**Introduction**:

Machine Learning (ML) by looking at the definition, we can say that it is a field of computer science. It is the learning and building of algorithms that can learn from it and make predictions on datasets. The problem of income inequality has been a major concern in the past few years. This model is built so that we can conduct an analysis and understand some of the important factors which are necessary to improve the individual’s income.

**Problem** **Definition**:

We will use Logistic Regression for this dataset to build the model. The data contains age, education, occupation, relationship, race, sex, work class etc. Here, we need to predict the income which has two values >50k and <=50k. There are 32560 rows and 15 attributes. The data contains numerical and categorical values. Most importantly, there are no missing values in this dataset.

**Exploratory Data Analysis (EDA):**

EDA is a phenomenon under data analysis is used for gaining a better understanding of data aspects like a.)Main features of data. b.) Variable and relationships that hold between them. c.) Identifying which variables are important for our problem.

There are 32560 samples in this dataset. There are both categorical and numerical values in this dataset. The target variable income is imbalanced. We will balance our target variable by using random oversampling which will add copies to the minority class and therefore, our class will be balanced. Fnlwgt is not having any pattern with the target variable income and it will be dropped before we build the model. The columns education is also a string representation of the columns education\_num. So, we will be dropping the education column. We will be plotting the distribution plots, box plots, count plots etc. While plotting the distribution plots, if there is any skewness present in it then we remove it by applying some transforms like sqrt, cbrt, log and then we need to make sure that the skewness range is between -0.5 to +0.5 and we plot the distribution plot again without any skewness. We plot the boxplots to detect if there are any outliers present in them or not. We will also plot the histogram plots for all the column to check whether there is any skewness present in each of the features or not. Multiple libraries are available to perform basic EDA such as pandas, matplot library and seaborn.

**EDA Concluding Remarks:**

After doing all the data cleaning process, I am successful in removing all the skewness from the data and I removed all the outliers which were present in the data by using Z-score.

**Pre-Processing Pipeline:**

The pre-processing of data involves three steps namely data cleaning, feature selection and data transformation. The data cleaning involves missing data. We need to replace the missing values by using either mean, median or model. Before, training the model feature selection is one of the important factors that can influence the model’s performance.

In the processing steps with the dataset, I have cleaned all the data with techniques like:

1. Handling the missing values which are not necessary as there are no missing values present in the dataset.

2. Encoding the categorical variables which were done by using a label encoder as the categories are assigned starting from 0.

3. I have shuffled the data such that all my training data is on one side and my testing data is on another side.

4. I have split the data into training and testing sets. I have made 70% of the data available for training and 30% of the data available for testing.

**Building Machine Learning Models:**

I have used Logistic Regression since it is one of the proper model of regression analysis to utilize when our dependent variable is binary. I have also used some of the classification algorithms to check the accuracy of our data and check which model is performing better when compared to the other models. Firstly, I am giving a basic introduction of all the models which I have used in my dataset.

1. Logistic Regression: It works when there are only two values i.e. 0 and 1. It is estimating the parameters of a logistic model which is a form of binary regression.

2. Gaussian Naïve Bayes: It is a supervised machine learning algorithm. It is a classification technique that works only when there are binary values.

3. K Nearest Neighbor: This algorithm is working by calculating the distance of the nodes and we call this distance Euclidian distance. Here, k= number of neighbors.

4. Support Vector Machine (SVM): In SVM, we have a parameter called a kernel. We can give different values to the kernel to improve the accuracy. Radius base function (RBF), linear and poly are kernels in svc.

5. Decision Tree: It is one of the classification techniques. It works better with classification problem. We have two parameters in the decision tree i.e. Gini and entropy.

6. Random Forest: It is a supervised machine learning technique. It is an ensemble learning method for classification. It also works very well with classification.

Firstly, I have imported all the algorithms which was required to build the model. Then, I have taken the logistic regression model to check on which random state I am getting the best accuracy score such that the random state can be finalized. Then, by using the fit method, I have trained all the models. By using predict method, I have predicted the values for all the models. Predicted data is nothing but the answer given by the x\_test model and y\_test is the actual data. We will use an accuracy score from sklearn to find the accuracy of the model. We will also use a confusion matrix and classification report from sklearn and because of the confusion matrix, we can tell what will be the accuracy of the model. By using a classification report, we will be getting scores like recall, precision, specificity and model performance.

The formulae are as follows:

TP-True Positive, TN-True Negative, FP-False Positive, FN-False Negative

Recall or Sensitivity=TP/TP+FN

Precision=TP/TP+FP

Specificity=TN/FP+TN

Model Performance=F1=2(Precision\*Recall/Precision + Recall)

The accuracy score for logistic regression is 75.76%, accuracy score for gaussian naive bayes is 76.06%, accuracy score for support vector machine is 77.65%, decision tree classifier is 88.55%, accuracy for K Neighbors classifier is 85.10% and accuracy score for random forest classifier is 90.63%. After checking the accuracy scores for all the models, the Random Forest model is giving me the highest accuracy score of 90.63%. But this accuracy score can be because of overfitting. Let’s check with cross-validation.

Cross-Validation: It is a resampling technique with a basic idea of dividing the dataset into training and testing. On one part we train the model and on the second part i.e. testing the data which is unseen for the model. If the model works with good accuracy on test data, it means that the model has not over fitted the training data and it can be trusted with the prediction whereas if it performs with bad accuracy then our model is not to be trusted and we need to tweak our algorithm. We will use cross val score for cross-validation. I have taken the mean of all the scores obtained in each iteration as the final score of our model. The cross val score for logistic regression model is 76.43%, cross val score for gaussian naïve bayes is 75.51%, cross val score for support vector machine is 74.54%, cross val score for decision tree classifier is 87.25%, cross val score for K Neighbors classifier is 81.02% and cross val score for random forest classifier is 88.12%. After checking the cross val score for all the models, the least difference between the highest accuracy score and cross val score says that random forest model is working better when compared to the other models. Now, let’s fine-tune the model by using hyper-parameter tuning.

Hyper Parameter Tuning: We use this technique to improve the accuracy score of the best model that we have chosen. Sklearn comes with Grid Search CV to do the search over specified parameter values for an estimator. It helps to optimize the model’s performance. With the best parameters, the model will identify the patterns within the dataset in a better way. I used the parameters like criterion, max depth, min samples split, max features and n estimators for the random forest classifier model to find the improved accuracy of the model. After tuning the model, the accuracy of the model with the best parameters is around 92.45% which suggests that there is an increase in the performance of the model by 3%.

Saving the model: We are done with testing the model. Now, we will save our model by importing joblib. Now, we need to load the model from the joblib.

**Conclusion:**

We have learned to build a complete machine learning project. We also learned to fine-tune our model and save it for further use. Finally, the validation accuracy which was obtained 91.35% has been the highest ever accuracy achieved so far by any income prediction model so far.