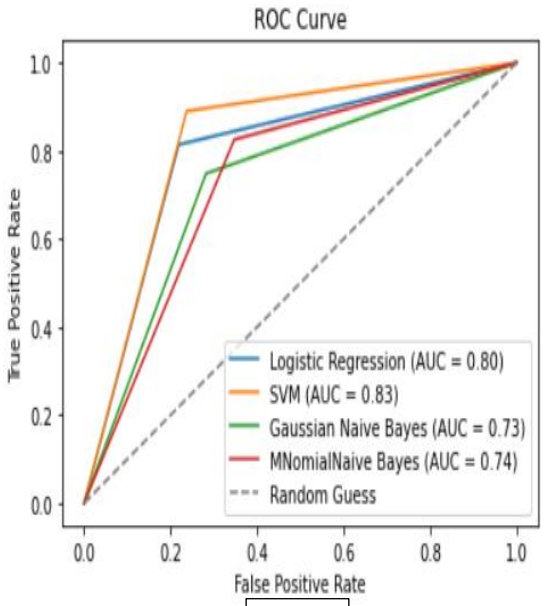


Fig.1

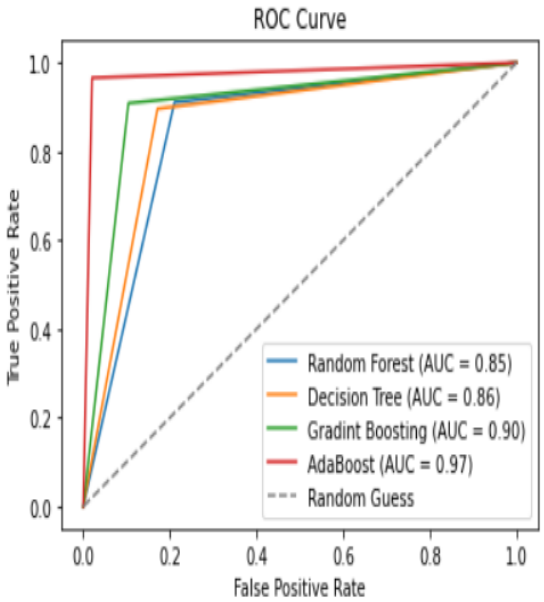


Fig.2

From Fig.1 SVM and logistic regression shows higher AUC score.

From Fig.2 Gradient Boosting and AdaBoosting shows higher AUC score.

8. Overview of best fitted model

AdaBoosting

Confusion Matrix:

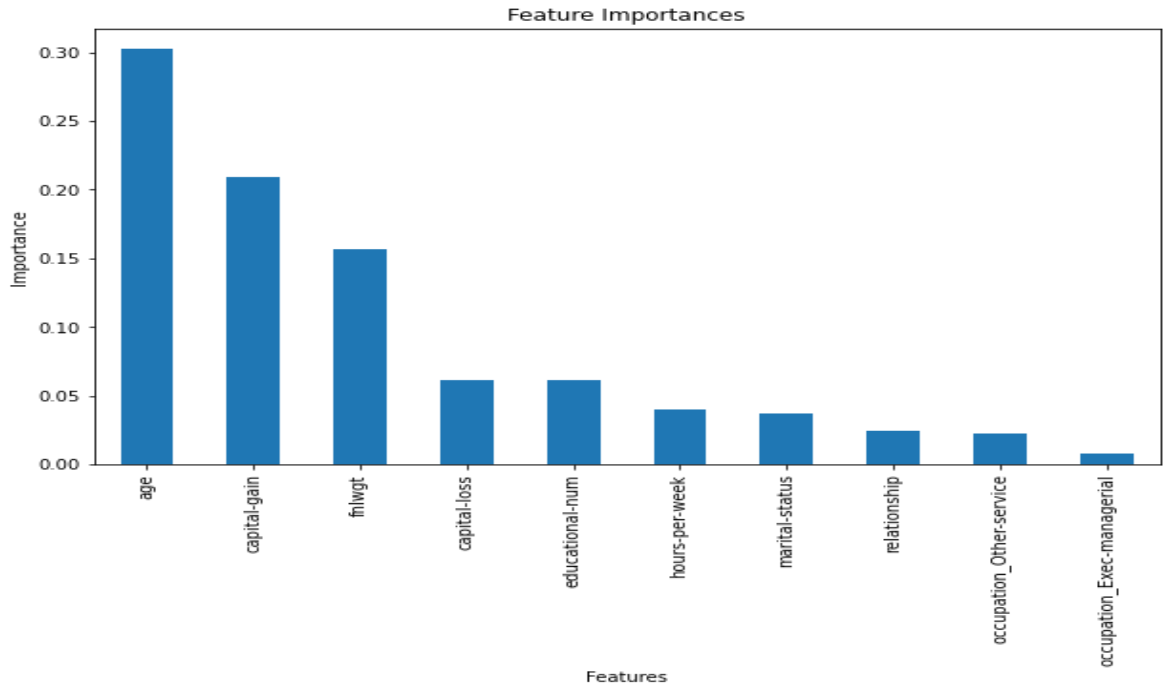
	Predicted Class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive = 24155	False Negative = 523
	Class = No	False Positive = 767	True Negative = 23911

Accuracy : 0.97

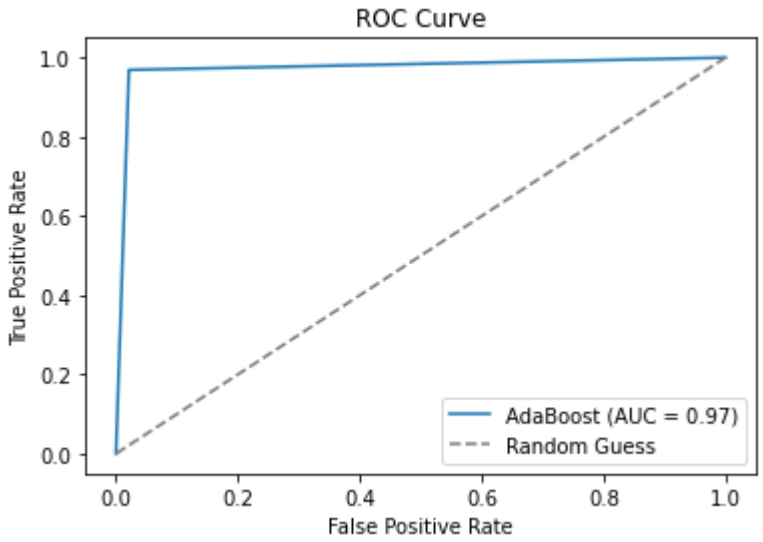
Precision : 0.98

Recall : 0.97

F1 Score : 0.97



Weak Classifier Parameters: {'ccp_alpha': 0.0, 'class_weight': None, 'criterion': 'gini', 'max_depth': 10, 'max_features': None, 'max_leaf_nodes': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'random_state': 1608637542, 'splitter': 'best'}



Odds Ratio : [0.37377079 0.81178254 0.79277712 ... 1.13261732 1.36431482 1.22622345]

The odds ratio represents the ratio of the predicted probabilities of the positive class (label 1) to the predicted probabilities of the negative class (label 0) for each instance in the test set.

- **Odds Ratio < 1:**

The model is more confident in classifying these instances as the negative class.

- **Odds Ratio > 1:**

This means that the model is more confident in classifying these instances as the positive class.

- **Odds Ratio = 1:**

The model does not favor one class over the other and considers them equally likely.

For the greatest Odds Ratio = **0.81178254 < 1**

Therefore, The model is more confident in classifying these instances as the negative class.

9. Conclusions

- With the help of feature selection technique in each model we have concluded that the features such as relationship , age , education , Marital status, etc . This are the most important features for determining the individuals income class.

- We apply logistic regression, Decision Tree , Random forest, Bagging, SVM, Naïve Bayes and boosting to solve this problem. After comparing the performance of these models, we find that boosting (Adaboost) provides the highest AUC value also have highest accuracy , which indicates that it is the best model for fitting the data.

10. Future Work

From this Adult UCI Data we Extract India's Data and fit AdaBoosting on that, we get good accuracy

Confusion Matrix :

	Predicted Class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive = 144	False Negative = 5
	Class = No	False Positive = 14	True Negative = 135

Accuracy: 0.94

Precision: 0.96

Recall: 0.91

F1 score: 0.93

So we will Fit this Model on the India's real time data.

11. Limitations

- We extracted India's data from the UCI adult dataset, and we predict the income from this India's data using AdaBoost, it gives better accuracy but this model is useful only for observations before the 2012.

- The dataset primarily focuses on the United States, which restricts the generalizability of the findings to other countries or regions with different socioeconomic conditions and cultural contexts. The insights derived from the analysis may not be applicable universally.

References

Chen, L.P. (2017). Supervised Learning for Binary Classification on US Adult Income. Journal of Big Data, 4(1), 1-13. Doi: 10.1186/s40537-017-0083-0.

Zaid, M., & Rajendra, T. (2019). Higher Classification Accuracy of Income Class Using Decision Tree Algorithm over Naïve Bayes Algorithm. International Journal of Scientific & Engineering Research, 10(3), 1399-1403.

Alina Lazar (2004). Income prediction via support vector machine,
DOI: 10.1109/ICMLA.2004.1383506 Source DBLP - ICMLA 2004,
16-18 December 2004, Louisville, KY, USA.

Gomez-Cravioto et al. Journal of Big Data (2022). Supervised machine learning predictive analytics for alumni income, Doi.org/10.1186/s40537-022-00559-6

Navoneel Chakrabarty and Sanket Biswas (2018). A Statistical Approach to Adult Census Income Level Prediction DOI: 10.1109/ICACCCN.2018.8748528