Alexis Lewis, Fola Ilori, Rafael Quiroz Portella
Professor Tasfia Mashiat
CS:3980:0002
13 September 2025

Final Report

Abstract

This paper investigates bias in machine learning models in the areas of criminal justice, education, and income prediction. Our goal is to measure and reflect on how these systems produce different and unequal outcomes across demographic groups. To do this we apply fairness evaluation techniques to the COMPAS dataset (recidivism prediction), the LSAC dataset (bar passage prediction), and the UCI Adult Income dataset (income prediction).

The three models that were trained on each dataset are Logistic Regression (Logreg), K-Nearest Neighbors (KNN), and K-Means Clustering (KMC). These models were evaluated using four fairness metrics: Demographic Parity, Equal Opportunity, Equalized Odds, and Disparate Impact, which help explain how models vary across sensitive attributes like race, gender, and age.

Key findings include evidence of higher false positive rates for African Americans in the COMPAS dataset, unfair predictions against lower income and minority students in the LSAC data, and major gender and racial disparities in the Adult Income predictions. Reweighting mitigation strategies helped reduce some of these disparities in the LSAC and COMPAS datasets but had small tradeoffs in model performance.

This work focuses the biases present in real world datasets and how the use of fairness evaluation in machine learning is crucial. Without this these types of algorithmic systems may perpetuate or worsen existing inequalities in society.

1. Introduction

As machine learning becomes more prevalent in decision making systems across industries, concerns about fairness and bias have also become more frequent. In important areas such as hiring, lending, education, and criminal justice, unfair predictions can have serious and permanent consequences. A model that systematically disadvantages a particular demographic group can increase existing inequalities and result in discriminatory outcomes.

This paper examines whether standard machine learning classifiers produce biased outcomes when applied to three widely studied datasets, each reflecting an area in the real world with ethical implications. In all three cases, we treat each classifier as a black box and evaluate its outputs to assess disparities in sensitive demographic attributes such as race, gender, income, and age.

The datasets we analyze are:

- The COMPAS dataset, used to assess recidivism risk in the U.S. criminal justice system, has previously been judged for racial bias in sentencing decisions.
- The LSAC dataset contains data on law school applicants and outcomes, including whether an individual passes the bar exam.
- The Adult Income dataset from the UCI Machine Learning Repository is a benchmark dataset used to predict whether an individual earns more than $50K per year based on demographic and occupational features.

Each of these datasets has the potential to reflect and reinforce inequities. Therefore, we apply fairness metrics and mitigation strategies to explore how bias comes to be and what can be done to address it.

2. Methods

2.1 Dataset Characteristics

COMPAS Dataset

The COMPAS dataset was sourced from Kaggle and contains 60,843 rows and 28 columns. It includes data on Individuals evaluated by the COMPAS system, with a focus on predicting the likelihood of reoffending. The primary outcome variable used in our analysis is a binary indicator of whether the individual was classified as "high risk," based on recoded supervision levels.

The key attributes in this dataset were gender, race, age. The target variable was a binary outcome of 0 for low/medium risk, 1 for high risk. Preprocessing involved parsing birth dates, filtering age (18-90), removing rows missing key variables or invalid data, and grouping ages. RawScore and DecileScore were retained.

Visual analysis revealed significant imbalances: 78% of individuals were male, and 44% were African American. Box plots showed that African Americans received higher decile scores on average, pointing to potential bias in the underlying COMPAS risk scoring system.

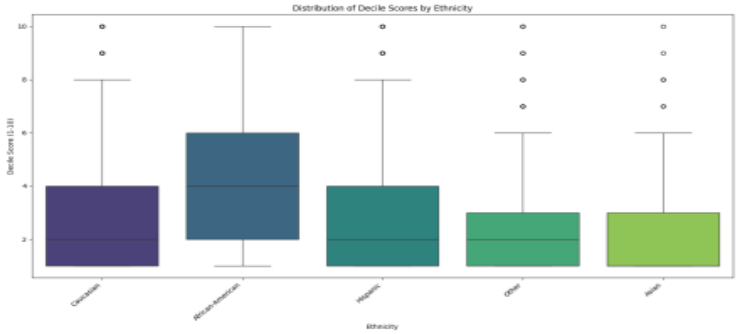


Figure 1: Box plot of decile scores by ethnicity: African American individuals had higher median risk scores, indicating disparities

LSAC Dataset

The LSAC dataset, originally collected by Linda F. Wightman (1998), contains 22,407 rows and 39 columns on law school applicants, including demographic information, LSAT scores, undergraduate GPA, and bar passage outcomes.

Sensitive attributes include gender (male; 0 = female, 1 = male), race, and income. The binary target is bar passage (pass_bar; 0 = fail, 1 = pass). After imputing missing values using the most common category and dropping irrelevant columns (graduation year, dropout status, identifiers, and clustering labels) one-hot encoding was used and categorical values were reindexed.

Statistics showed high pass rates overall (94.8%), with racial disparities: 96.6% of White students passed, compared to only 77.8% of Black students. Disparities were also seen in bar passage by family income, law school tier and gender, supported by figures A7, A8, and A9 in the appenndix.

Adult Income Dataset

From the UCI Machine Learning Repository, this dataset has 32,561 training rows, 16,281 testing rows, and 15 columns. It predicts whether income exceeds $50K.

Sensitive features were sex and race. After replacing missing values with mode, we dropped skewed columns, one-hot encoded categorical variables, and scaled numeric features.

Analysis revealed that men were predicted to earn >$50K over 3.5x more than women. There were also racial disparities with Whites and Asians being much more likely to be predicted in the >$50K group compared to other races, as seen in figure A3 and A4 in the appendix.

2.2 Model Training

To answer our research question and assess fairness, we intentionally selected three vastly different algorithms to broaden the scope of our results. Logistic Regression (Logreg) is a parametric linear model with high interpretability, K-Nearest Neighbors (KNN) is a non-parametric instance model, and K-Means Clustering is an unsupervised learning model.

We used the following libraries: Pandas for dataset and DataFrame handling and manipulation, mathplotlib, seaborn, and IPython for visualizations and plotting, and statistical analysis was performed using numpy, scipy, and statsmodels. For the machine learning algorithms, preprocessing, and metrics we used scikit-learn. Imbalanced-learn (Imblearn) was used to address class imbalance. Finally, fairness was assessed using fairlearn, and holisticai.

The dataset split strategy varied among datasets. COMPAS used a 75/25 train-test split and stratified by HighRisk and Race to preserve demographic balance. Adult Income used a 75/25 train-test split while LSAC used a 70/30 split.

COMPAS pipeline included scaling numerical features Age, RawScore, and DecileScore. Categorial attributes Race, Gender, and Marital Status were one-hot encoded. The LSAC pipeline involved scaling numeric features (upga, lsat) using MinMaxScaler; one-hot encoding the categorical attribute race1; and re-encoding family income from 1-6 to 0-5. Adult Income pipeline used ColumnTransformer to scale numeric and encode categorical features.

Model performance evaluation remained consistent across datasets and included drawing confusion matrices and calculating accuracy, precision, recall, F1, and ROC AUC for each model. We then compared the performance between Logreg, KNN, and KMC.

LSAC dataset exhibited high class imbalance between students passing or failing the bar exam. To address this, we trained a second Logreg model using the Synthetic Minority Over Sampling Technique (SMOTE) to oversample minority class (failures). This second Logreg model followed the previously described procedure.

2.3 Fairness Assessment

To conduct the fairness assessment the following four metrics were used:

First, Demographic Parity (DP) measures the equal positive prediction rate across groups. DP was chosen because it helps detect whether the model favors one group over another.

Second, Equal Opportunity (EO) measures the equality of true positive rates across groups. It was selected because it evaluates if individuals from different groups have the same chance of being correctly classified by the model.

Third, Equalized Odds Difference (EOD), measures the differences in both the true positive rate (TPR) and false positive rate (FPR) across groups. Was included in the fairness analysis as it extends EO to also account for disparities in the FPR.

Finally, Disparate Impact (DI), measures the ratio of positive prediction rates against the reference group. It was added because DI has a clearly defined legal threshold values below 0.8 or above 1.25 are considered unfair.

Fairlearn and HolisticAI (holisticai) libraries were used to conduct these fairness metrics.

2.4 Bias Mitigation

Bias mitigation was performed on the LSAC and COMPAS datasets by applying pre-processing techniques that targeted racial bias, using the Logreg model and AI Fairness 360 (afi360) library, focusing exclusively on the race attribute. Four models were then compared: First, the baseline Logreg model, trained on the unmodified dataset. Second, Reweight only, which applies the Reweighing bias mitigation technique without modifying the data distribution. Third, SMOTE only, which applies SMOTE to balance class labels before training, without any fairness reweighting. Finally, SMOTE plus Reweight combines both techniques by applying SMOTE to rebalance classes, followed by Reweighting to mitigate bias. Reweighting improved fairness across all metrics, while SMOTE alone introduced new disparities, emphasizing the importance of selecting mitigation strategies aligned with fairness goals.

In the COMPAS dataset, only two models were compared: a baseline Logreg model and a Reweight-only model that adjusted sample weights based on racial group membership. Reweighting effectively reduced the false positive rate for African American individuals and improved the model's Disparate impact score, indicating a more equitable outcome. However, similar to LSAC, these fairness gains came with a slight trade-off in performance, reflected in a reduced F1 score.

3. Results

3.1 Model Performance

Table 1 presents a summary of model performance metrics for models trained on the COMPAS dataset. The Logreg model achieved an accuracy of 0.88, a precision of 0.75, a recall of 0.61, an F1 Score of 0.67, and an AUC of 0.91. Appendix B Figure B1 shows the confusion matrix and Figure B2 displays the ROC curve for the Logreg model. The KNN model displayed an accuracy of 0.90, precision of 0.80, recall of 0.67, F1 Score of 0.73, and an AUC of 0.91. Appendix B Figure B3 shows the confusion matrix and Figure B4 displays the ROC curve for KNN. In contrast, the KMC model had an accuracy of 0.60, with precision at 0.20, recall at 0.33, and an F1 Score of 0.25. Appendix B Figure B5 shows the KMC confusion matrix.

Table 1 – Model Performance Metrics – COMPAS dataset

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logreg	0.88	0.75	0.61	0.67	0.91
KNN	0.90	0.80	0.67	0.73	0.91
KMC	0.60	0.20	0.33	0.25	

Table 2 illustrates the model performance metrics for those trained on the LSAC dataset. The baseline Logreg model achieved an accuracy of 0.95, with precision also at 0.95, a perfect recall of 1.00, an F1 Score of 0.97, and an AUC of 0.78. The KNN model similarly had an accuracy of 0.90, precision of 0.80, recall of 0.67, F1 Score of 0.73, and an AUC of 0.91. Meanwhile, the KMC model recorded an accuracy of 0.60, a precision of 0.20, a recall of 0.33, an F1 Score of 0.25, along

with an AUC of 0.61. Appendix B Figure B6 shows the confusion matrices across models, Figure B7 displays the ROC curves across models Figure B8 depicts model performance metrics across models (including SMOTE Logreg results).

Table 2 – Model Performance Metrics – LSAC dataset					
Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Baseline Logreg	0.95	0.95	1.00	0.97	0.78
Logreg + SMOTE	0.74	0.98	0.75	0.85	0.78
KNN	0.95	0.95	1.000	0.97	0.72
KMC	0.42	0.96	0.41	0.57	0.61

Table 3 outlines the performance metrics for models trained on the Adult Income dataset. The Logreg model achieved an accuracy of 0.83, with a precision of 0.70, a recall of 0.56, an F1 Score of 0.62, and an AUC of 0.89. Appendix B Figure B9 shows the confusion matrix and Figure B10 displays the ROC curve for this model. The KNN model had an accuracy of 0.82, a precision of 0.63, a recall of 0.60, an F1 Score of 0.61, and an AUC of 0.85. Appendix B Figure B11 shows the confusion matrix and Figure B12 displays the ROC curve for KNN. Lastly, the KMC model recorded an accuracy of 0.68, with a precision of 0.42, recall at 0.89, and an F1 Score of 0.60.

Table 3 – Model Performance Metrics – Adult Income dataset					
Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logreg	0.83	0.70	0.56	0.62	0.89
KNN	0.82	0.63	0.60	0.61	0.85
KMC	0.68	0.42	0.89	0.60	

3.2 Fairness Metrics

To evaluate fairness in model predictions, we applied four well-established fairness metrics across all datasets: Demographic Parity (DP), Equal Opportunity (EO), Equalized Odds (EOD), and Disparate Impact (DI). Demographic Parity refers to the requirement that the probability of a favorable outcome (e.g., being predicted to have a high income) is equal across different demographic groups, regardless of actual outcomes. Equal Opportunity requires that true positive rates be similar across groups. Disparate Impact quantifies the ratio of favorable outcomes between groups, with a range of 0.8 to 1.25 generally considered acceptable.

In the COMPAS dataset, disparities were especially evident across race, gender, and age groups. African Americans had a false positive rate (FPR) of 8.2% compared to 1.87% for Caucasians (see Figure 2). Their DI score was 1.67- well outside of the fairness threshold- indicating they were disproportionately labeled high risk. Males were 1.6 times more likely to receive a high-risk label than females, and the age group 26-35 showed the highest high-risk labeling rate, with a DI of 1.88 (see Appendix A, Figure A1). Visualizations such as bar plots of FPR and DI across these demographic subgroups further supported these findings.

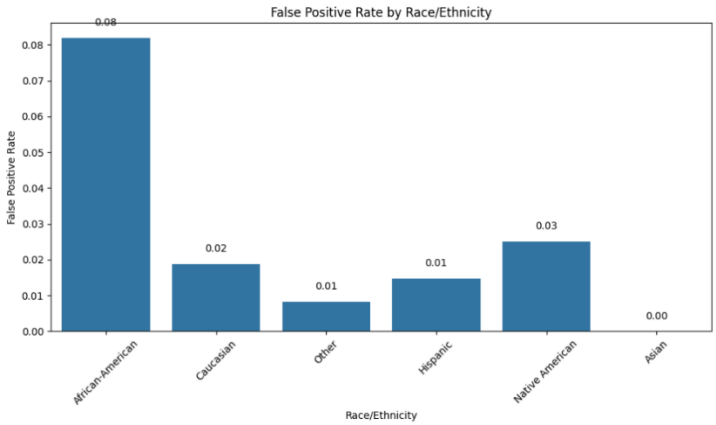


Figure 2: False Positive Rate by race in the COMPAS dataset, showing significant disparities between African American and Caucasian individuals.

In the LSAC dataset, gender disparities were minor. Difference in true and false positive rates between males and females under 1% and the fairness metrics aligned with this observation (DI=1.004). However, more significant disparities were found by race and family income level. African American students exhibited a false positive rate of 19.7% compared to 3.5% for White students. Model performance varied across racial groups, with higher F1 scores observed for White (0.983) and Asian (0.951) students, and the lowest for African American students (0.875), as shown in Appendix A, Figure A2. DI remained above the legal fairness threshold (0.8) across all racial groups but was lowest for African American (0.94) and Hispanic (0.98) individuals.

For the Adult Income dataset. Disparities were substantial and multifaceted. Males were predicted to have high income more frequently than females, resulting in a DP of 0.184 and EO difference of 0.131. The DI for females was 0.43- indicating significant disadvantage (see Appendix A, Figure A3). Racial disparities were even more pronounced, with a DP difference of 0.244, EOD of 0.494, and DI scores ranging from 0.01 (American Indian-Eskimo) to 2.06 (Asian-Pacific Islander) (see Appendix A, Figure A4). These results suggest the model's outcomes were highly sensitive to demographic attributes and that fairness issues warrant serious attention.

3.3. Bias Mitigation

Bias mitigation strategies were applied to the COMPAS and LSAC datasets, while the Adult Income dataset was only evaluated.

In the LSAC dataset, we implemented three approaches- reweighting, SMOTE, and a hybrid of both- on a Logreg model to evaluate their impact on racial fairness. The baseline model had moderate fairness, with a DP of 0.056, EO of 0.032, EOD of 0.138, and DI of 0.944. Reweighting alone slightly improved fairness (DP decreased to 0.044, EO to 0.025, EOD to 0.106, and DI increased to 0.956), suggesting that strategic reweighting can balance subgroup influence without altering the dataset itself. SMOTE, in contrast, yielded sharp increases in DP (0.392), EO (0.675), and EOD (0.675), but the DI plummeted to 0.147. This distortion suggests that while SMOTE addresses class imbalance, it exacerbates subgroup prediction disparities. Combining SMOTE with reweighting produced somewhat moderated but still problematic results: DP of 0.465, EO of 0.418, EOD of 0.418, and DI of 0.373.

Model	DP	EO	EOD	DI
Baseline Logistic Reg.	0.056	0.032	0.138	0.944
Reweight Only	0.044	0.025	0.106	0.956
SMOTE Only	0.692	0.675	0.675	0.147
SMOTE + Reweighting	0.465	0.418	0.418	0.373

Figure 3: Comparison of fairness metrics across baseline and mitigation strategies applied to the LSAC dataset.

For the COMPAS dataset, reweighting was used to address racial bias. Sample weights were generated to rebalance outcomes across racial subgroups, especially for African Americans. This resulted in a lower FPR for African Americans and improved overall fairness metrics (see Figure 4). However, there was a trade-off: the model's F1 score dropped from 0.73 to 0.667 post-mitigation. This illustrates a well-known challenge in fairness-aware machine learning- enhancing equity often comes at the expense of model precision or recall.

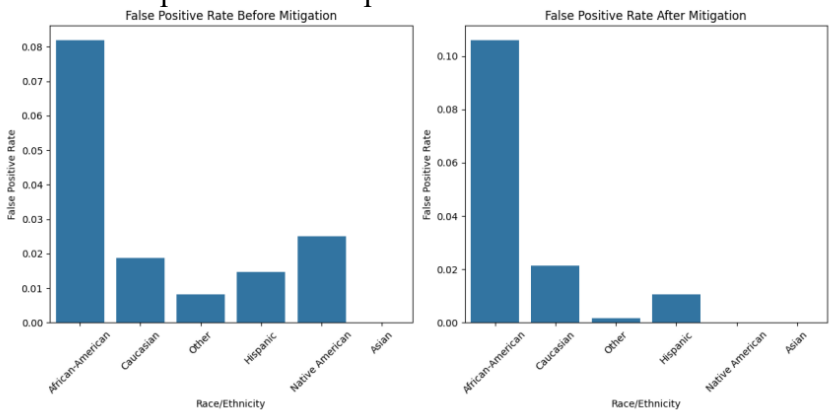


Figure 4: FPR by race before and after reweighting in the COMPAS dataset, demonstrating improved fairness for African American individuals.

The Adult Income dataset was not subjected to any bias mitigation efforts in this analysis. Despite evidence of demographic disparities, the dataset's mitigation remains a recommendation for future work. Techniques like reweighting, post-processing, or adopting fairness-aware learning algorithms could potentially address the stark inequalities observed.

4. Discussion

Fairness assessments across the three datasets reveal consistent disparities that disproportionately affect historically marginalized groups. These disparities reflect both systemic inequalities in the data and the impact of model design choices. Despite strong performance metrics overall, fairness metrics exposed group-specific shortcomings that warrant close examination.

In the COMPAS dataset, African American individuals and younger adults were disproportionately labeled as high-risk. This aligns with broader critiques of the COMPAS system and suggests the model inherited historical bias patterns. Applying reweighting helped reduce the false positive rate for African Americans, though at a slight cost to performance (F1 score dropped from 0.73 to 0.667). This reflects a common fairness-performance trade-off: efforts to equalize outcomes can shift model optimization away from overall accuracy.

In the LSAC dataset, while gender disparities were minimal, race and income-based gaps were evident. African American students had lower F1 scores and higher false positive rates. Students from lower-income quintiles were less likely to be correctly identified as passing. Reweighting proved to be an effective mitigation technique, improving fairness metrics with minimal disruption. SMOTE, while intended to address class imbalance, introduced additional disparities, emphasizing the importance of selecting mitigation strategies that align with dataset characteristics.

In the Adult Income dataset, disparities were most pronounced. Women were predicted to earn more than $50K significantly less often than men. Racial disparities were also extensive, with Disparate Impact scores ranging from 0.01 to over 2.0. Although fairness metrics clearly indicated unequal treatment, no mitigation strategy was implemented. Future work should prioritize this dataset for fairness intervention.

These results suggest that fairness-aware machine learning requires not just accurate models, but equitable ones. Ensuring equal treatment across groups is essential in domains like criminal justice, education, and employment, where biased predictions can lead to real-world harm.

5. Conclusion and Best Practices

This analysis demonstrates that fairness disparities exist across model predictions, even when performance metrics appear strong. All three datasets—COMPAS, LSAC, and Adult Income—showed inequities tied to race, gender, or income. While mitigation improved fairness in COMPAS and LSAC, no action was taken in Adult Income, highlighting an urgent need for future fairness mitigation efforts in this domain.

The findings affirm that machine learning models must be evaluated not just on accuracy, but on equity. Responsible modeling includes auditing fairness, selecting appropriate mitigation strategies, and being transparent about limitations.

Recommended Best Practices for Fairness in Machine Learning:

1. Use representative and balanced data. Avoid datasets that underrepresent or misrepresent certain groups.
2. Audit for fairness early and often. Use multiple metrics like DP, EO, EOD, and DI throughout the model development lifecycle.
3. Apply context-aware mitigation techniques. Choose strategies like reweighting or in-processing based on dataset balance and model use case.
4. Document trade-offs transparently. Clearly report any decrease in performance when improving fairness.
5. Maintain human oversight. Keep people in the loop for decisions made in high-impact areas.
6. Disclose all fairness evaluations. Use tools like model cards or datasheets to communicate fairness checks.

Fairness is not just a technical metric—it is an ethical responsibility. Integrating these practices into ML development is essential to ensuring models benefit all users equitably.

Works Cited

Kohavi, Ron, and Barry Becker. *UCI Adult Income Dataset*. UCI Machine Learning Repository, 1996. https://archive.ics.uci.edu/dataset/2/adult. Accessed 6 May 2025.

DanOfer. *COMPAS Recidivism Risk Score Data and Analysis*. Kaggle, 2021. https://www.kaggle.com/datasets/danofer/compass. Accessed 6 May 2025.

DanOfer. *Law School Admissions & Bar Passage Data*. Kaggle, 2021. https://www.kaggle.com/datasets/danofer/law-school-admissions-bar-passage. Accessed 6 May 2025.

Wightman, L. F. (1998). *LSAC National Longitudinal Bar Passage Study*. Law School Admission Council.

Appendix A: Supplementary Figures

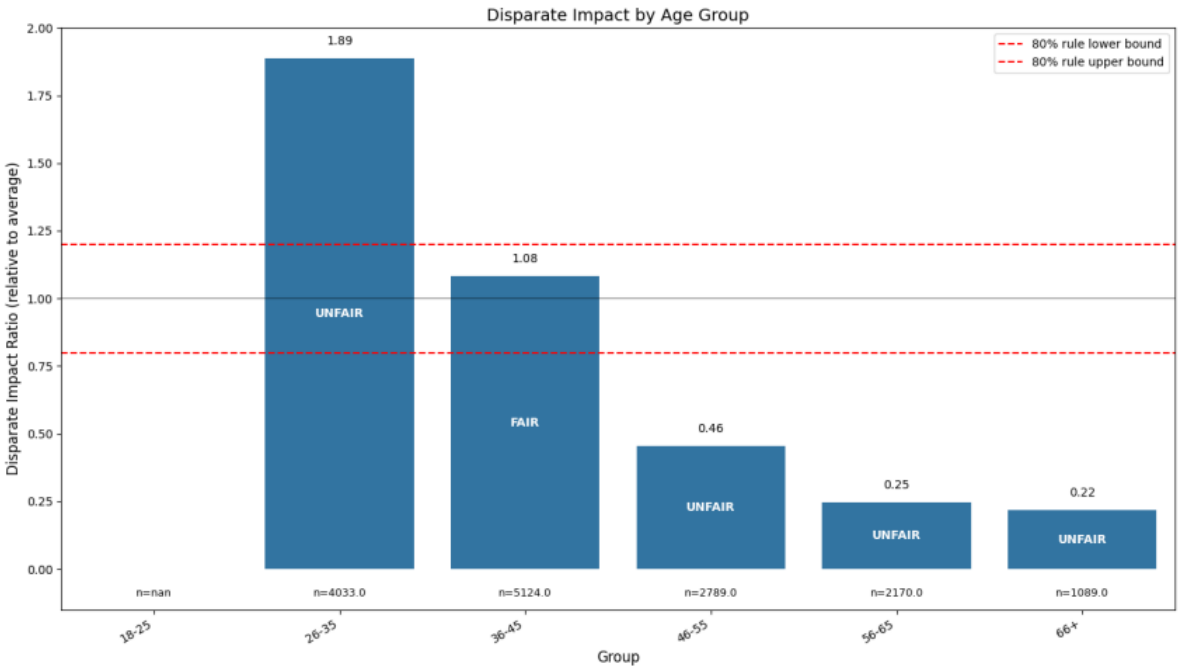


Figure A1: Disparate Impact by age group in the COMPAS dataset, with individuals aged 26–35 receiving the highest proportion of high-risk labels.

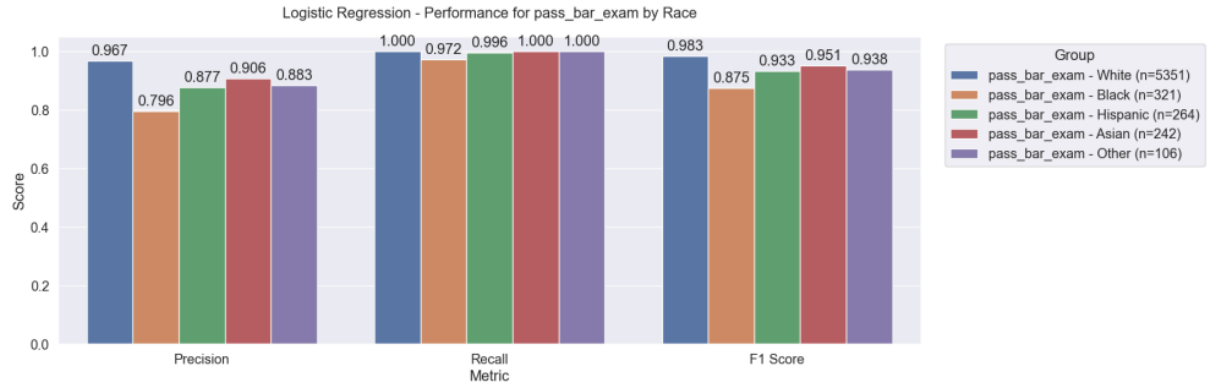
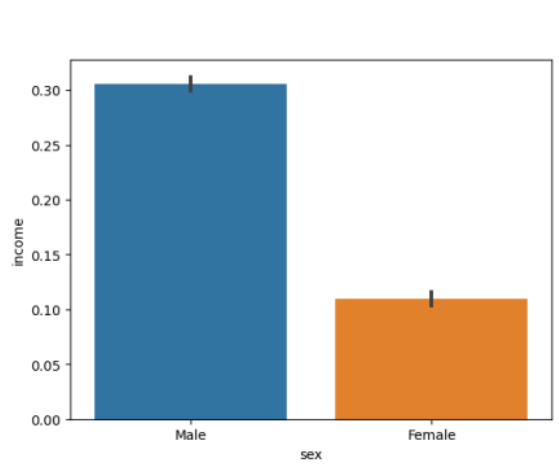


Figure A2: Precision, recall, and F1-score of the logistic regression model by race for bar exam predictions in the LSAC dataset. Performance is highest for White and Asian students and lowest for Black students.



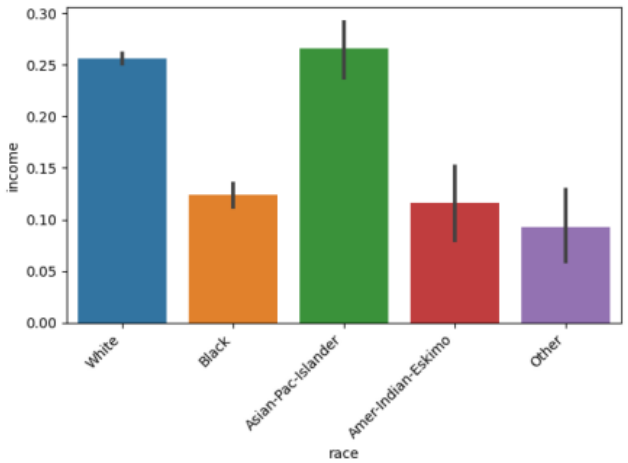


Figure A3 (left): Proportion of individuals predicted to earn more than $50K by sex in the Adult Income dataset, illustrating disparity in positive prediction rates.

Figure A4 (right): Disparate Impact across racial groups in the Adult Income dataset, highlighting significant variability in positive prediction outcomes.

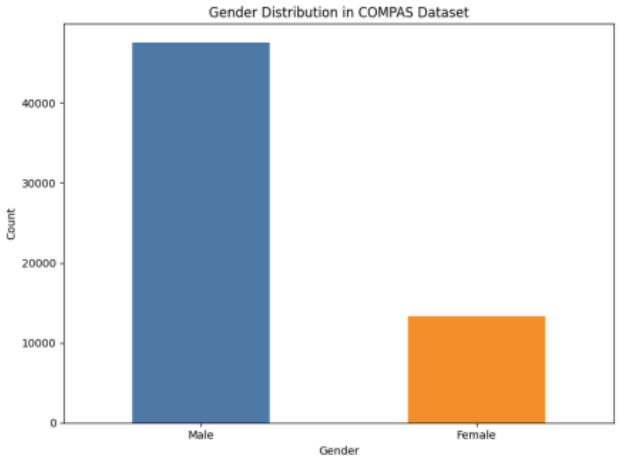


Figure A5: Gender distribution in the COMPAS dataset, the dataset is predominantly male (78%)

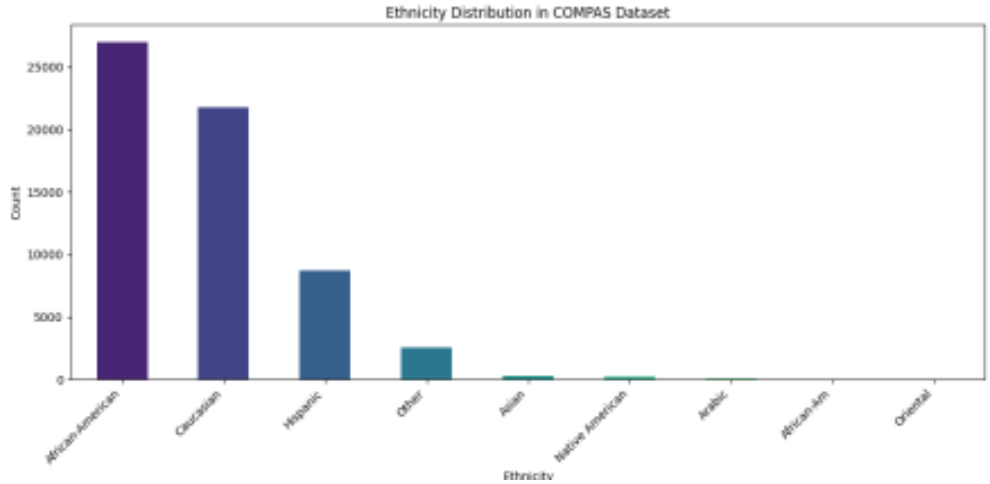


Figure A6: Ethnicity distribution in COMPAS dataset. The largest groups are African American and Caucasian, with notable underrepresentation of others.

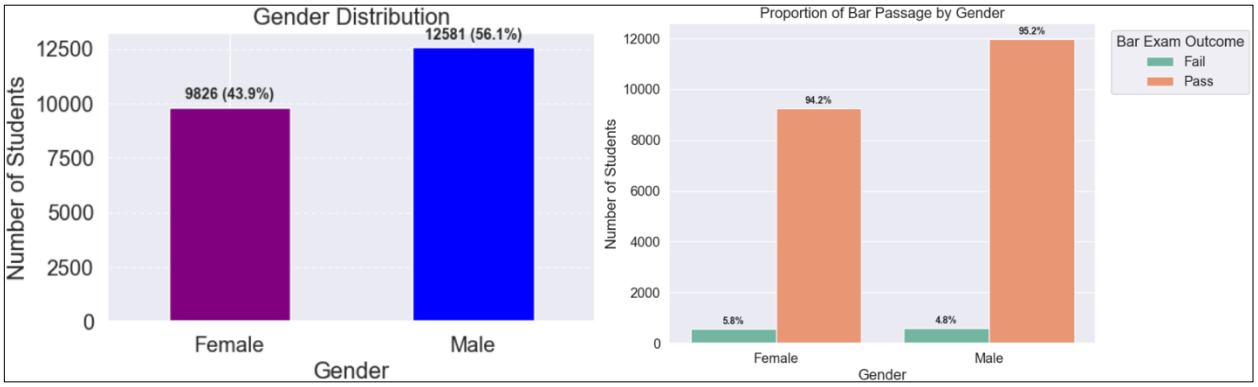


Figure A7: Gender Distribution and Bar Passage Rates in LSAC Dataset
Males slightly outnumber females (56%) and had marginally higher bar passage rates (95.2% vs. 94.2%).

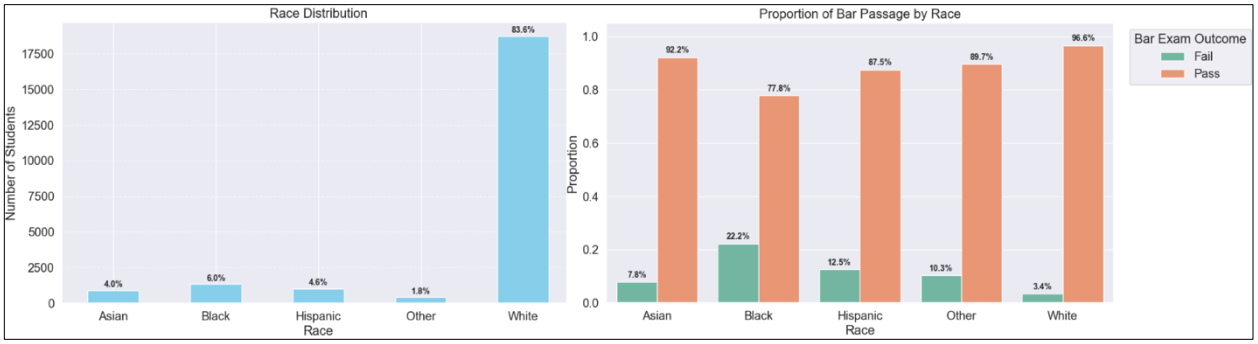


Figure A8: Race Distribution and Bar Passage by Race
White students dominate the dataset (83.6%), with significantly higher pass rates than Black (77.8%) and Hispanic students.

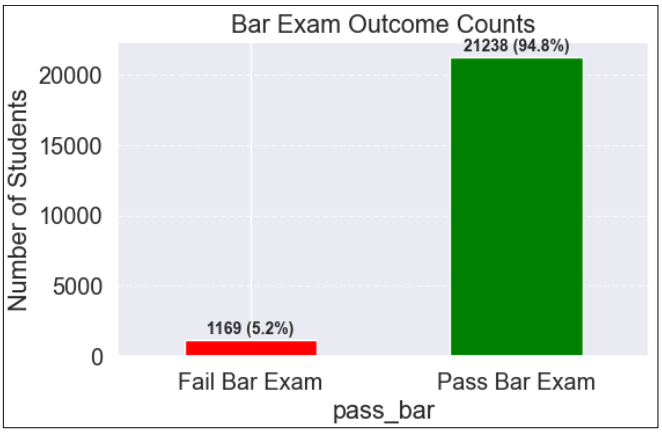


Figure A9: Overall Bar Exam Outcome Counts 94.8% of law students passed the bar, while 5.2% failed—highlighting class imbalance in the target variable.

Appendix B: Model performance visuals and confusion metrices

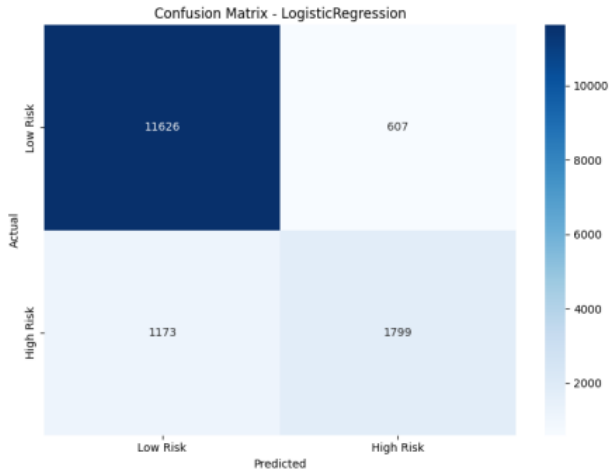


Figure B1: COMPAS Logreg confusion matrix.

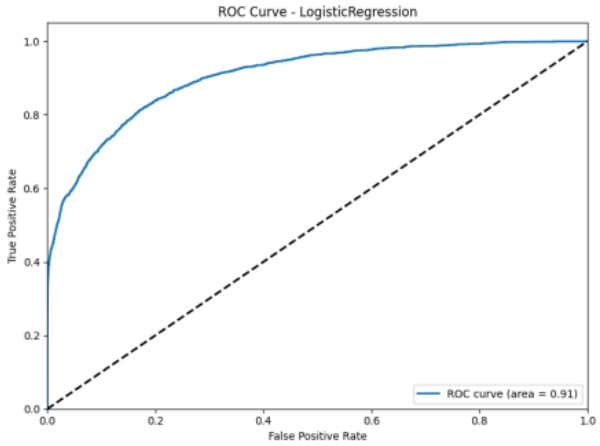


Figure B2: COMPAS Logreg ROC curve.

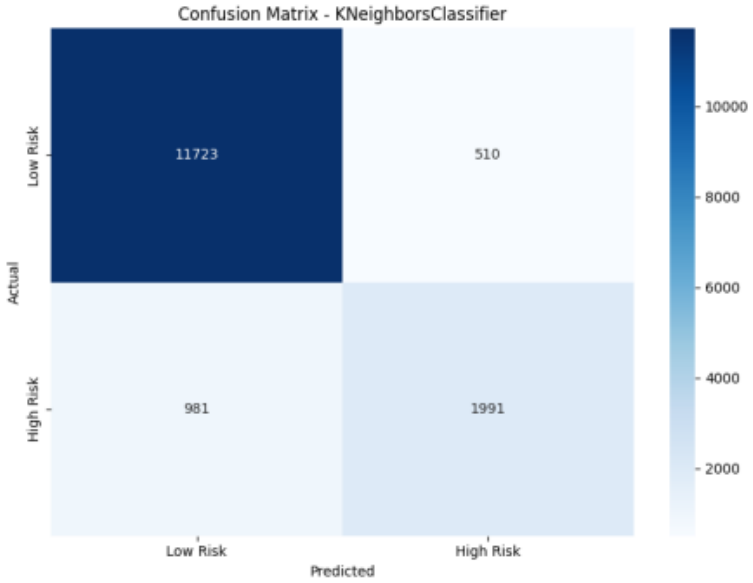


Figure B3: COMPAS KNN confusion matrix.

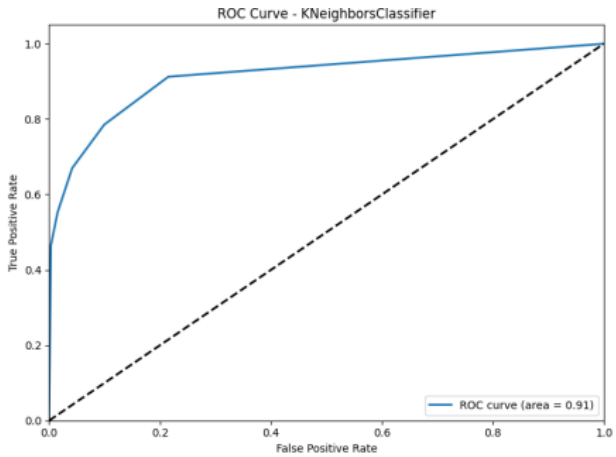


Figure B4: COMPAS KNN ROC curve.

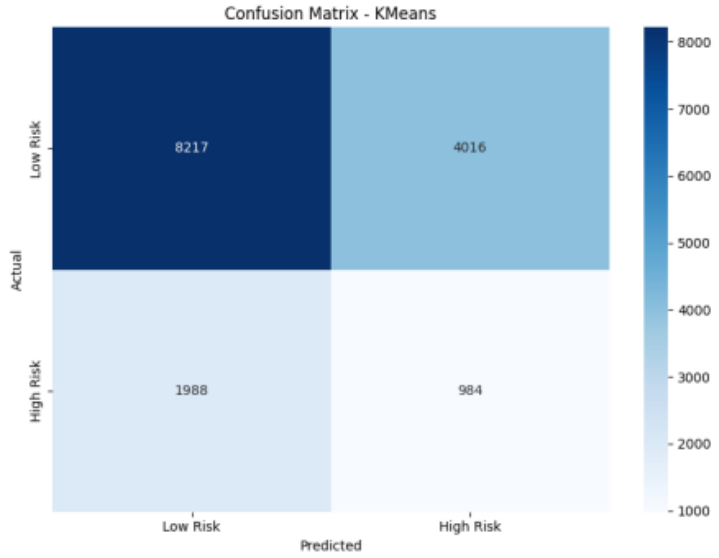


Figure B5: COMPAS KMC confusion matrix.

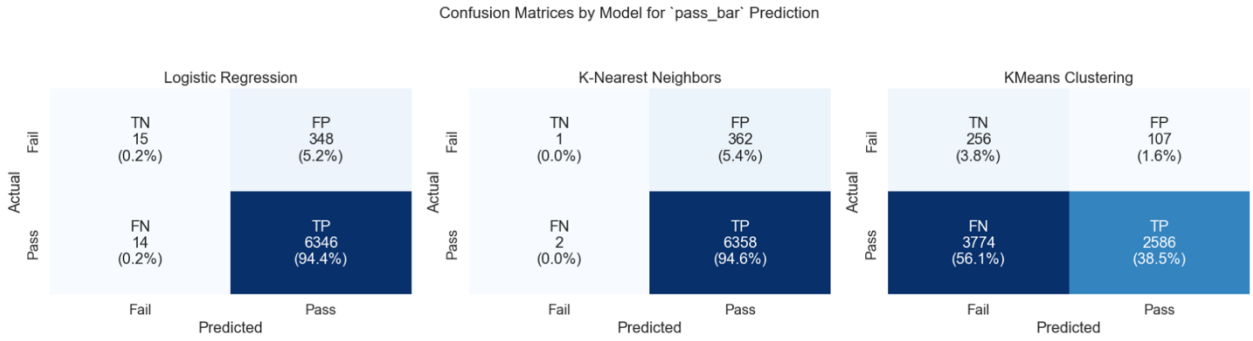


Figure B6: LSAC overall model performance comparison.

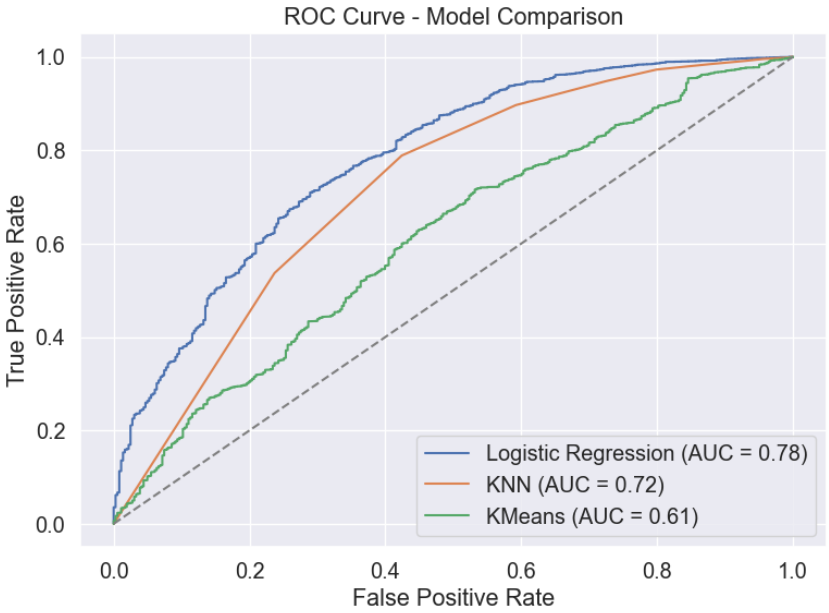


Figure B7: LSAC ROC curves.

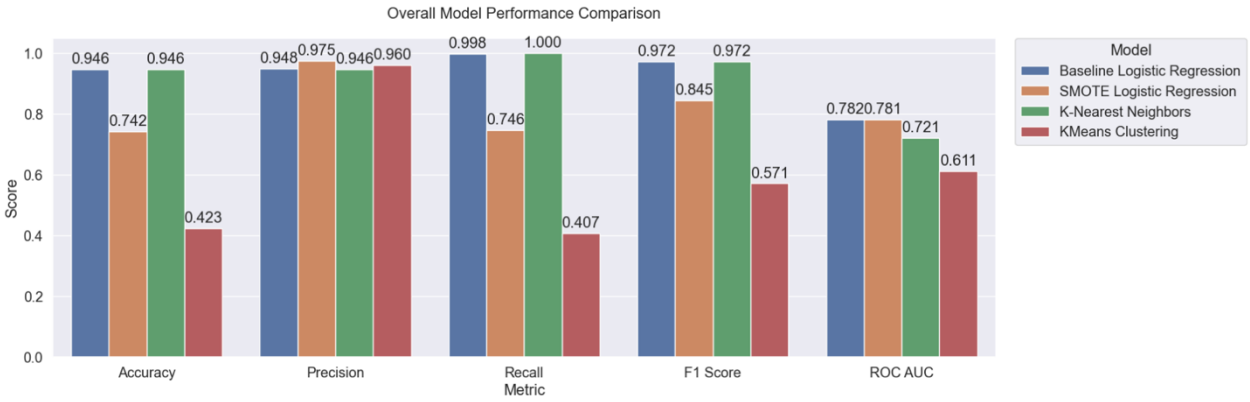


Figure B8: LSAC overall model performance comparison.

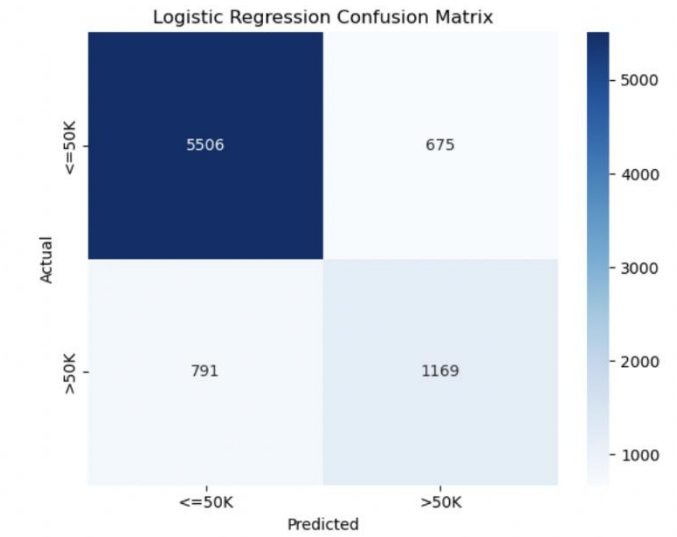


Figure B9: Adult Income Logreg confusion matrix.

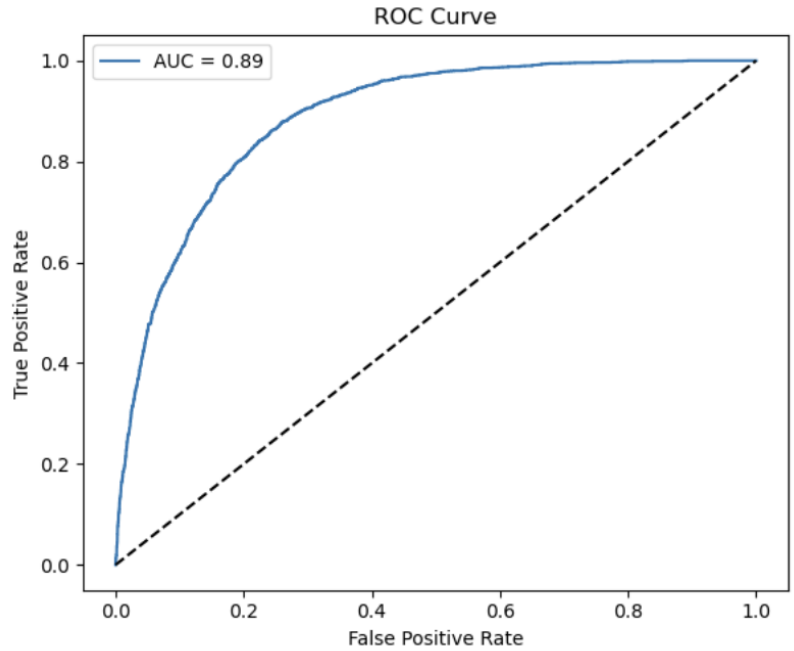


Figure B10: Adult Income Logreg ROC curve.

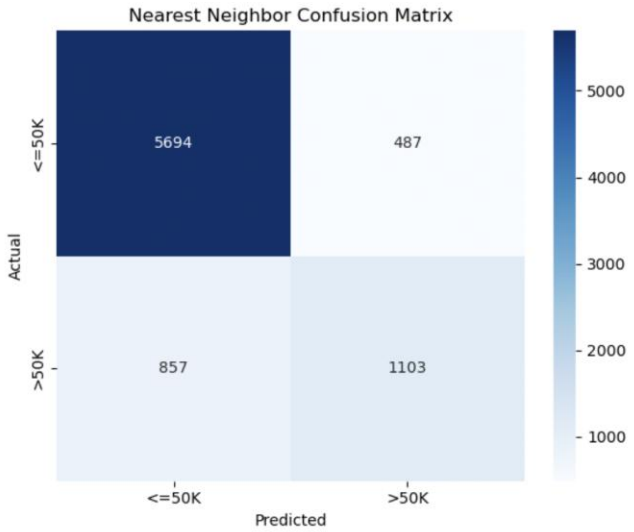


Figure B11: Adult Income KNN confusion matrix.

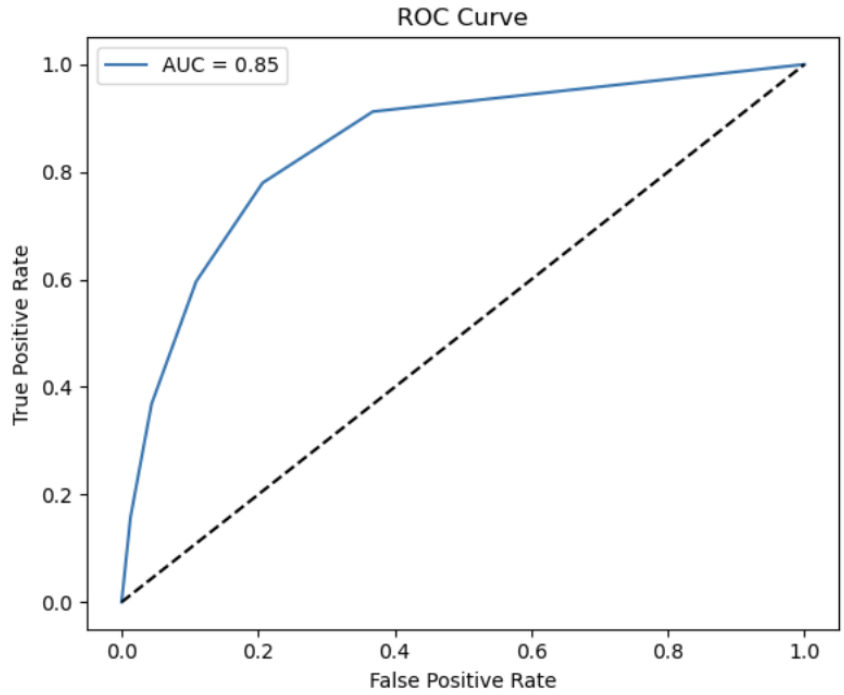


Figure B12: Adult Income KNN ROC curve.