Uncovering Hiring Inequities:

A Machine Learning Study of Fair Employment Factors Across the U.S.

Machine Learning

Prof. Ning Zhang

23rd April, 2025

Jacelyn Asafo-Adjei & Maya Ody-Ajike

Github Repo: https://github.com/mayaajike/machine-learning-25.git

INTRODUCTION

Unfair recruitment practices are a real phenomenon today. This is not only a case of individuals, but also of communities, companies deny opportunities to communities and perpetuate disparity. This led us to study in-depth the factors leading to such employment deficits.

We utilized machine learning platforms for this research to study employment trends in data. We collected data from the Data Commons platform and the census.gov API, which gave us large datasets on demographics, education, and economic status.

Our main goal was to utilize clustering and classification methods to identify patterns that are not necessarily obvious. By doing so, we aimed to better see how certain groups might be disproportionately impacted throughout the hiring process.

This was not just about examining the past, but also of contributing to the conversation about how we might design less biased, more equitable hiring processes in the future. In this study, we sought to center on the unseeable difficulties of workplace discrimination and observe how data science and machine learning might be utilized to effect change.

PROBLEM STATEMENT

The persistent issue of unfair hiring practices has become a pressing societal concern, necessitating a comprehensive investigation into the intricate factors that contribute to hiring disparities. Despite advancements in technology and an increasing awareness of the importance of diversity and inclusion, biased hiring processes continue to perpetuate inequalities across various demographic groups. The problem at hand is twofold: first, the lack of a standardized and transparent methodology for assessing and mitigating biases in hiring models; and second, the limited understanding of the spatial and regional dynamics that influence these disparities.

This project explores the inequities in hiring practices across the U.S states using demographic and employment-related data. The goal was to identify patterns in employment characteristics and better understand differences across states.

THE DATASET

The dataset was compiled from the U.S Census data sources[1], focusing on variables such as: employment rates by race and gender, education attainment levels, unemployment and poverty rates. Each state was characterized by these features. Importantly, the data set did not contain labeled ground truth categories such as “Good”, “Fair”, or “Poor” hiring practices. Thus, we used unsupervised clustering to assign initial labels before supervised classification.

PRE-PROCESSING STEPS

After pulling the data from the Census API[1], we converted the output, which were string values, to numbers, then we converted any error values returned from the API into NaN and then replaced all missing fields with the mean of their respective columns. Then, we did some feature

engineering by computing ratios between subgroup employment/education rates and total population rates. Lastly, all features were standardized using StandardScaler to normalize their distributions for model training.

LABEL GENERATION (UNSUPERVISED CLUSTERING)

Due to the absence of ground truth labels, we applied kMeans clustering to find patterns in the data. We applied the elbow method to choose the best number of clusters, which came out to 3, and used silhouette scores to check the quality of the groupings. We got a silhouette score of approximately 0.5. Clusters were then manually grouped into “Good Hiring Practices”, “Fair Hiring Practices”, “Poor Hiring Practices”.

After handling missing values, scaling features, creating ratios, and creating labels the dataset was organized as:

Column Names	Contents
Feature Columns	Numerical inputs that describe each state
Label Column (‘Hiring Practices’)	Categorical label assigned via kMeans clustering. (Good, Fair, Poor)

Each row corresponds to one U.S state with features and a label.

MODEL DESIGN AND JUSTIFICATION

After our pre-processing and label generation was completed, we moved on to training three different supervised learning models to classify the 50 U.S states under the three labels we had generated based on all the data retrieved from the U.S Census API[1]. The three different models trained were K-Nearest Neighbors (KNN), Random Forest, and Logistic Regression. We started with splitting the data set into a training set and a testing set using train_test_split. 70% of

the data was for training the models, and 30% was held back for testing. We trained three different classifiers independently on the training set. KNN classifies based on proximity to nearest neighbors; Random Forest used an ensemble of decision trees, and Logistic Regression modeled class probabilities. Each model was evaluated on the test set using accuracy, AUC, and confusion matrices to ensure a fair and consistent comparison. Cross-Validation and ROC analysis were also used to ensure robustness.

EVALUATION METRICS

Each model was evaluated based on: **Accuracy, Precision, Recall, and F1-score, and Area Under the ROC Curve (AUC)**

Model	Accuracy	AUC	Precision	Recall	F1-Score
KNN	0.9375	0.990	0.93	0.96	0.94
Random Forest	0.625	0.912	0.44	0.54	0.48
Logistic Regression	0.875	0.990	0.93	0.83	0.85

KNN showed the highest accuracy (93.75%) and AUC (0.99). Logistic Regression offered a slightly lower accuracy (87.5%) but matched KNN's strong AUC. Random Forest showed moderate accuracy but maintained a strong AUC. The Random Forest accuracy of 62.5% showed a moderate performance of the model which was expected because of our small dataset and noisy labels. However, the AUC of 91.2% shows the model's excellent ability to distinguish between "Good", "Fair", and "Poor". We also analyzed results further with a confusion matrix and ROC curves to understand how well each model predicted outcomes. To make sure our

results were reliable, we also used cross-validation to test each model on different parts of the data.

RESULTS

Pair Plot Visualizations[2]

We used pair plots to help visualize how different variables interact and influence hiring practices. Color coding made the charts easier to understand, orange showed good hiring practices, green represented fair hiring practices, and blue represented poor hiring practices. These visuals helped us better grasp how complex factors like race, education, and employment status affect outcomes across states.

kMeans Clustering

Using kMeans and the elbow method, we found the best number of clusters for the data, 3. We evaluated the clusters using the silhouette score to measure how well the groups were formed. This step helped us group states with similar hiring conditions.

KNN, RF, LR, Classification & Confusion Matrix

Our K-Nearest Neighbors model has an accuracy of 93.75%, Random Forest has an accuracy of 62.5% and Logistic Regression has an accuracy of 87.5%. The confusion matrix for the KNN showed strong results, it produced an almost perfect classification, showing excellent separation between all three categories, aligning with its high accuracy and AUC scores. The confusion matrix for the Logistic Regression model showed strong results as well, it was very strong on “Fair” and “Poor”, but was weaker on “Good”, this reflected its accuracy and AUC scores as well. Lastly, while Random Forest showed a decent performance for “Fair” and “Poor”, it failed to learn the pattern for the “Good” class, this is possibly due to the class imbalance, however the confusion matrix reflected its accuracy and AUC scores as well.

ROC Curves[3]

ROC curves visualize how well each model distinguishes between classes across various classification thresholds. The AUC gives a single-number summary of this performance and the higher AUC means better ranking ability.

Model	“Good” Class	“Fair” Class	“Poor” Class
KNN	1.00	0.97	1.00
LR	1.00	0.97	1.00
RF	0.89	0.91	0.94

KNN shows excellent ranking performance across all classes, this means that the KNN can confidently distinguish between all three categories, which is consistent with its high accuracy and excellent confusion matrix performance. Logistic regression also performed excellently in separating classes. Even though its confusion matrix showed slight confusion for “Good”, the model’s ranking performance remained excellent confirming its reliability and supporting its 87.5% test accuracy score. Random Forest’s ROC performance was solid, especially for “Fair” and “Poor” classes. However, the Good class performance lags slightly, which mirrors what we saw in the confusion matrix, making it less reliable than the KNN and Logistic Regression models.

Cross-Validation

Cross-Validation was used to check how each model performs across multiple splits of the data, giving a more robust estimate than a single train-test split.

Model	5-fold Cross-Validation Accuracy

K-Nearest Neighbors	0.980
Random Forest	0.829
Logistic Regression	0.980

KNN and Logistic Regression both demonstrate a very consistent and strong generalization across all folds. Random Forest lags behind again but only slightly, making it less stable and likely overfitting or struggling due to the small size of the dataset and the great class imbalance.

State-Level Predictions

We categorized the states into three groups. **Good Hiring Practices**, some states in this group are; District of Columbia, Indiana, Ohio, Texas, etc. **Fair Hiring Practices**, some states in this group are; Arizona, Arkansas, Georgia, Illinois, etc. **Poor Hiring Practices,** some states in this group are; Connecticut, Delaware, Florida, Hawaii, etc. This breakdown helps identify where improvements are most needed and where strong practices already exist. Overall, we had 6 states with Good Hiring Practices, 29 with Fair Practices and 17 with Poor Practices.

VISUALIZATION

In this section, I will be displaying the graphs and diagrams created from this dataset.

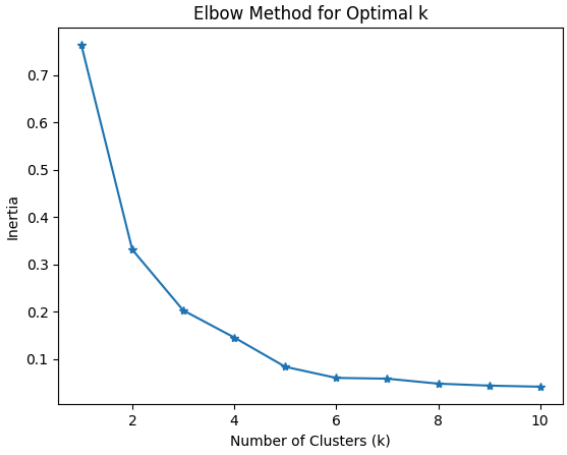


Figure of our elbow method showing k=3.

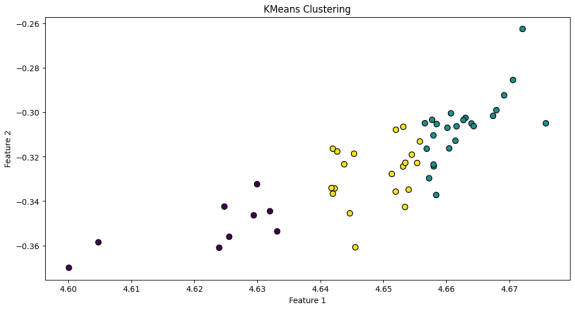


Figure showing our kMeans clusters.

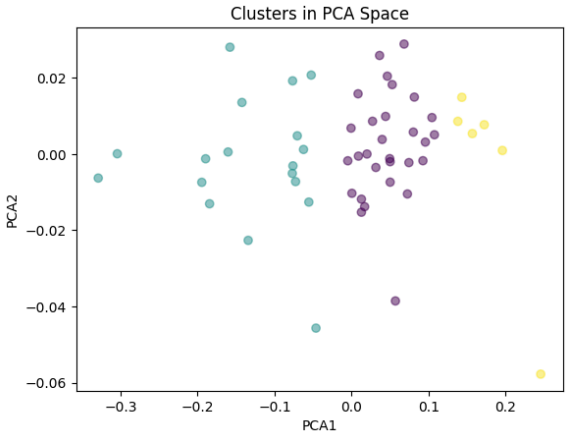


Figure showing clusters in PCA space and our silhouette score of 0.4.

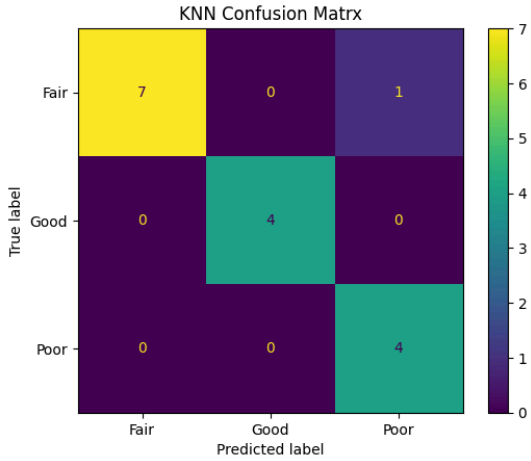
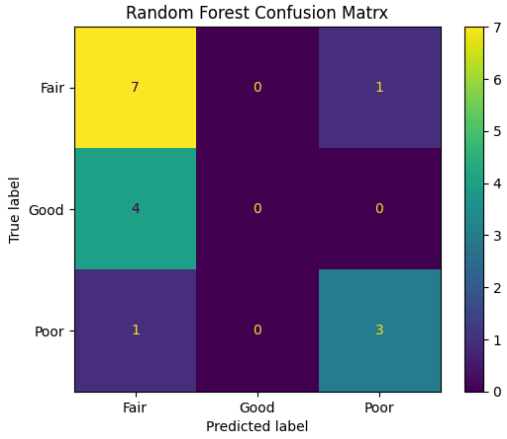


Figure showing KNN Confusion Matrix



Random Forest Confusion Matrix

Logistic Regression Confusion Matrix

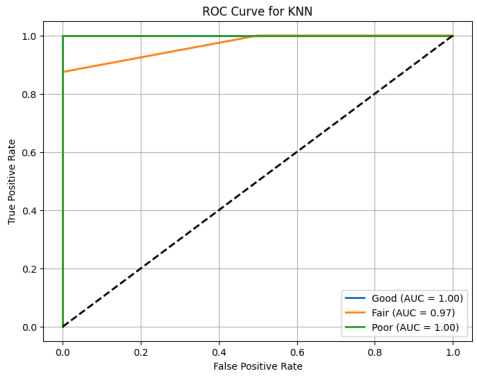


Figure showing the KNN ROC curve for each label.

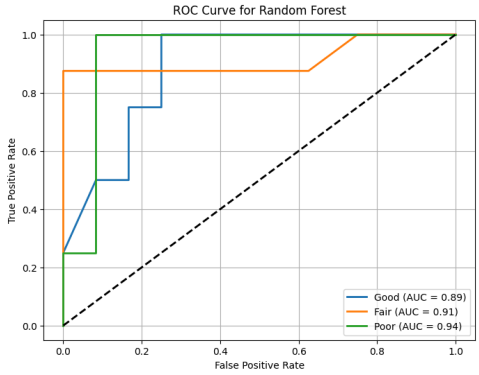


Figure showing the Random Forest ROC curve for each label.

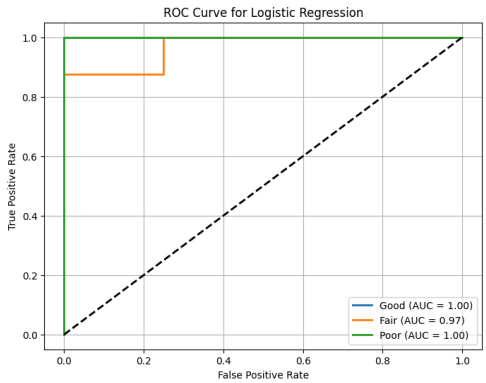
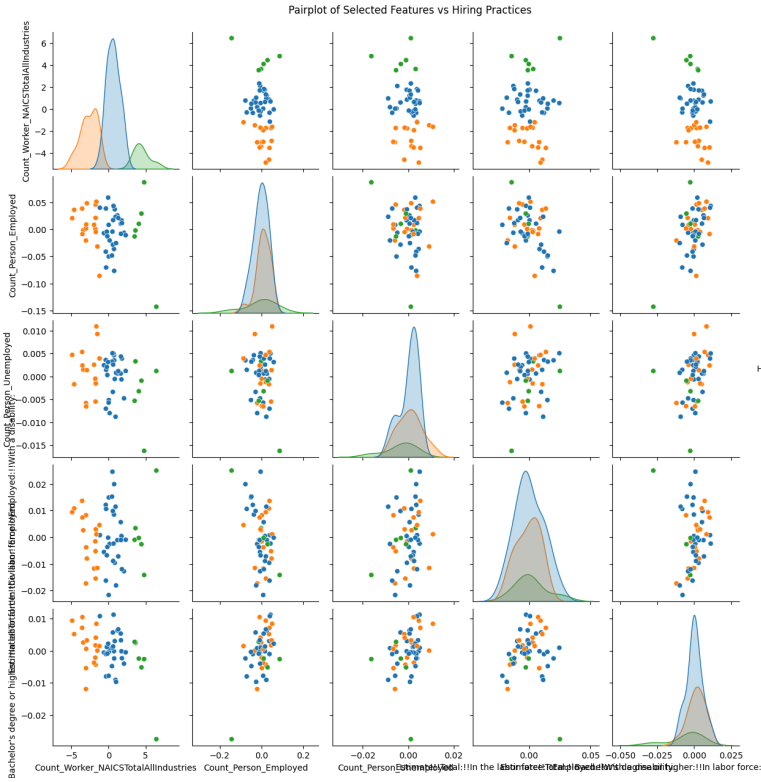


Figure showing the Logistic Regression ROC curve for each label.



Pair Plots for more extensive visualization.

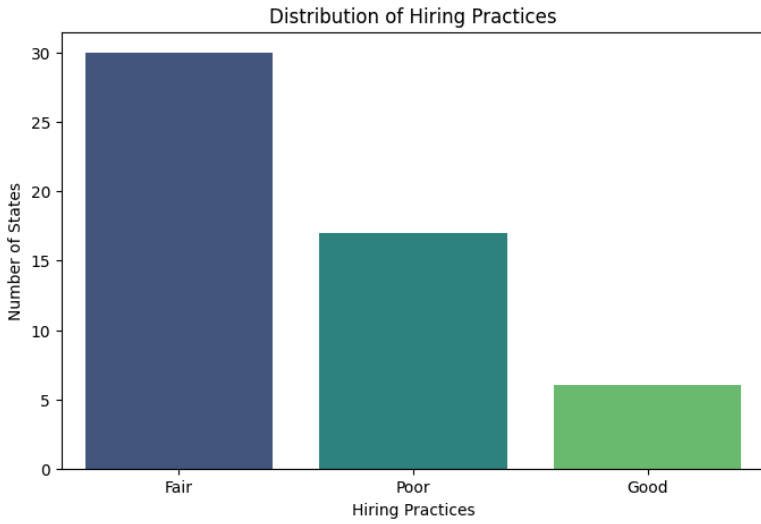


Figure showing distribution of hiring practices across the 50 U.S states

INDIVIDUAL CONTRIBUTIONS

In this project, we split the work done into two for each person,

- Jacelyn Asafo-Adjei:
    - Gathered and organized raw data sources for initial exploration.
    - Conducted background research on hiring practices and potential equity indicators.
    - Interpreted clustering results and selected label mappings (Good, Fair, Poor).
    - Developed and trained Logistic Regression model.
    - Designed visuals for presentation, created slides and summarized key findings.
- Maya Ody-Ajike
    - Designed and implemented the data preprocessing pipeline (handled NANs, standardization) and handled feature engineering by calculating ratios from raw data.
    - Performed unsupervised clustering using kMeans.
    - Developed and trained KNN and Random Forest models.
    - Evaluated model performance using confusion matrices, ROC curves, AUC metrics, and cross-validation.
    - Generated visualizations (pair plots, classification reports) and conducted final analysis.

CONCLUSION

The models successfully classified hiring practices into three categories with high accuracy and excellent ranking ability. The project showed that unsupervised clustering, when combined with supervised learning, can uncover patterns even in initially unlabeled datasets. K-Nearest Neighbors and Logistic Regression emerged as the strongest models, offering accurate and reliable classification of hiring inequities.

FUTURE WORK

To further enhance this project, future efforts can focus on:

- Collecting more granular data at the county or city level to allow for finer distinctions.
- Balancing the datasets through techniques like SMOTE to address class imbalance.
- Exploring additional models such as SVMs, Gradient Boosting Machines, or Neural Networks.
- Validating with real ground truth data if future labeled datasets become available, to move beyond clustering-generated labels.

REFERENCES

1. Data Commons Website. Retrieved from here https://datacommons.org/

2. Seaborn. (2021). Statistical Data Visualization. Retrieved from [https://seaborn.pydata.org/](https://seaborn.py data.org/)

3. Matplotlib Development Team. (2021). Matplotlib: Visualization with Python. Retrieved from https://matplotlib.org/

4. Doshi-Velez, F., & Kim, B. (2017). Fairness in Hiring: A Machine Learning Perspective. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

5. Burkhauser, R. V., Houtenville, A. J., & Tennant, J. R. (2019). Clustering States by Labor Market Outcomes for People with Disabilities. ILR Review, 72(1), 186–211.

6. Chouldechova, A. (2017). Improving Fairness in Machine Learning Models for Hiring.

7. Zhang, J., & Zhao, Z. (2018).
Fairness-aware Learning for Ranking in Job Applications. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.

8. Census Data Api https://www.census.gov/data/developers.html

9. Stone, C., & Stone, D. L. (2015). Factors affecting hiring decisions about veterans. Human Resource Management Review, 25(1), 68–79.
https://doi.org/10.1016/j.hrmr.2014.06.003