**ALY 6020**

**Report**

**Predictive Analytics**

By: Muhammad Hassan Zahoor

11th January 2025

**Introduction**

This report outlines the process and results of building a k-Nearest Neighbors (k-NN) model to classify income levels (≤50K or >50K) using a dataset containing demographic and occupational attributes. The primary goals of this project are:

1. To analyze the relationships between features and income levels.
2. To identify patterns and insights within the data using visualizations.
3. To develop and evaluate a k-NN model for income classification.

**Analysis**

Below is a detailed analysis of the dataset and the graphs produced during the exploration and modeling process. Please add the provided screenshots in the spaces indicated.

**1. Missing Values Analysis**

The dataset was checked for missing values to ensure data quality. Columns containing missing values were identified, and appropriate preprocessing steps (e.g., removing rows or imputing values) were taken to prepare the data for analysis.

**2. Correlation Heatmap**

A correlation heatmap was generated to examine relationships between numerical features and the target variable (income). This visualization highlights features with significant positive or negative correlations that could be important predictors for the model.

* **Key Observations**:
  + Features like Education\_Num and Hours\_per\_Week show moderate positive correlations with income (>50K).
  + Variables like Capital\_Gain also exhibit strong positive correlations with income.

**3. Categorical Feature Distribution vs. Income**

Several bar plots were created to show the distribution of categorical variables (e.g., Workclass, Education, Occupation) with respect to income levels.

* **Key Observations**:
  + Education: Higher education levels (e.g., Bachelors, Masters) are associated with higher income (>50K).
  + Workclass: Private-sector employees form the largest group in both income categories, but public-sector and self-employed individuals are more likely to have income >50K.
  + Relationship: Individuals classified as "Husband" dominate the >50K category.

*A graph of different colored bars

Description automatically generated with medium confidence*

**A graph with blue and orange bars

Description automatically generated**

**4. Accuracy vs. K Value**

A line plot was generated to analyze the accuracy of the k-NN model as the number of neighbors (“k”) varied between 1 and 20. The purpose of this graph is to identify the optimal value of “k” that balances underfitting and overfitting.

* **Key Observations**:
  + Accuracy increases initially as “k” increases, reaching a peak at , after which it stabilizes or slightly declines.
  + was chosen as the optimal value for the final model.

*A graph with a line going up

Description automatically generated*

**5. Confusion Matrix**

A confusion matrix was generated to evaluate the performance of the final k-NN model. It provides counts of true positives, true negatives, false positives, and false negatives.

* **Key Observations**:
  + The model performs well in identifying individuals with income ≤50K (high true negatives).
  + The classification of individuals with income >50K is slightly less accurate, indicating room for improvement in detecting high-income earners.

*A blue squares with white text

Description automatically generated*

**Results**

**Classification Report**

The classification report provides precision, recall, f1-score, and support for both income categories (≤50K and >50K):

* **Precision**: Indicates the proportion of true positives out of all predicted positives.
  + For income ≤50K: 85%
  + For income >50K: 72%
* **Recall**: Indicates the proportion of true positives out of all actual positives.
  + For income ≤50K: 93%
  + For income >50K: 52%
* **F1-Score**: Harmonic mean of precision and recall, providing a balanced measure.
  + For income ≤50K: 0.89
  + For income >50K: 0.60
* **Support**: The number of actual occurrences of each class in the test set.
  + For income ≤50K: 6,757
  + For income >50K: 2,288

**Overall Model Performance**

* **Accuracy**: The overall accuracy of the model is 82.76%.
* **Macro Avg**: Averaging precision, recall, and f1-score across both classes results in 79% precision and 72% recall.
* **Weighted Avg**: Weighted by the support of each class, the precision and recall are both approximately 83%.

This classification report highlights that the model performs better for the majority class (≤50K) but struggles with the minority class (>50K). This imbalance could be addressed using techniques like oversampling, undersampling, or more advanced algorithms.

*A screenshot of a computer

Description automatically generated***Conclusion**

This analysis highlights the effectiveness of k-NN for income classification using demographic and occupational data. The visualizations provided valuable insights into the relationships between features and income levels. While the model performs well overall, future improvements could focus on:

1. Exploring advanced feature engineering or dimensionality reduction techniques.
2. Addressing class imbalance to improve recall for the >50K category.
3. Comparing k-NN with other classification algorithms like logistic regression, decision trees, or support vector machines.

By incorporating these steps, the model’s predictive performance and interpretability could be further enhanced.

**References**

** Alpaydin, E. (2020). Introduction to Machine Learning (4th ed.). MIT Press..**

** García, S., Luengo, J., & Herrera, F. (2015). Data Preprocessing in Data Mining. Springer.**

** Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.**

** Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.**

** Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.**