AIT614_005

Project Abstract
Analysis of the Adult Dataset

Team – 5
Pravallika Avula
Saipriya Bethi
Sai Roopesh Diddi
Snehita Moturu
Sameeksha Muralidhar Gupta

Professor – Dr. Lindi Liao
AIT614 - 005: Big Data Essentials

George Mason University

24 April 2023

Project Abstract
Team 5

Abstract:

The employment rate and general health of a nation's population are important indicators of its economy. Many people use their income to determine their well-being of the people. In the past, the household's standard of living was assessed using its income. There are other indicators as well, though. In addition to these factors, age, race, family, gender, and education also have an impact on income. A person's income may change every day for a variety of reasons. As a result, it is critical to assess an individual's financial situation and estimate their earnings.

In this project, we will analyze the adult dataset which is retrieved from the census conducted in 1994. Our analysis will mainly focus on the prediction of income levels based on the data across different countries and other factors like age group, education, occupation, working hours per week, and so on. Initially, we will be doing exploratory and descriptive data analysis to gain insights on different characteristics of data like summary statistics, and correlation between different variables to identify patterns. Then, we will perform the predictive data analysis using databricks so that we can predict the income of a person based on their demographic and socioeconomic variables. After this, we will be creating models using machine learning and regression where the dataset is split into testing and training data so that we can predict the best models for classification with the lowest error rate, highest accuracy, and highest precision. Further, we will be performing classification and clustering analysis to identify characteristics of different groups based on income level and cluster them based on similarities. For this analysis, we will be using pySpark as it enables fast and flexible processing of large-scale data sets. Finally, after the analysis, we will make recommendations on how to improve the financial status of citizens based on the conclusions from the analysis using different attributes.

Keywords:

1. **Exploratory Analysis:** A process of examining and visualizing data to gain insights and identify patterns or trends.
2. **Descriptive Analysis:** A process of summarizing and presenting data in a format that is easy to understand and interpret like visualizations.
3. **Predictive Analysis:** A process of using data and statistical algorithms to make predictions about future outcomes.
4. **Machine Learning Algorithms:** A set of mathematical models and algorithms that enable computers to learn and make predictions from data.
5. **Regression:** A statistical analysis technique used to model the relationship between a dependent variable and one or more independent variables.
6. **Classification Analysis:** A type of machine learning analysis that categorizes data into different classes, based on its attributes.
7. **Clustering Analysis:** A type of machine learning analysis that groups data into clusters based on similarities in their attributes.
8. **Databricks:** A cloud-based platform that provides an analytical workspace for data engineering, data science, and business analytics.
9. **pySpark:** A Python-based API for Apache Spark. will be used as the programming language for interfacing with both MongoDB and Apache Spark.

References:

[1] Dataset link: *UCI Machine Learning Repository: Adult Data Set.* [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Adult.

[Accessed: 24-Apr-2023].

[2] G. L. Team, "Understanding data visualization techniques," *Great Learning Blog: Free Resources what Matters to shape your Career!*, 02-Sep-2022. [Online]. Available: https://www.mygreatlearning.com/blog/understanding-data-visualization-techniques/.

[Accessed: 24-Apr-2023].

[3] E. Stevens, E. S. O. from England, E. Stevens, and O. from England, "The 7 most useful data analysis techniques [2023 guide]," *CareerFoundry*, 17-Apr-2023. [Online]. Available: https://careerfoundry.com/en/blog/data-analytics/data-analysis-techniques/.

[Accessed: 24-Apr-2023].

[4] "Data Mining and visualization group Silicon Graphics, inc.." [Online]. Available: http://robotics.stanford.edu/people/ronnyk/nbtree-talk.pdf.

[Accessed: 24-Apr-2023].

[5] U. S. C. Bureau, *Census.gov*, 06-Apr-2023. [Online]. Available: https://www.census.gov/.

[Accessed: 24-Apr-2023].