Faculty of Engineering and Technology

Department of Electrical and Computer Engineering

ENCS5341, MACHINE LEARNING AND DATA SCIENCE

Assignment #3

Prepared by:

Sara Ammar-1191052

Al-Ayham Maree-1191408

Instructor: Yazan Abu Farha

Section: 2

Date: 25/1/2024

Introduction

The objective in this assignment is to apply different machine learning models to address a real problem. This assignment will address the problem of predicting income levels using an adult income dataset. The dataset obtained from UCI's machine learning repository, contains various features such as age, education, occupation, and marital status. The task involves a binary classification, determining whether an individual earns more than $50,000 per year or less. The task divides in multiple stages: establishing a baseline with nearest neighbor model using k=1 and k=3, exploring the potential of Random Forest and XGBoost models by tuning different parameters to get the best performance, performing detailed performance analysis, and evaluating using f1, confusion matrix and area under the precision-recall curve to select the best model.

Table of Contents

List of figures

List of tables

I. Data Set

The Adult Income dataset comprises around 48842 records and 15 features, each detailing an individual's demographic and employment information [1]. The features encompass age, workclass specifying the type of employer, fnlwgt representing the population weight, education level, education-num indicating years of education, marital status, occupation type, relationship status, race, gender, capital gains and losses, hours worked per week, and native country. These features are pivotal in predicting whether an individual earns more than $50,000 annually, serving as the target variable for binary classification. The data has both categorical and numerical features, as shown below:

```
Data columns (total 15 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   age              48842 non-null  int64
 1   workclass        46043 non-null  object
 2   fnlwgt           48842 non-null  int64
 3   education        48842 non-null  object
 4   educational-num  48842 non-null  int64
 5   marital-status   48842 non-null  object
 6   occupation       46033 non-null  object
 7   relationship     48842 non-null  object
 8   race             48842 non-null  object
 9   gender           48842 non-null  object
 10  capital-gain     48842 non-null  int64
 11  capital-loss     48842 non-null  int64
 12  hours-per-week   48842 non-null  int64
 13  native-country   47985 non-null  object
 14  income           48842 non-null  object
```

Figure 1: Adult Income dataset information

To use the data in correct way and get best results, preprocessing steps were applied on it. Firstly, the duplicate records were removed, and the missing values "?" were dropped. Before proceeding with the preprocessing processes, descriptive statistics and visualizations were developed to better understand the nature of the data. In the following the Descriptive statistics for numerical and categorical features:

	age	fnlwgt	educational-num	capital-gain	capital-loss	hours-per-week
count	45175.000000	4.517500e+04	45175.000000	45175.000000	45175.000000	45175.000000
mean	38.556170	1.897388e+05	10.119314	1102.576270	88.687593	40.942512
std	13.215349	1.056524e+05	2.551740	7510.249876	405.156611	12.007730
min	17.000000	1.349200e+04	1.000000	0.000000	0.000000	1.000000
25%	28.000000	1.173925e+05	9.000000	0.000000	0.000000	40.000000
50%	37.000000	1.783120e+05	10.000000	0.000000	0.000000	40.000000
75%	47.000000	2.379030e+05	13.000000	0.000000	0.000000	45.000000
max	90.000000	1.490400e+06	16.000000	99999.000000	4356.000000	99.000000

Figure 2: Descriptive Statistics for Numerical Features

	workclass	education	marital-status	occupation	relationship	race	gender	native-country	income
count	45175	45175	45175	45175	45175	45175	45175	45175	45175
unique	7	16	7	14	6	5	2	41	2
top	Private	HS-grad	Married-civ-spouse	Craft-repair	Husband	White	Male	United-States	<=50K
freq	33262	14770	21042	6010	18653	38859	30495	41256	33973

Figure 3: Descriptive Statistics for Categorical Features

Then the distribution of the target variable (Income) was drawn, and the result suggested an imbalanced classification problem that included two categories: a negative case (income less than $50K annually) with the majority of examples and a positive case (income greater than $50K annually) with a minority of examples:

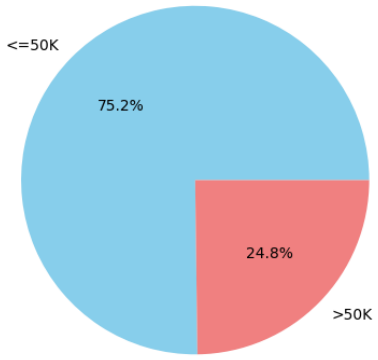
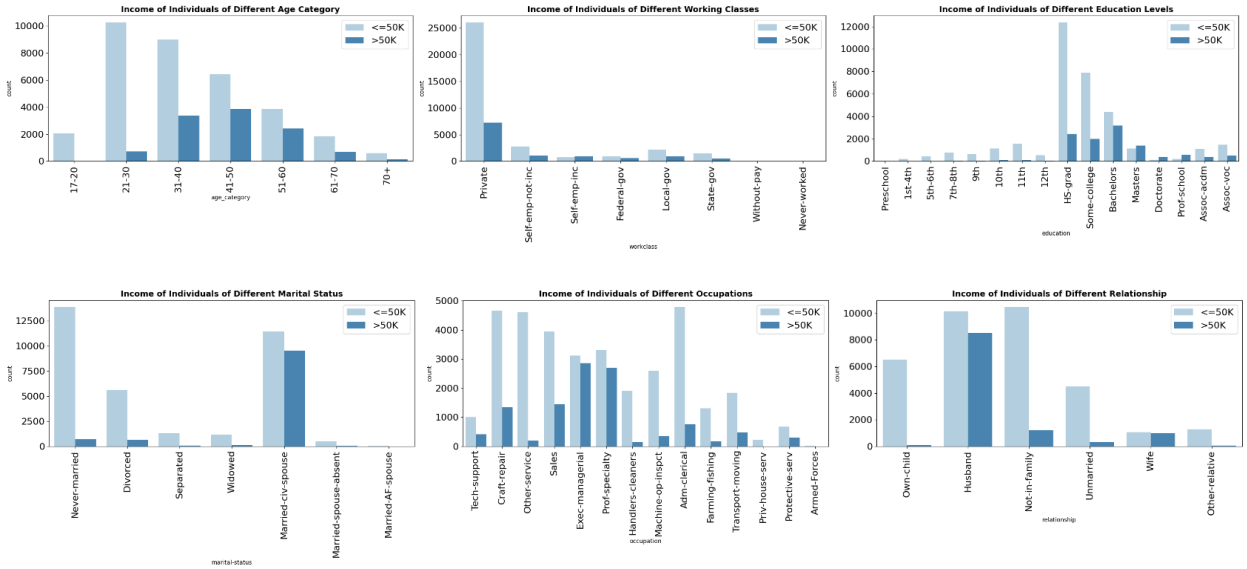


Figure 4: Distribution of Incomes

After that, categorical features and age feature were plotted with income, to make an interesting observation about the relations:



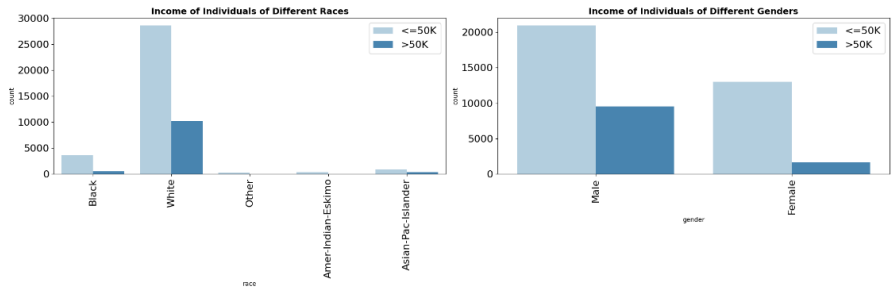


Figure 5: Categorical fetures analysis

People earning over $50K increase with age, with 17-30 earning significantly less, while 41-50 and 51-60 earn nearly to those earning less.

In working classes, Self-employed individuals earn more than $50K, while 75% of private sector workers earn less than $50K annually. Federal government employees earn less than $50K annually, while those without pay and never-worked have minimal data.

Only a few individuals earning over $50K annually after 12th Standard education, while most earn below $50K. Therefore, the school education categories (high school and lower) were combined to be summarized into one category, “School”. Those in Bachelors, Masters, Doctorates, and Prof-schools earn more than $50,000 annually. In case of Assoc-acad or Assoc-voc, only a few earn over $50K annually.

In Marital status, Married-civ-spouses have a comparable number of people in both categories, while less than 25% of adults earn over $50,000 annually.

Adults in Exec-managerial positions are similarly likely to make more than $50K per year, with a 33% probability for Prof-specialty in this field. However, those in farming, fishing, machine-operating, other services, administrative, and transport-moving industries are less likely to earn over $50K annually, while around 25% of sales professionals do.

Wives and husbands are equally likely to earn over $50K annually, while only a few unmarried individuals earn more than $50K annually, indicating a significant disparity in income levels.

In Races, except Whites there are very few people of different races, which may lead to a lack of understanding of the percentage and relationship of individuals earning over $50K annually. Therefore, all races except whites were combined to become “others” category.

The earnings gap between males and females is significant, with less than 10% of adults earning over $50K annually, while nearly 33% of males earn over $50K annually.

The correlation map was plotted for numerical features to determine the relationship between independent and dependent features, revealing that most features are positively correlated with the Income Variable:

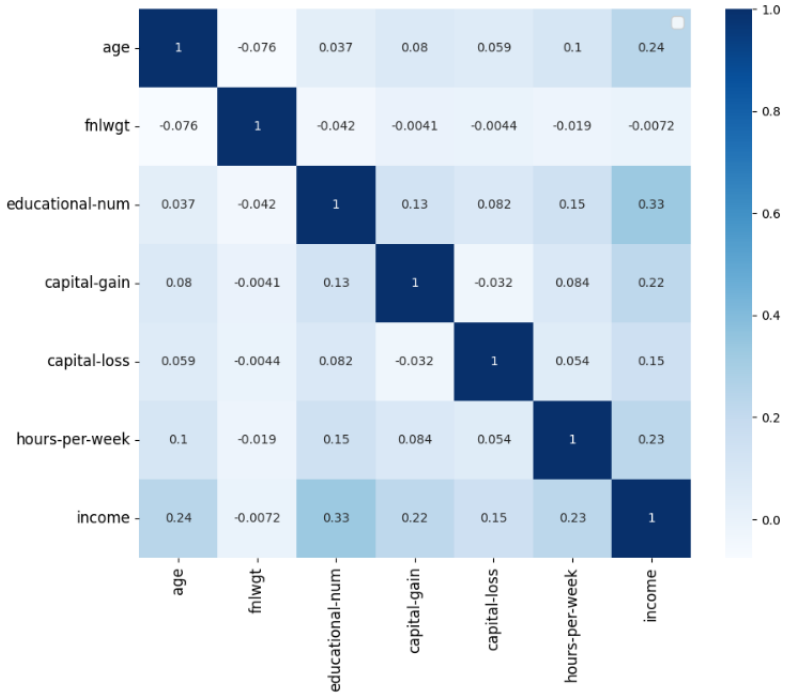


Figure 6: Correlation between numerical features of the dataset

To encode categorical variables, in the 'Gender' category, Male and Female were replaced by 1 and 0, respectively, while in "Education" category Label Encoding was used, and for the remaining categorical variables, One-Hot Encoding was applied.

Then, Normalization was used, to scale the values of different features to a standard range, and ensuring that all features contribute equally to the model training process. Two methods for normalization were used, Min-Max Scaling was applied for k-Nearest Neighbors model, and Standardization (Z-score normalization) was applied for Random Forest and XGBoost algorithms.

Finally, the data was divided randomly into training data and testing data, at a rate of 70% and 30%, respectively:

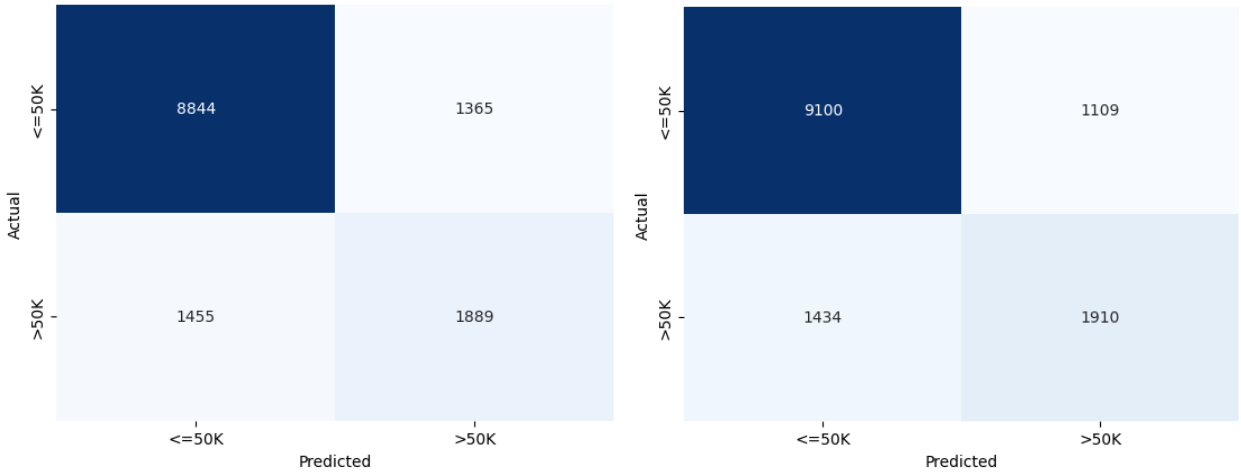
```
X_train shape: (31622, 84)
X_test shape: (13553, 84)
y_train shape: (31622,)
y_test shape: (13553,)
```

Figure 7: Training and testing data shapes

II. Experiments and Results

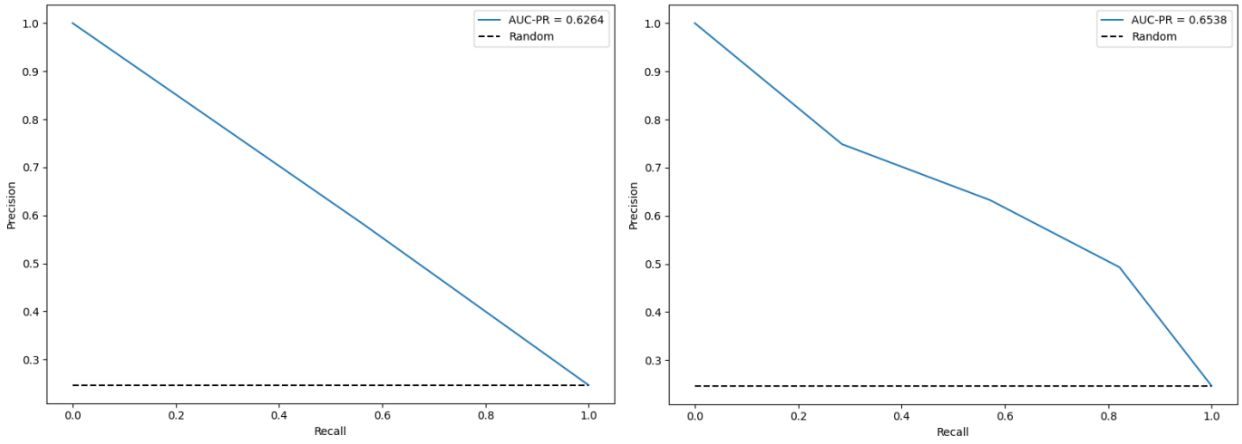
II.1. K nearest neighbors' model

The K-Nearest Neighbors (KNN) algorithm, an instance-based learning method, was utilized to predict features based on the majority class of its k-nearest neighbors. The Minkowski distance metric was employed for K=1 and K=3, and the results shown below:



a) Confusion Matrix for K = 1 *b) Confusion Matrix for K = 3*

Figure 8: Confusion Matrices for KNN



a) Precision-Recall Curve for K = 1 *b) Precision-Recall Curve for K = 3*

Figure 9: Precision-Recall Curve for KNN

Table 1: Evaluation metrics for KNN

	Precision	Recall	F1-score	Accuracy	AUC-PR
KNN, K = 1	0.58	0.56	0.57	0.79	0.62
KNN, K = 3	0.63	0.57	0.6	0.81	0.65

From the evaluation metrics above, accuracy was relatively high, but accuracy alone may not be suitable for evaluation, especially in this case with an imbalanced dataset. A more comprehensive evaluation considers precision, recall, F1-Score, and AUC-PR, providing a more subtle understanding of the model's strengths and weaknesses. Precision and recall are slightly higher for k = 3, indicating that the model with k = 3 is better at identifying instances with '>50K' income. The F1-Score and AUC-PR provide a subtle evaluation of the trade-off between precision and recall, and both are slightly higher for k = 3. The performance of the KNN models, particularly with k = 1 and k = 3, is considered weak. Despite achieving a reasonable balance between precision and recall, the overall predictive capability of the models is not robust. This weakness is evident in the modest F1-Score and AUC-PR values.

II.2. Random Forest Classifier

Random Forests are composed of several Decision Trees that were trained on random selections of features and data. The Random Forests model was chosen because it improves generalization by decreasing overfitting and increases robustness against imbalanced datasets by combining many trees. Two parameters were tuned to improves the model's performance:

- **n_estimators** parameter which represents the number of trees in the forest was tuned at values (10, 50, 100, 150, 200, 250, 300):

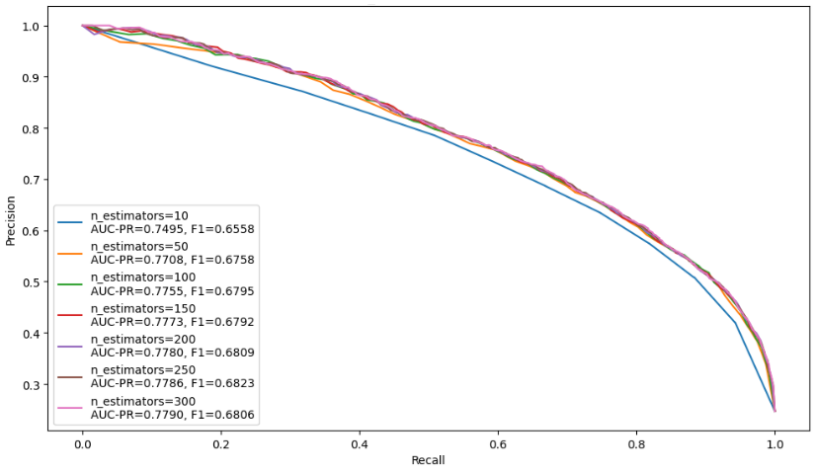


Figure 10: n_estimators Tuning in Random Forest Classifier

- **max_depth** parameter which controls the maximum depth of each tree was tuned at values (None, 5, 10, 15, 20, 25, 30):

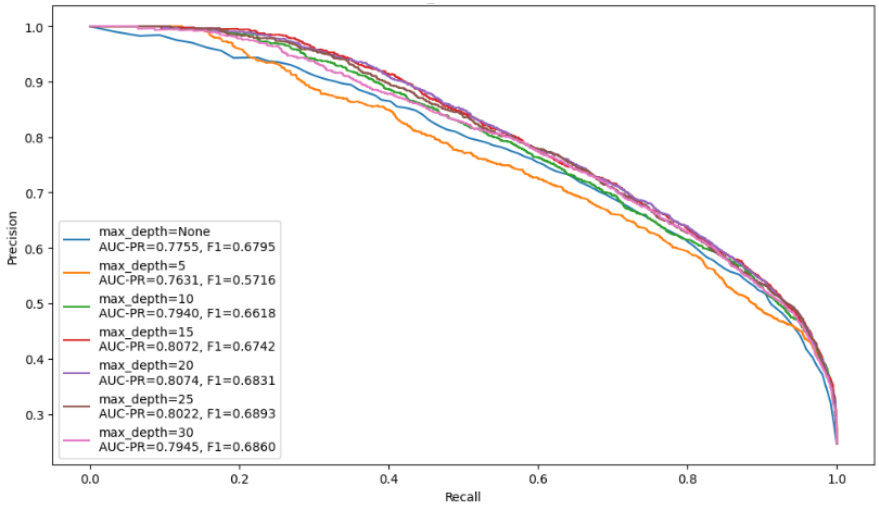
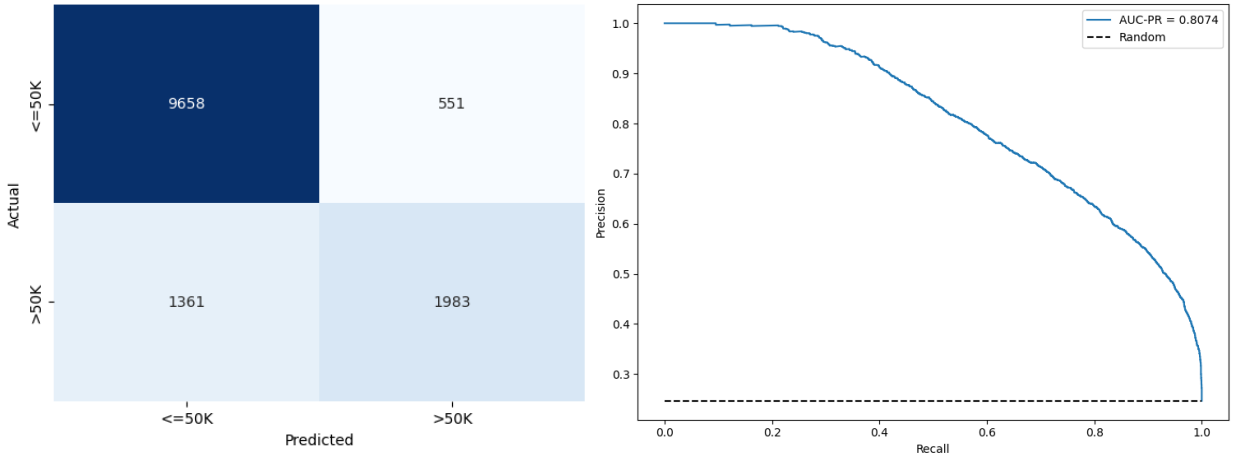


Figure 11: max_depth Tuning in Random Forest Classifier

The resultant curves above show that the values of AUC-PR and F1 begin to stabilize or vary insignificantly at 150 trees and max depth 15; hence, 150 trees and 15 for max depth were chosen for the model to get better performance and limit computational complexity, and the following shows the results of the chosen model:



a) Confusion Matrix for Random Forest

b) Precision-Recall Curve for Random Forest

Figure 12: Confusion Matrix and Precision-Recall Curve for Random Forest

Table 2: Evaluation metrics for Random Forest

	Precision	Recall	F1-score	Accuracy	AUC-PR
Random Forest	0.78	0.59	0.67	0.85	0.8

The Random Forest model's performance is mixed across evaluation metrics. Its precision is 0.78, indicating a significant portion of positive instances were true, but recall is 0.59, suggesting it may have missed some. The F1-score is 0.67, balancing precision and recall. The AUC-PR value of 0.8 suggests a reasonable trade-off between precision and recall. Improvement in recall is needed for a more comprehensive positive instance capture.

II.3. XGBoost Classifier

eXtreme Gradient Boosting uses the boosting ensemble approach to create a strong predictive model by integrating the predictions of numerous weak models. It develops trees in a sequential manner, with each new tree focused on the faults made by the preceding ones. XGBoosted was chosen because of its capacity to focus on misclassified occurrences via sequential learning under imbalanced conditions. Two parameters were tuned to improve the model's performance:

- **learning_rate** parameter which controls the contribution of each tree to the final prediction was tuned at values (0.01, 0.1, 0.2, 0.3, 0.5).

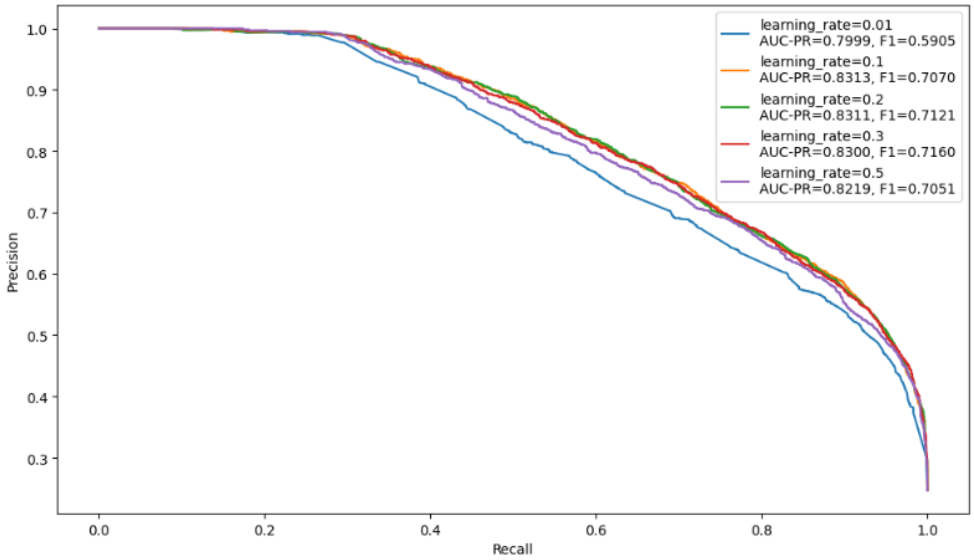


Figure 13: learning_rate Tuning in XGboost Classifier

- **max_depth** parameter which controls the maximum depth of each tree was tuned at values (3, 5, 7, 9, 11):

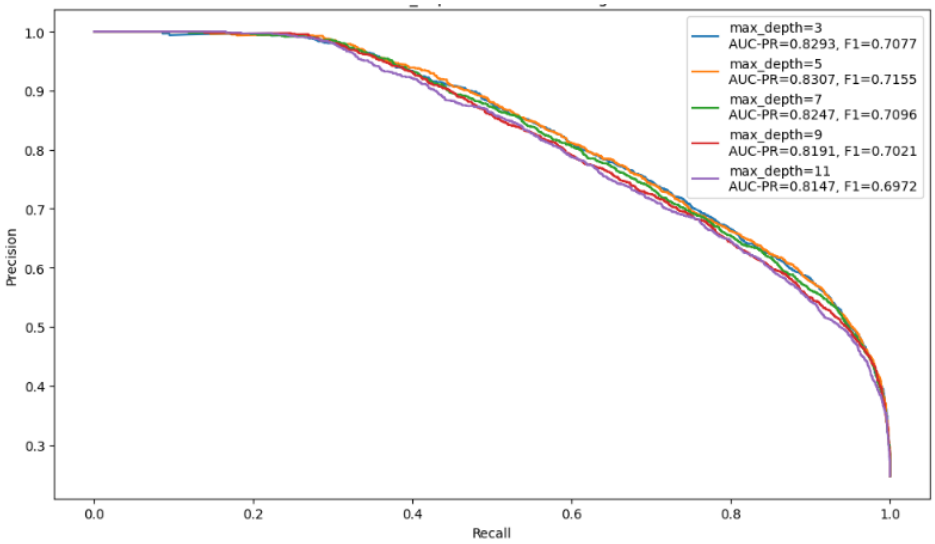
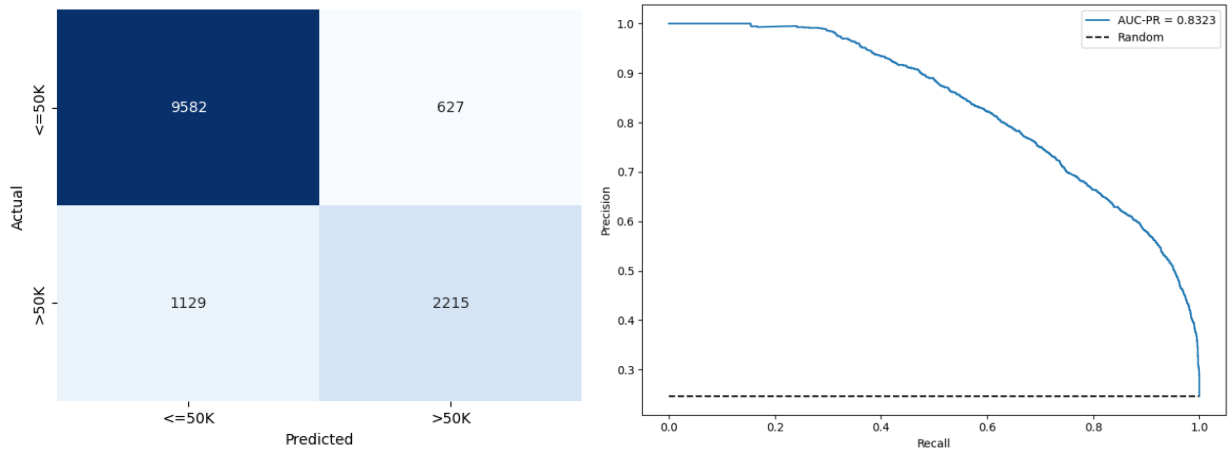


Figure 14: max_depth Tuning in XGboot Classifier

The resultant curves above demonstrate that the values of AUC-PR and F1 were the best on a learning rate of 0.2 and max depth of 5, which were chosen for the model to improve performance and reduce computational complexity, and the results of the chosen model are shown below:



a) Confusion Matrix for XGboot

b) Precision-Recall Curve for XGboot

Figure 15: Precision-Recall Curve for XGBoost

Table 3: Evaluation metrics for XGBoost

	Precision	Recall	F1-score	Accuracy	AUC-PR
XGBoost	0.78	0.66	0.71	0.87	0.83

The XGBoost classifier performs well across a variety of evaluation measures. It has a precision of 0.78, which means that 78% of the cases are true positives (>50K). It captures 66% of all (>50,000) instances, which, while not extremely high, indicates a reasonable sensitivity to positive cases.The F1-score of 0.71 balances Precision and Recall, indicating a positive trade-off. The high accuracy of 0.87 indicates general correctness, however accuracy may be affected by the majority class (<=50K). The AUC-PR of 0.83 assesses the model's ability to discriminate positive and negative events at various probability levels. Overall, the XGBoost model performs well across several criteria.

III. Analysis

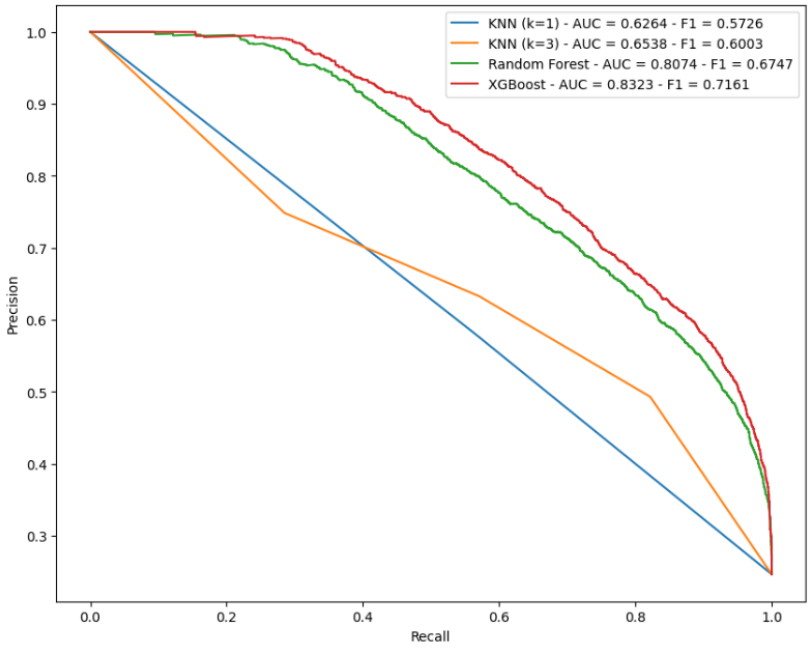


Figure 16: Precision-Recall Curves for all Models used

First, based on the figure shown above, the models were used as previously explained in the previous part were the K nearest neighbor and Random Forest and XGboost Classification models, and the PR-curve was plotted to show that the best model was the XGboost model, because of the area under the curve was the best value of it and it was 0.8323 and the f1 score was 0.7161, and this mean that the model has a good ability to distinguish between the two classes (>50) and (<=50). Especially, when the area under the curve close to 1 as it shown in our model, so it gives an indication that has excellent discrimination and the model was very effective at recognizing between the positive and the negative classes. In other hand, the value of f1 score was 0.7161 and it proves that the XGboost model was the best also, and in our case the value got of f1 score indicates that the model with balanced precision and recall, so it means that the value good at both correctly identifying positive instances and avoiding false positives.

Moreover, cases of misclassification were analyzed, as it was noted that the instances that were incorrectly classified and given an income classification (> 50), most of them had an educational degree of Bachelor's degree and an occupation label of Prof-specialty. It was also noted that the instances that were incorrectly classified and given an income classification (<= 50), most of them had an educational degree of high school degree and an occupation label of Craft-repair.

IV. Conclusion

In conclusion, the analysis of the Adult Income dataset gave useful data on predicting individual yearly incomes over $50k based on employment and demographic variables, so the complete preprocessing, which included solving missing values and removing duplicates, cleared the path for an in-depth understanding of the nature of the dataset. Also, the connections between different features are made clear by visualizations and descriptive statistical information, which also reveal patterns like the impact of age, occupation, and education on income levels.

Based on the requested tasks in the assignment, there are three classification models were used in the experimental part like: XGboost, Random Forest, and K-Nearest Neighbors (KNN). Although KNN models showed some accuracy, their poor predicting ability was revealed by their weaknesses in precision, recall, F1-Score, and AUC-PR. While the Random Forest Classifier gained good precision, its decreased recall raised concerns, so its performance was variable. On the other hand, the XGboost Classifier continuously beat the other models, showing a strong balance between recall and precision. The models had additional improvements by hyper parameter tuning, with XGboost emerging as the best performance with high recall, F1-Score, precision, and AUC-PR.

The analysis revealed areas of misclassification, particularly in regards to occupation and education, and also demonstrated how well the XGboost model predicts income levels, so this analysis highlights how important it is to do thorough preprocessing, carefully select models, and fine tuning parameters when creating reliable machine learning models for binary classification tasks.

This project faced challenges associated with imbalanced data, where the majority of individuals earned less than $50k annually. Handling imbalanced datasets required careful consideration of evaluation metrics, with precision, recall, F1-Score, and AUC-PR proving essential for nuanced performance assessment. The choice of choosing models played a pivotal role in overcoming imbalanced data challenges. The models demonstrated superior performance by effectively balancing precision and recall, highlighting the importance of selecting models and metrics capable of navigating the complexities posed by imbalanced data, ultimately ensuring accurate and reliable predictions in real-world scenarios.

V. References

[1] UCI Machine Learning Repository. (n.d.). Retrieved January 22, 2024, from https://archive.ics.uci.edu/dataset/2/adult