Data Mining and Discovery

Introduction:

Data mining and discovery is a larger domain to extract requisite patterns and information from a substantial dataset [1]. Within lying Census Income dataset (available by UCI Machine Learning Repository) [2], which is a meaningful way to simplify not only socioeconomic patterns from a larger dataset but also demonstrates demographics features in detail pertaining to income levels of that data [3]. In addition to Census Income dataset, we explore deep studies of Python language and analyzed statistical configurations, correlations and displaying of data. Such a vast exploration underpins the clarity of patterns in enhancing knowledge of Census Income data and in incorporating other languages. The following sections encompass the pre-processing of data, analysis of regression and classification of data by using other language such as Machine Learning.

Data Analysis:

For the analysis of Census Income dataset, descriptive statistics, correlation analysis, and visualization techniques were employed. The histplot() method illustrated an overview of descriptive histogram of the feature (Figure 1-a). Correlation analysis underwent corr() method to signify the relationships between numeric variables (Figure 2-b).

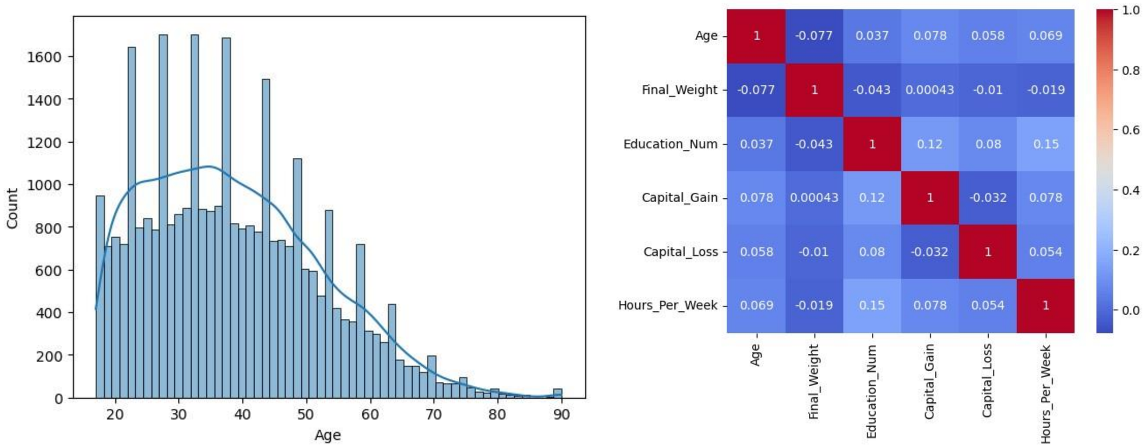

Figure 1: Histogram of the Age Feature (a) and Correlation Matrix of Dataset (b)

Data Preprocessing:

A comprehensive data preprocessing was pre-requisite before employing Machine Learning models in a dataset. This included meticulous examination of missing values using isnull(). To combat variations in regression, one-hot encoding method helped employed to transform categorical features in numerical values. Furthermore, handling outliers and normalization, the other steps of data pre-processxing, were entered to split data subsequently into training and testing sets in 80:20. During preprocessing phase for machine learning formatting, categorical features were converted into numeric format with the assistance of label encoder and one-hot encoder.

Linear Regression:

To predict income levels in the Census Income dataset, Linear Regression model was employed. With the help of evaluation metrics to evaluate model's performance, a scale that monitored 5.58 Mean Squared Error (MSE) was employed with the R-squared (R^2) value of 0.11 entered. A lower value indicates more accuracy. The MSE identifies the average squared difference between predicted and actual values. With an R^2 value of 0.11, the model appears to be able to explain a moderate amount of the data's variability. Various visualization techniques were used to subsequent results interpretation and pattern identification in model's performance. To meet the purpose, Scatter plot with Regression Line and a Residual Plot method were applied and results were interpreted accordingly (Figure 2).

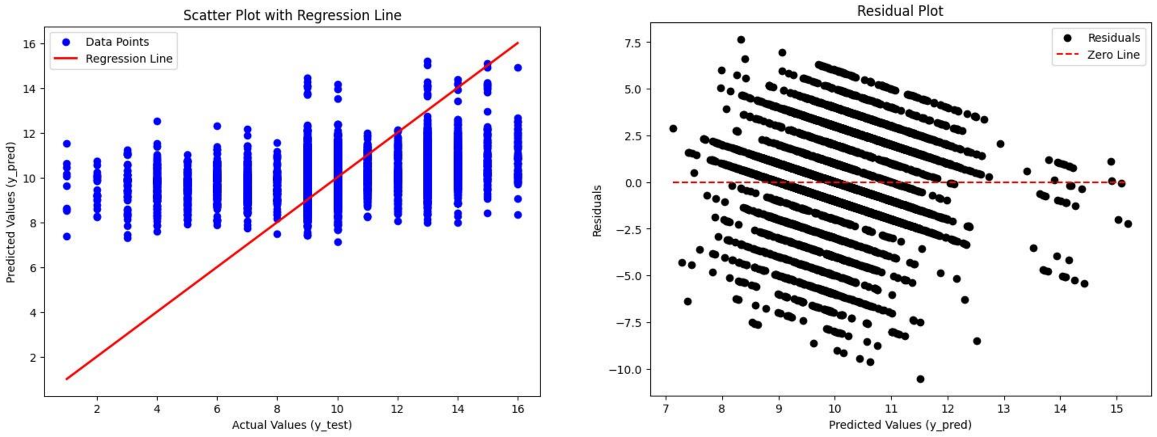

Figure 2: Regression Line and Residual Plot

Classification:

An effective neural network model that has the capacity to meet classification task is MLP Classifier (Multilayer Perception Classifier), hence was applied in this model pertaining to demand. As aforementioned, the assistance of label encoder and one-hot encoder in regression were employed to transform categorical variables into numerical counterparts. The conversion format was made applicable for classification task for working of MLP classifier. A hidden layer design of (100, 50) was used to configure the model, implying that there were 100 neurons in the first hidden layer and 50 in the second. For reproducibility, the random state was retained at 42 and the max_iter parameter was set to 500. Using the processed dataset as training data, the model achieved an accuracy of 70.29% (Figure 3).

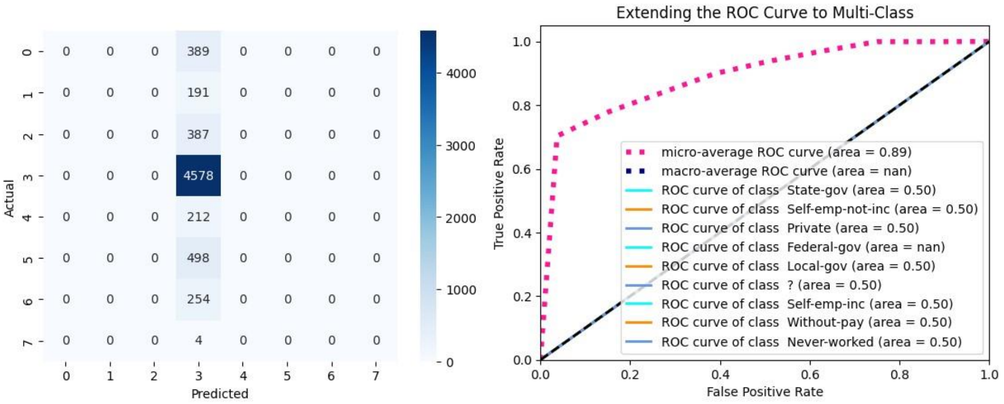

Figure 3: Classification Prediction and ROC Curve

Conclusion:

To sum up our study, it is apparent that Census Income dataset holds an importance in simplifying the socio-economic patterns from a large dataset. Revisiting classification problem, the prediction of Work-class was made across nine other classes with an accuracy of 70.29% using the MLP classifier that illuminated the complex patterns and defined several professional categories, such as "State-gov," "Private," and "Self-emp-inc.". Together, in the light of regression issue, the results of our focus on predicted “Age” variable via linear regression scored a mean squared error of 5.85, keeping R-squared value, 0.11. Such comprehensive analysis of both classes through liner regression underpins insights of income levels and contributes to the wider application of data mining techniques in simplifying the socio-economic intricate patterns.

References:

[1]. Rehman, A. U., Saleem, R. M., Shafi, Z., Imran, M., Pradhan, M., & Alzoubi, H. M. (2022, February). Analysis of Income on the Basis of Occupation using Data Mining. In 2022 International Conference on Business Analytics for Technology and Security (ICBATS) (pp. 1-4). IEEE.

[2]. Kohavi, R. (1996). Census Income. UCI Machine Learning Repository. (https://doi.org/10.24432/C5GP7S)

[3]. Gomez-Cravioto, D. A., Diaz-Ramos, R. E., Hernandez-Gress, N., Preciado, J. L., & Ceballos, H. G. (2022). Supervised machine learning predictive analytics for alumni income. Journal of Big Data, 9(1), 11.