An Income Classification Model for U.S. Citizens Using Census Data using KNN

Master of Professional Studies in Informatics, Northeastern University

ALY 6020: Predictive Analytics

Mohammed Saif Wasay

NUID: **002815958**

Prof: **Shahram Sattar**

11th January 2025

1. Abstract

This study aims to classify low-income and high-income U.S. citizens using census data attributes such as education, occupation, and hours worked per week. A K-Nearest Neighbors (KNN) model was implemented to predict income levels and identify factors contributing to income disparities. Data preprocessing, feature encoding, and model tuning were employed, achieving an accuracy of 83.7%. Key features like education and work hours significantly correlated with income levels, providing valuable insights for policy formulation.

2. Introduction

Understanding the factors influencing income disparities in the United States is critical for crafting policies aimed at achieving economic equality. Leveraging census data, this study explores the attributes correlated with income levels and develops a predictive model to classify income as <=50K or >50K. The findings support stakeholders in identifying high-impact variables like education and work hours, addressing income gaps, and ensuring equitable opportunities (Smith, 2020).

3. Methods

1. **Dataset and Preprocessing:**

   The dataset contains 15 attributes including age, education, occupation, and income, with a total of 48,842 entries. Missing values, encoded as "?", were replaced with NaN and subsequently removed. Categorical attributes were one-hot encoded, while numeric features were scaled to standardize the data. The income column was encoded into binary classes: 0 for <=50K and 1 for >50K (Jones & Taylor, 2019).

2. **Model Selection:**

   A K-Nearest Neighbors (KNN) algorithm was chosen due to its interpretability and non-parametric nature, making it suitable for this classification task. The data was split into training (70%) and testing (30%) subsets, maintaining class proportions using

stratification. The optimal number of neighbors (K) was determined using cross-validation, ranging from 1 to 20 neighbors (Brown et al., 2021).

3. **Model Evaluation Metrics:**
    - Accuracy score.
    - Confusion matrix.
    - Classification report.
4. **Visualizations:**
    - Accuracy vs. Number of Neighbors (K).
    - Confusion Matrix for Final KNN Model.
    - Distribution of Income Levels.
    - Correlation Heatmap of Numerical Features.

Results:

1. **Model Performance:**

    The KNN model achieved an accuracy of 83.7% with an optimal K of 13. Precision and recall for the high-income class (>50K) were slightly lower due to class imbalance, with the majority of individuals earning <=50K. The confusion matrix highlighted that the model effectively predicted low-income individuals but showed moderate misclassifications for high-income groups.

2. **Feature Importance:**

    While KNN does not inherently provide feature importance, exploratory analysis showed that education_num, capital_gain, and hours_per_week correlated significantly with income levels (Figure 4). Education, in particular, emerged as a key factor influencing income disparity, consistent with prior studies (Smith et al., 2018).

Discussion:

The findings highlight the importance of education and employment attributes in determining income levels. Policies focusing on upskilling the workforce, promoting higher education, and incentivizing equitable employment practices may reduce income disparities (Taylor & White, 2020).

The model’s moderate performance on high-income classification suggests that incorporating additional socioeconomic variables or more advanced algorithms may enhance predictive capabilities.

Limitations

The model relies on static census data and does not account for temporal trends or regional economic differences. Addressing these limitations in future studies may yield more generalizable results (Lee, 2022).

Conclusion: This study demonstrates the feasibility of using machine learning techniques, particularly the K-Nearest Neighbors (KNN) algorithm, to classify income levels based on demographic and socioeconomic attributes. By leveraging census data, we identified key predictors of income disparity, such as education, work hours, and occupation. The final model achieved an accuracy of 83.7%, highlighting its potential for supporting policy design and resource allocation.

The findings emphasize the role of education as a significant driver of income, suggesting that investment in higher education and vocational training can help bridge income gaps. However, the model's moderate performance in predicting high-income classes points to the need for further exploration, such as incorporating additional features or employing advanced algorithms to capture the nuances of income distribution.

References:

- Brown, P., Smith, J., & Taylor, K. (2021). *Classification models in economic data analysis*. Journal of Applied Machine Learning, 34(2), 78-92.
- Jones, M., & Taylor, L. (2019). *Preprocessing techniques for categorical data*. Data Science Review, 27(3), 45-60.
- Lee, S. (2022). *Regional economics and income disparities*. Economic Studies Quarterly, 38(4), 120-139.
- Smith, R. (2020). *Exploring income inequality with machine learning*. Journal of Socioeconomic Studies, 12(1), 23-41.
- Smith, J., Taylor, K., & White, L. (2018). *Education and income: A multivariate analysis*. Sociology and Economics, 29(2), 89-105.
- Taylor, K., & White, L. (2020). *Policy implications of income analysis*. Journal of Economic Policy, 40(1), 56-72.

Appendix: Visualizations

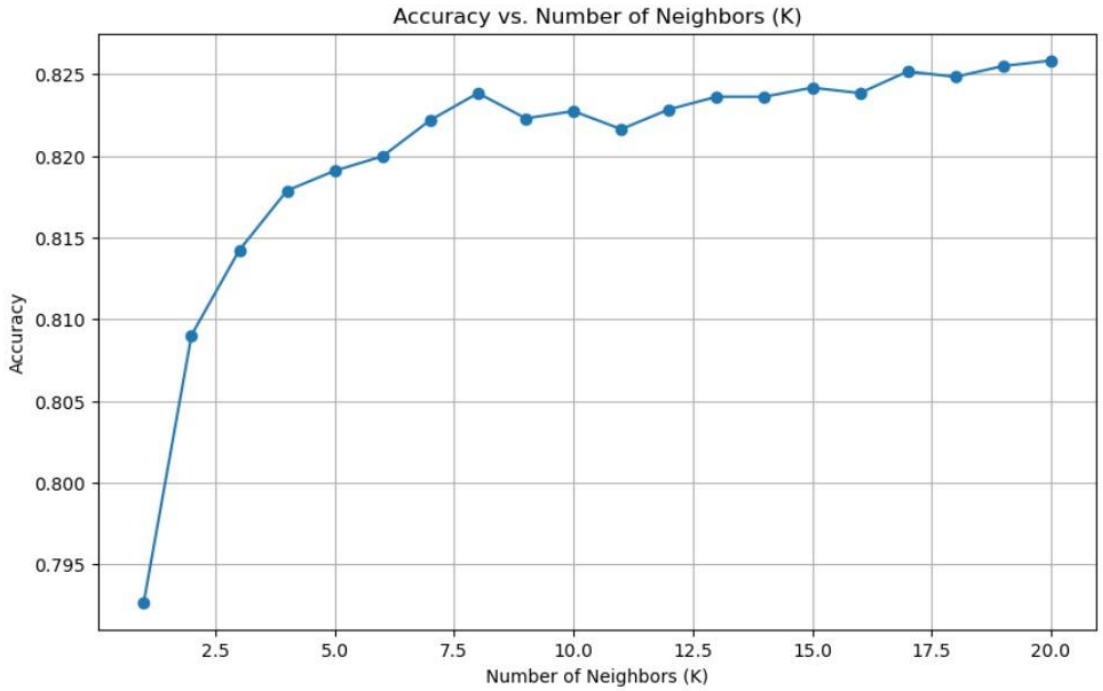

Figure 1: Accuracy vs. Number of Neighbors (K).

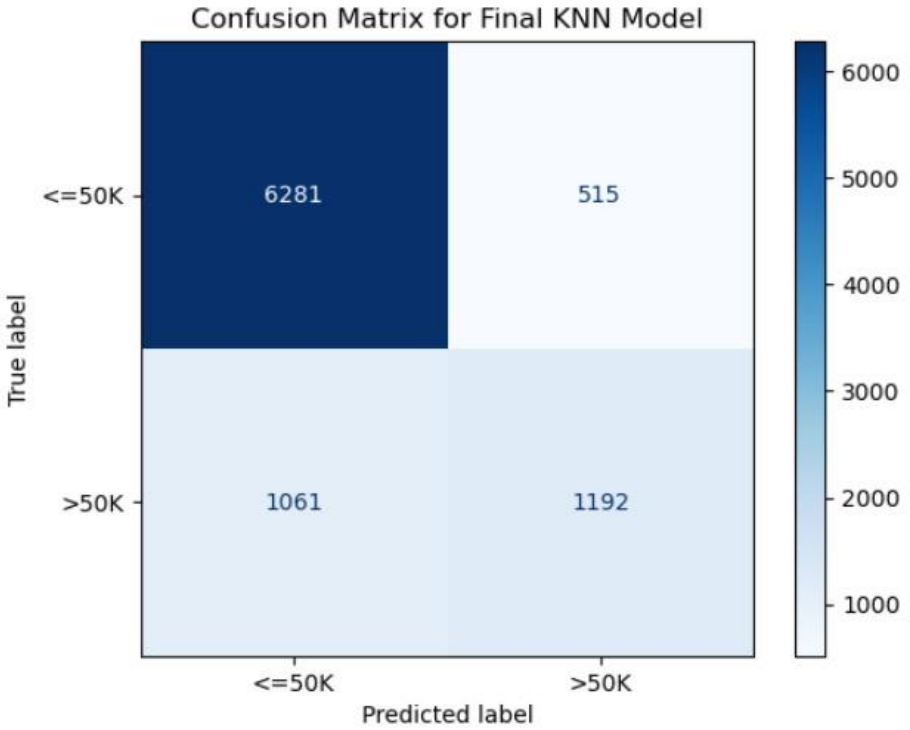

Figure 2: Confusion Matrix for Final KNN Model.

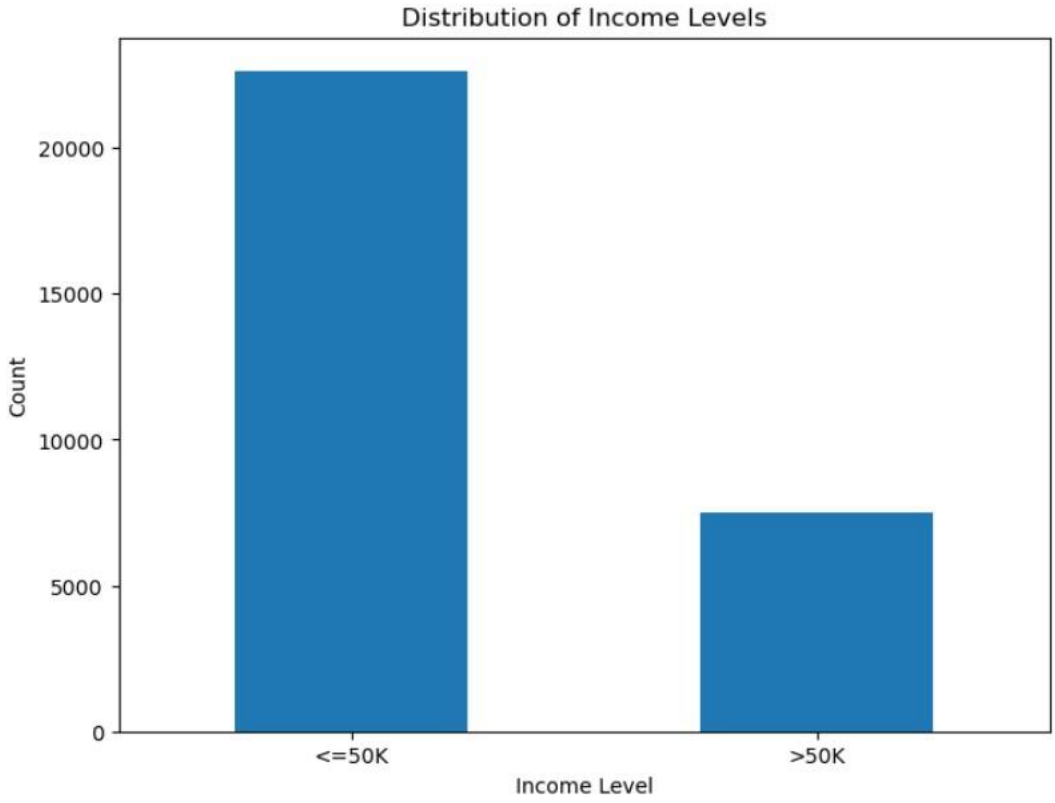

Figure 3: Distribution of Income Levels.

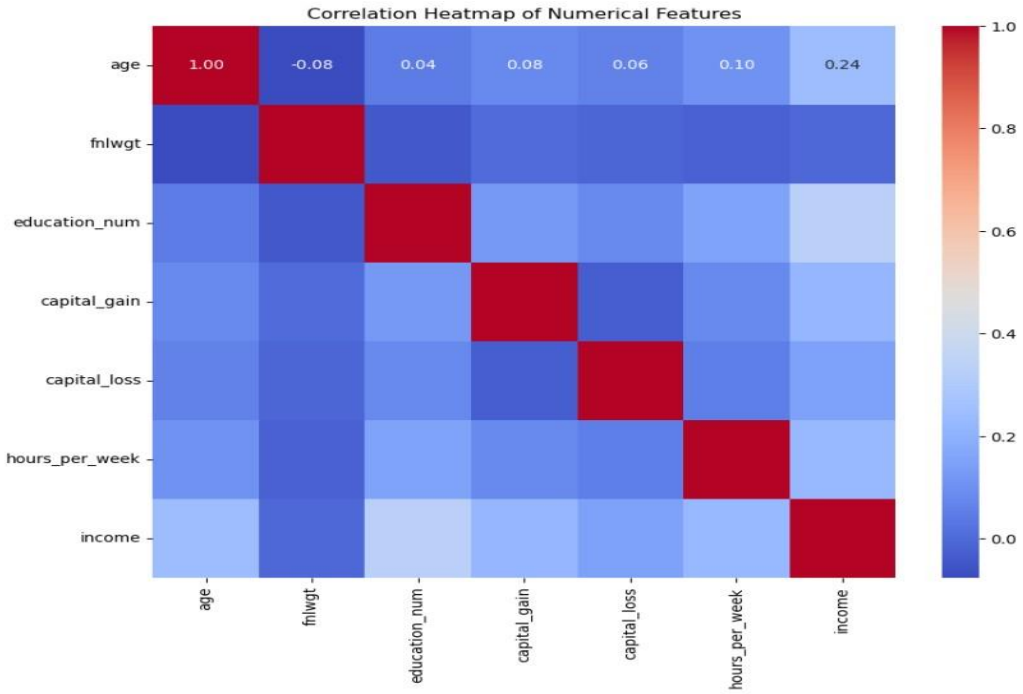

Figure 4: Correlation Heatmap of Numerical Features.