COMP 6721-Project Proposal: Adult Census Income Analysis

Dina Omidvar
Concordia University
40239883

Niloofar Tavakolian
Concordia University
40220767

Nastaran Naseri
Concordia University
40215694

SeyedehMojdeh Haghighat Hosseini
Concordia University
40171693

1 PROBLEM STATEMENT AND APPLICATION

The Adult Census Income dataset provides valuable information for analyzing factors that contribute to income levels in the adult population [1]. The problem of predicting whether an individual earns more than $50,000 per year based on various socioeconomic attributes is important in understanding income inequality and developing strategies for economic empowerment.

This dataset can be used to explore the relationship between demographic factors such as age, education, occupation, marital status, and income levels. By identifying influential factors, we can gain insights into the socioeconomic dynamics and potential challenges faced by different demographic groups.

The associated challenges of this problem application include:

- Handling missing values: The dataset may contain missing data, which needs to be addressed through appropriate handling techniques.
- Dealing with class imbalance: The dataset may have an imbalance in the number of instances belonging to different income classes, requiring techniques such as oversampling or undersampling to ensure model performance.
- Feature engineering: Extracting meaningful information from the dataset and selecting relevant features that can best predict income levels.
- Interpreting and explaining results: Providing meaningful insights and interpretations of the model's predictions to understand the factors driving income disparities and potential avenues for intervention.

The expectations and goals throughout the development of the application may include:

- Developing a predictive model: Building a machine learning model capable of accurately predicting income levels based on the provided features.
- Feature importance analysis: Identifying the most influential factors affecting income levels to understand the socio-economic dynamics.
- Evaluating model performance: Using appropriate evaluation metrics to assess the model's accuracy, precision, recall, and F1-score.
- Comparing different methodologies: Exploring various machine learning algorithms and deep learning approaches (such as CNN or RNN models) to evaluate their performance in predicting income levels.
- Generating actionable insights: Providing insights that can inform policymakers, social scientists, and economists in formulating strategies to address income disparities and promote economic equality.

2 DATASET SELECTION

The Adult Census Income dataset was obtained from Kaggle [2] : https://www.kaggle.com/datasets/uciml/adult-census-income?resource=download.
It consists of approximately 48,842 instances, each with 15 attributes (features) and a target variable indicating income level (<=50K or >50K). The features include demographic information such as age, education, occupation, marital status, race, gender, and work-related factors like hours per week and native country. The dataset provides a comprehensive representation of the socioeconomic characteristics of individuals in the United States.

3 POSSIBLE METHODOLOGY

The possible methodology for analyzing the Adult Census Income dataset involves the following steps:

- Data preprocessing: This step includes handling missing values, encoding categorical variables, and normalizing numerical features. It ensures data compatibility and quality for subsequent analysis.
- Exploratory data analysis: Conducting statistical analysis and visualizations to gain insights into the dataset's characteristics, understand relationships between variables, and identify potential patterns or anomalies.
- Model selection: Exploring a range of models such as logistic regression, decision trees, random forests, support vector machines, and deep learning models (e.g., CNN or RNN) to determine the most suitable approach.
- Model evaluation: Employing evaluation metrics such as accuracy, precision, recall, and F1-score to assess the performance of different models.
- Cross-validation and hyperparameter tuning: Employing techniques like k-fold cross-validation to assess model performance on different subsets of data. Optimizing hyperparameters to improve model accuracy and generalization.
- Comparisons and analysis: Comparing the performance of different models, evaluating the impact of feature selection, and examining the effects of varying algorithmic approaches. Analyzing and interpreting the results to extract meaningful insights.
- Potential applications: The analysis and comparisons can be useful for researchers, economists, and policymakers in understanding the factors driving income disparities, designing targeted interventions, and formulating evidence-based policies for economic empowerment [3].

REFERENCES

[1] R. Kohavi, “Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid,” *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, September 1996.

[2] J. Weston, S. Bengio, and N. Usunier, “Large scale image annotation: Learning to rank with joint word-image embeddings,” *Machine Learning*, vol. 81, no. 1, 2010.

[3] M. D. Skowronski and K. M. Carley, “Extracting patterns of behavior in sociolinguistic context„” *Social Networks*, vol. 35, no. 3, 2013.