DATA MINING PROJECT - ISM6136.004F23

Prof. TIMOTHY C SMITH

TEAM BLEED BLUE

TEAM MEMBERS –

Priyaanka Reddy Boothkuri - U70184580
Niharika Mullangi - U57336324
Manogna Tummanepally - U39202669
Pranathi Cheemarla - U22927056
Santhoshini Bojanapally - U88362375

Project Proposal: Predicting Income Levels from Census Data

Data Set - https://archive.ics.uci.edu/dataset/20/census+**income**

DOI:10.24432/C5GP7S

Introduction:

The project aims to predict whether an individual's income exceeds $50,000 per year based on census data. This task falls under the domain of classification in machine learning and can have significant social and economic implications. The dataset, commonly referred to as the "Adult" dataset, contains both categorical and integer features, making it a challenging and interesting problem to address.

Problem Statement:

The primary goal of this project is to build a machine learning model that can accurately classify individuals into two income groups: those earning more than $50,000 per year and those earning $50,000 or less. This predictive model can help in identifying the factors that contribute to higher incomes, which can be valuable for targeted policy-making, resource allocation, and economic analysis.

Dataset:

The dataset used for this project contains the following characteristics:

Type: Multivariate

Subject Area: Social Science

Associated Task: Classification

Feature Types: Categorical and Integer

Number of Instances: 48,842

Number of Features: 14

Approach:

Data Collection: The first step is to collect and load the dataset into the project environment. The dataset can be obtained from a reliable source, such as the UCI Machine Learning Repository.

Data Preprocessing: Data preprocessing is crucial to handle missing values, encode categorical features, and scale/normalize numerical features. Exploratory data analysis (EDA) will also be conducted to understand the dataset's distribution and relationships between variables.

Feature Selection: Feature selection techniques will be applied to identify the most relevant features that have the most significant impact on an individual's income.

Model Building: Various classification algorithms will be considered, including logistic regression, decision trees, random forests, support vector machines, and gradient boosting. Multiple models will be built and evaluated for accuracy and robustness.

Model Evaluation: The models' performance will be assessed using metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve. Cross-validation will be performed to ensure the models generalize well.

Hyperparameter Tuning: Grid search or randomized search will be used to optimize the model's hyperparameters, further improving predictive performance.

Model Deployment: The best-performing model will be deployed as a web application, API, or other suitable platforms, making it accessible to users for income predictions.

Expected Outcomes:

A machine learning model that can predict income levels based on census data.

A better understanding of the factors that influence income levels.

Insights for policymakers to target interventions for individuals with lower incomes.

Conclusion:

Predicting income levels from census data is a valuable task with applications in social science and economics. The project will involve data collection, preprocessing, feature selection, model building, evaluation, and deployment. The ultimate goal is to create a reliable and interpretable model for predicting income levels and generating insights for societal and economic improvement.