CITIZENS INCOME PREDICTION

Venkata Sriram Rachapoodi
ID: 01897282
Department of Computer Science
Kennedy College of Sciences
University of Massachusetts Lowell
venkatasriram_rachapoodi@student.uml.edu

Kamal Yeshodhar Shastry Gattu
ID: 02007505
Department of Computer Science
Kennedy College of Sciences
University of Massachusetts Lowell
kamalyeshodharshastry_gattu@student.uml.edu

Aditya Santhoshkumar Karnam
ID: 02003014
Department of Computer Science
Kennedy College of Sciences
University of Massachusetts Lowell
aditya_karnam@student.uml.edu

1. OBJECTIVE:

For any government or an organization to develop and implement benefit plans for its citizens, it is important to have a rough knowledge of the mode of living of its citizens and how much income citizens of a country have. Having a rough picture of the information helps the government to formulate plans based on the requirements of the people.

This paper proposes to estimate the income of citizens and classify people into two categories based on various dependent properties of a person collected during a Census. We use different Classification algorithms like Naïve-Bayes Classifier, Logistic Regression, Support Vector Machine, K Nearest Neighbors Classifier, Decision Tree, and Random Forest algorithms to perform this task.

Furthermore, Compare the results obtained from the above approaches and find which algorithm is more suitable for the current scenario to produce the most accurate prediction.

2. CURRENT STATE OF ART

The problem has been solved using one customized algorithm [1]. Using any one algorithm on a dataset may not give the most accurate solution in a scenario. Hence it is required to test multiple algorithms and find out which algorithm is best suited for a particular case and use that to solve the problem.

3. APPROACH:

In the system, we collect data from the census dataset and preprocess and clean the data. Then we follow Supervised Learning to train the system to predict the income class of a person.

We train six different systems on the dataset individually by applying different algorithms for each system. We then test the trained systems on the same test dataset and map the accuracy of the systems using based on True Positive Rate and True Negative rates using Confusion Matrix and Classification Report for each of the outcomes.

Based on the results of these, we conclude which algorithm is more accurate.

4. DATASET:

The dataset being used is taken from the UCI Dataset Repository [2] which contains the data of citizens above age 16 from the 1994 Census database.

5. TIMELINE:

Dataset Collection and Data Preprocessing	8 February - 13 February
1st Algorithm of each teammate status check	13 February - 27 February
1st Algorithm Completion	27 February - 13 March
Project Progress Presentation	13 March - 22 March
2nd Algorithm of each teammate status check	22 March - 27 March
2nd Algorithm Completion	27 March - 10 April
Comparison of results and aggregating code	10 April - 15 April
Preparing Project Report	15 April – 20 April
Project Submission	22 April

6. ROLES AND TASKS:

Logistic Regression Naïve Bayes Classifier	Kamal Yeshodhar Shastry
Decision Tree Support Vector Machine	Venkata Sriram
K Nearest Neighbors Random Forests	Aditya

[1] Ron 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, 1996
https://dl.acm.org/doi/10.5555/3001460.3001502

[2] UCI Census Income Dataset
https://archive.ics.uci.edu/ml/datasets/census+income