**Predicting Income Levels with Machine Learning**

A Course Project report submitted

in partial fulfillment of the requirement for the award of the degree

**BACHELOR OF TECHNOLOGY**

**IN**

## ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

**By:**

**Zuhayr Zia (2203A52065)**

Under the guidance of

**D.Ramesh**

Associate professor, Department of CSE.

Introduction

In this project, we explore the application of machine learning and artificial intelligence techniques to answer a key question: Can we accurately predict income levels based on a combination of demographic and economic attributes? Income classification plays a key role in a variety of real-world scenarios, including credit approval, marketing targeting, and financial planning. Our project aims to provide insight into the potential of using machine learning for accurate income classification.

Abstract

This project revolves around classifying income levels and categorizing individuals into income groups ("<50K" or ">50K") using their demographic and economic attributes. The use of machine learning and AI methodology forms the core of our analysis. By assessing the effectiveness of different classification models and evaluating their performance, we contribute to the body of knowledge in revenue prediction and classification.

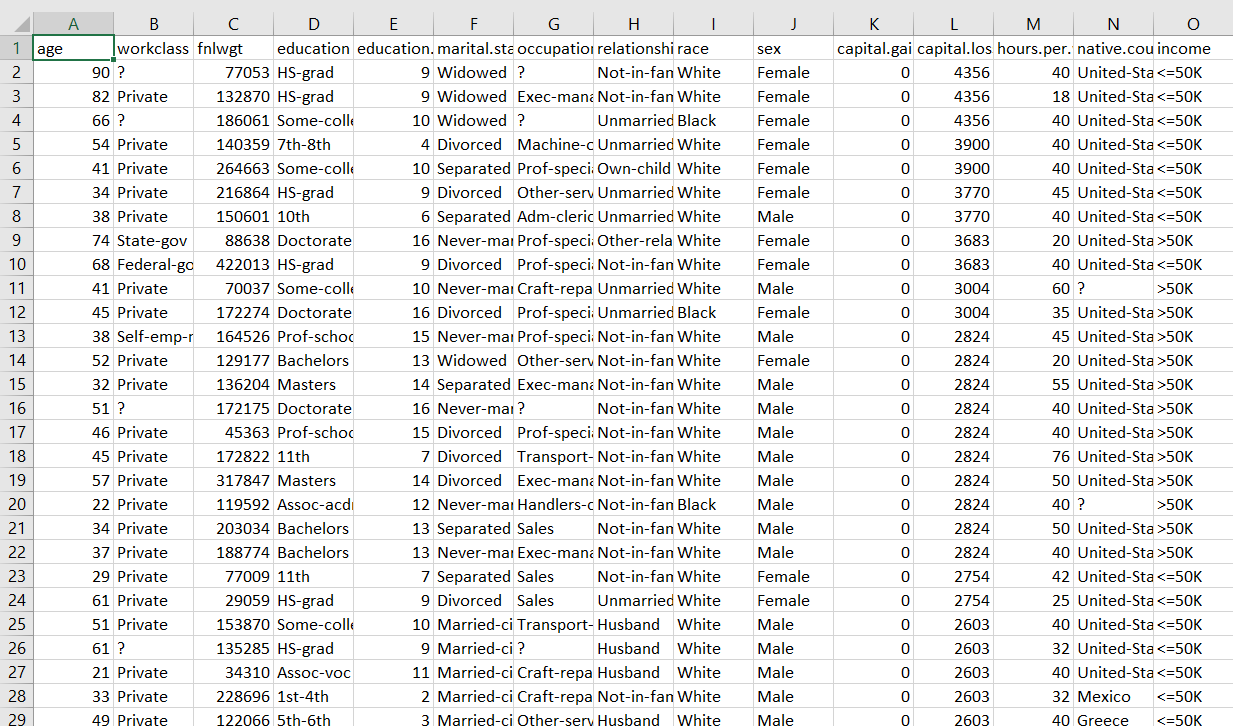
Literature Survey

To inform our project, we conducted a literature review to understand the existing research on income prediction and classification. By examining related work, we gained insight into the importance and challenges of income classification and how machine learning techniques have been applied in this context.

Data Pre-Processing

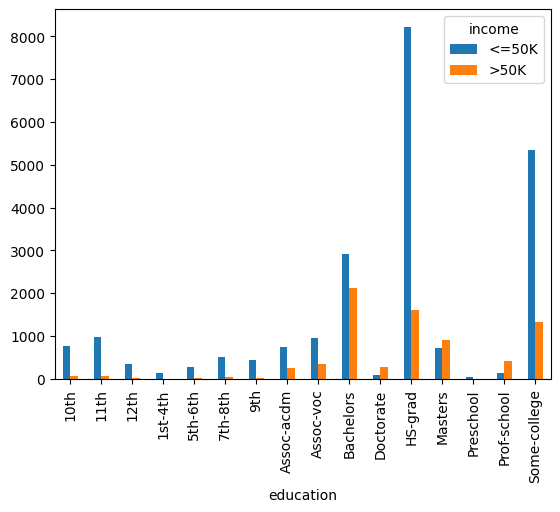
Data quality is paramount and starts with data preprocessing. In this project, we take care to clean the data carefully, including removing missing values ​​and removing redundant columns. This crucial step is critical to structuring our dataset to meet the requirements of machine learning algorithms and to ensure the strength of our analysis.

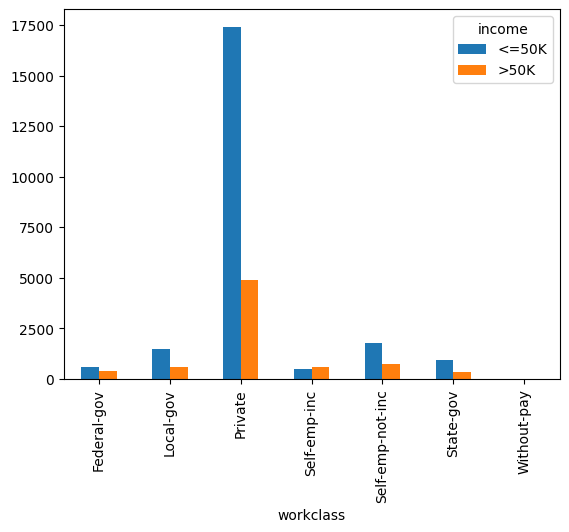
Data-Set

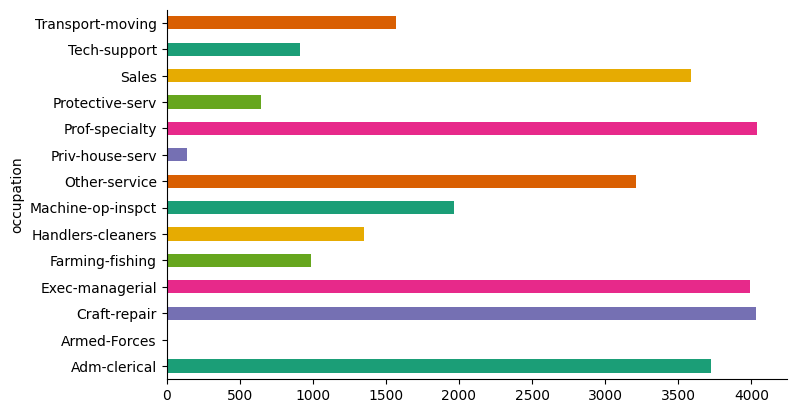


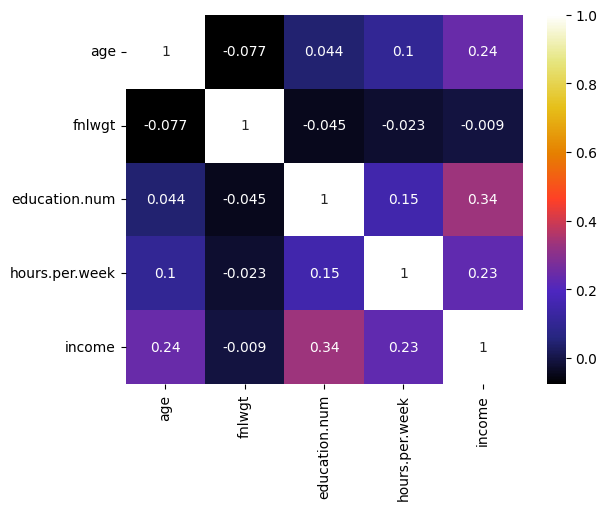
Data Visualization

Visualizations serve as powerful tools to reveal the complex relationships between demographic and economic factors and their impact on income levels. Our project uses data visualization techniques that provide a clear view of the role of these attributes in classifying individuals into income categories ("<50K" or ">50K"). These visualizations improve our understanding of the complex interplay between variables.









Methodology

**Logistic Regression:**

Logistic regression is a basic and interpretable algorithm that we use to categorize people into income levels. It predicts the probability that an instance belongs to a certain class. In our case, it estimates the probability that an individual will have an income above a certain threshold. During model training, we leveraged its predictive capabilities to produce accurate income predictions for each individual in our data. This approach allows us to provide not just a binary classification of "above the threshold" or "below the threshold", but rather a precise estimate of income that may have significant real-world applications.

**Perceptron:**

The perceptron is a basic but necessary model in binary classification tasks. By training this model on our data set, we took advantage of its ability to make binary predictions and distinguish between individuals with income below and above a certain threshold. Perceptron's simplicity and efficiency make it a valuable addition to our classification toolkit.

**Support Vector Machine (SVM):**

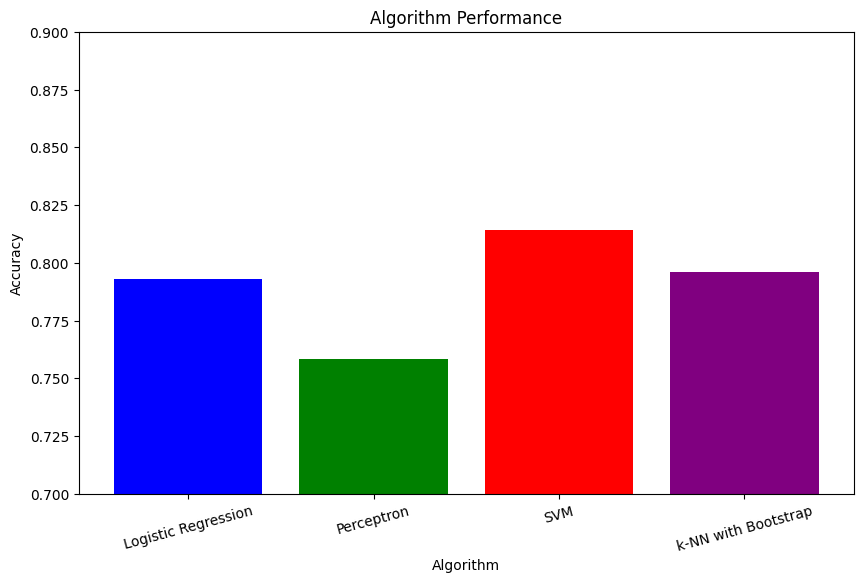
The Support Vector Machine is a robust model that excels at separating data points into distinct classes and is particularly valuable for income classification. During model training, the SVM learned complex decision boundaries, allowing it to distinguish between individuals with income levels below and above a specified threshold. Its ability to capture complex relationships between elements contributes to the accuracy of model classification.

**k-Nearest Neighbors (k-NN) with Bootstrap:**

We used the k-Nearest Neighbors (k-NN) algorithm with Bootstrap resampling to evaluate how our models perform under different conditions. By resampling the training data several times and training the k-NN classifiers, we obtained an overview of the model's stability and generalization ability. The mean accuracy across resampled models provides a robust estimate of k-NN classifier performance.

Outputs:

* Logistic Regression Accuracy: 0.7932634929054502
* Perceptron Accuracy: 0.7582548733589709
* SVM Accuracy: 0.8140830128630155
* k-NN with Bootstrap Mean Accuracy: 0.7962538124917125



Conclusion

In conclusion, we can say that our project is embarking on a journey into the field of income classification through the lens of machine learning and artificial intelligence. A diverse set of methodologies and rigorous evaluations highlight the potential of machine learning to produce accurate revenue forecasts. The implications of this project extend to real-world applications in areas such as financial decision-making, credit rating, and targeted marketing.

Links & References:

* Dataset: <https://www.kaggle.com/datasets/uciml/adult-census-income/data>
* Github link: <https://github.com/2203A52065/Project-Stat-ML>
* <https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/>
* <https://towardsdatascience.com/perceptron-learning-algorithm-d5db0deab975>