Ahliya Journal of Business Technology and MEAN Economies
[Published by Palestine Ahliya University]
Vol. 03 Issue 01 (2026)
ISSN: 3007-9691

The Impact of Advanced Preprocessing Techniques on Machine Learning Models for Income Prediction

Mutaz Rasmi Abu Sara[1] and Andres Emaya[2]
[1] Faculty of Engineering and Information Technology, Palestine Ahliya University (Palestine)
✉ moutaz.a@paluniv.edu.ps
[2] Palestine Ahliya University (Palestine)
✉ Andreselias262@gmail.com

Received:14/04/2026 Accepted:23/05/2026 Published:31/05/2026

Abstract:

Data preprocessing plays a fundamental role in improving the reliability and predictive performance of machine learning models, particularly when dealing with real-world tabular datasets containing missing values, outliers, redundant features, skewed distributions, and heterogeneous data types. This study investigates the impact of advanced preprocessing techniques on income prediction using the Adult Income dataset. A comprehensive preprocessing pipeline was developed by integrating missing value imputation, variance-based feature selection, IQR-based outlier treatment, Yeo–Johnson transformation, standard scaling, one-hot encoding, Singular Value Decomposition (SVD), and a Column Transformer-based workflow to eliminate data leakage. Four machine learning models—Histogram-Based Gradient Boosting (HistGB), Random Forest, Logistic Regression, and Linear Support Vector Classifier—were trained and evaluated using stratified k-fold cross-validation and an 80/20 train-test split. Performance was assessed using Accuracy, Precision, Recall, F1-score, ROC-AUC, and Log-loss. The results demonstrate that the proposed preprocessing pipeline consistently improved the performance of all models, with Histogram-Based Gradient Boosting achieving the highest test accuracy of 86.8% and a ROC-AUC of 92.1%, indicating excellent predictive capability and strong generalization with minimal overfitting. The originality of this study lies in the development of a unified and reproducible preprocessing framework that systematically integrates multiple advanced preprocessing techniques and applies them consistently across different machine learning models, enabling a fair comparative evaluation. Unlike previous studies that primarily emphasize algorithm selection, this research demonstrates that a carefully designed preprocessing pipeline can substantially enhance predictive performance and produce competitive results without relying on complex ensemble architectures or extensive model tuning.

Keywords: *Data Preprocessing; Machine Learning Models; Adult-Income Dataset; Feature Engineering; Histogram-Based Gradient Boosting; Binary Classification.*

Ahliya Journal of Business Technology and MEAN Economies
[Published by Palestine Ahliya University]
Vol. 03 Issue 01 (2026)
ISSN: 3007-9691

تأثير تقنيات المعالجة المسبقة المتقدمة على نماذج التعلم الآلي للتنبؤ بالدخل

معتز رسمي أبو سارة[1]، أندريس اعميه[2]

[1] كلية الهندسة وتكنولوجيا المعلومات، جامعة أهلية فلسطين (فلسطين)

moutaz.a@paluniv.edu.ps

[2] جامعة أهلية فلسطين (فلسطين)

Mohd.siwad25@gmail.com

تاريخ الاستلام:2026/04/14 تاريخ القبول:2026/05/23 تاريخ النشر:2026/05/31

ملخص:

هدفت هذه الدراسة إلى التعرف على أثر تقنيات المعالجة المسبقة المتقدمة للبيانات في تحسين أداء نماذج التعلم الآلي المستخدمة في التنبؤ بالدخل بالاعتماد على مجموعة بيانات الدخل. ولتحقيق ذلك، تم تطوير إطار متكامل للمعالجة المسبقة يعالج أبرز مشكلات البيانات الواقعية، مثل القيم المفقودة، والقيم المتطرفة، والخصائص غير المؤثرة، والتوزيعات غير المتوازنة، واختلاف أنواع البيانات. وشملت عملية المعالجة تعويض القيم المفقودة، واختيار الخصائص، ومعالجة القيم المتطرفة باستخدام المدى الربيعي، وتحويل البيانات لتقليل الانحراف، وتوحيد المقاييس، وترميز المتغيرات الفئوية، وتخفيض الأبعاد، مع دمج جميع الخطوات ضمن إطار موحد يضمن منع تسرب البيانات أثناء التدريب. كما جرى تقييم أربعة نماذج للتعلم الآلي باستخدام أسلوب التحقق المتقاطع وتقسيم البيانات إلى مجموعة للتدريب وأخرى للاختبار. واعتمد التقييم على عدد من مؤشرات الأداء، شملت الدقة، والاسترجاع، والدقة الإيجابية، والمتوسط التوافقي بين الدقة والاسترجاع، ومنحنى التمييز بين الفئات، وخسارة التنبؤ. وأظهرت النتائج تفوق نموذج التعزيز التدريجي المعتمد على المدرجات التكرارية، إذ حقق أعلى دقة بلغت 86.8%، وأفضل قدرة على التمييز بين الفئات، مما يعكس كفاءة عالية في التنبؤ وقدرة جيدة على التعميم مع انخفاض احتمالية فرط التخصيص. وتتمثل الأصالة العلمية لهذه الدراسة في تطوير إطار متكامل وقابل لإعادة التطبيق يجمع بين مجموعة من تقنيات المعالجة المسبقة المتقدمة وتطبيقها بصورة موحدة على جميع النماذج، بما يضمن عدالة المقارنة بينها، ويؤكد أن تحسين جودة البيانات قبل بناء النموذج يسهم بصورة جوهرية في رفع كفاءة التنبؤ وتحسين الأداء دون الحاجة إلى نماذج معقدة أو عمليات ضبط موسعة للمعلمات.

الكلمات المفتاحية: *معالجة البيانات المسبقة؛ نماذج التعلم الآلي؛ مجموعة بيانات دخل البالغين؛ هندسة الميزات؛ تعزيز التدرج القائم على المدرج التكراري؛ التصنيف الثنائي.*

1. Introduction

Machine learning models, especially ones that are built on real-world data, rely tremendously on input data. Real-world data is often not complete, It is heterogenous, messy, has heaps of missing data, outliers, and redundant features. Without an adequate and robust preprocessing pipeline, the models that are produced would have performance issues, which would lead to inaccurate conclusions when solving tasks and making predictions. This study focuses on a binary classification problem. It is derived from the 1994 U.S. census database (Becker & Kohavi, 1996). Using the models from this dataset provides adequate understanding of financial and economic stability. It provided more insight on whether people make more or less than 50k dollars per year. The usage of machine learning and preprocessing techniques provides a robust solution for the income inequality problem from the dataset. The dataset provided several preprocessing challenges, including missing values, skewed distributions for numerical data, class imbalances, and some outliers. A strong preprocessing pipeline was created using scikit-learn's pipeline and column transformer framework (Pedregosa et al., 2011), which ensured high-quality data, reliability, and better decision making. The best model clocked the highest accuracy at 86.8%, which beats other benchmarks that used substandard preprocessing techniques. A comprehensive analysis was made to compare between different models and techniques, which helps in focusing on the important areas of preprocessing. Eventually, this may lead to improve the livable wages based on the individual jobs in the market (Kuhn & Johnson, 2013).

1.2 Related work

Recent studies have been applying machine learning approaches to the Adult Income Dataset. For example, Islam et al. (2023) performed a comparative analysis of eleven machine learning models for annual income prediction using demographic features and achieved the highest accuracy of 87% using an XGBoost–ANN ensemble model, outperforming individual machine learning classifiers. Jo (2024) on the other hand, got accuracies ranging from 84 to 86% after utilizing several pre-processing techniques and feature engineering steps. Thapa (2023) found that tree-based models such as Random Forest or Gradient Boosting were found to have higher accuracies than liner models, which is true for this paper too. The proposed work in this paper achieved an accuracy of 86.8% with Histogram-Based Gradient Boosting, and a robust preprocessing pipeline. When compared to other literature, our approach achieved one of the highest results, only falling behind Chakrabarty & Biswas (2018), who achieved an accuracy of approximately 88%.

2. Methodology

2.1 The dataset

The adult income dataset consists of approximately 32500 observations, each presenting a person from the United States, and 14 features including both numerical and categorical inputs. The features are represented in the table below:

Table 1: Types of data for the adult-income dataset

Feature ID	Feature Name	Type of Data
1	Age	Continuous
2	Work class	Categorical
3	Fnlwg	Continuous
4	Education	Categorical
5	Education-num	Continuous
6	Marital-status	Categorical
7	Occupation	Categorical
8	Relationship	Categorical
9	Race	Categorical
10	Sex	Categorical
11	Capital-gain	Continuous
12	Capital-loss	Continuous
13	Hours-per-week	Continuous
14	Native-country	Categorical

The target variable is binary:
<= 50k: Income is less than or equals 50k dollars.
>50k: Income is more than 50k dollars.
Several columns have missing values denoted by "?", and outliers which need to be dealt with before modeling.

2.2 Missing value standardization and imputation

In the adult-income dataset, there were multiple missing values across different rows and columns. They came in multiple forms such as missing symbols "?" and empty strings. Instead of deleting them, an imputation technique for numerical and categorical data was implemented call SimpleImputer. This technique fills the missing numerical data using the strategy of "median" and fills the missing categorical data by using the "most frequent strategy." Most researchers implemented the deletion of missing data, which negatively impacted the model's accuracy.

2.3 Removal of redundant features and duplicates

Columns that contained a single-value should be removed. In this dataset one redundant feature was removed, which is the capital gain feature. It contains numerical values that had near zero variance. Keeping this feature negatively impacted the models' performances and made them give inaccurate reading for the data. After consulting experts, some variables pertaining to that feature were missing but imputed as "9999", which is why the column was removed. The following code was used for this task: ("var", VarianceThreshold (1e-6)), OneHotEncoder(handle_unknown= "ignore" , sparse_output =True)

2.4 Target variable cleaning and encoding

Missing target values were filtered out for this binary classification problem. Since the target is categorical, we encoded the classes numerically:
0 → Lower income (<=50k)
1 → Higher income (>50k)
The numeric encoding used above enables compatibility with the different machine learning models we created.
Target variable encoding → (0,1) → Compatible with Linear models, Decision trees, Neural Networks, and Distance-Based models.

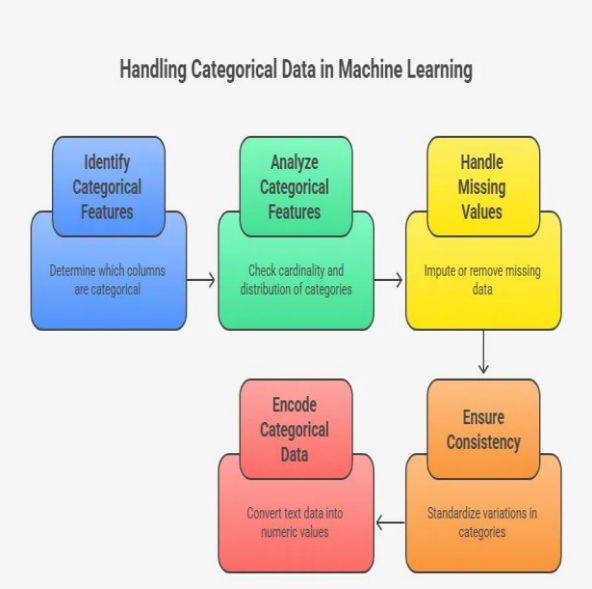

Figure 1: Handling categorical data using preprocessing techniques

2.5 Feature type identification

In the adult-income dataset, features are divided into two:
Numerical features: Mainly continuous data
Categorical features: Mainly nominal data
This separation allows us to do preprocessing techniques on numerical and categorical data separately. Later, the data is used in the ColumnTransformer() that passes each data type to its corresponding preprocessing technique.

2.6 Outlier consideration using IQR

Numerical features in the adult-income dataset were susceptible to extreme values. IQR (Inter Quartile Range) was utilized to clip the extreme values to boundary values. The range is: Q1 – 1.5 IQR, Q3 + 1.5 IQR. Any value outside the range would get a value of either Q1 – 1.5 IQR or Q3 + 1.5 IQR. This technique is used to reduce the influence of extreme values for all the machine learning models that were created.

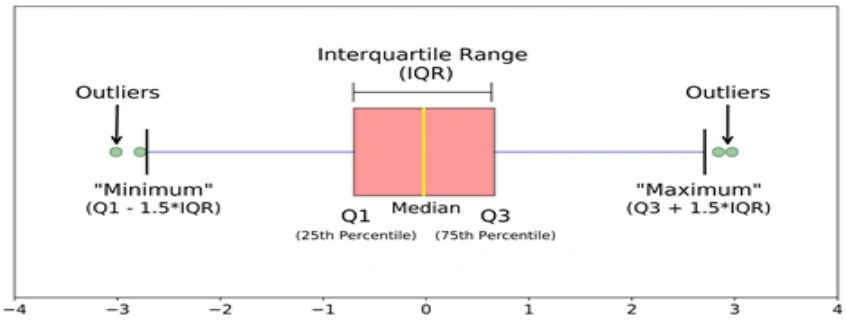

Figure 2: Dealing with outliers using IQR (Inter-Quartile Range method)

2.7 Preprocessing on different datatypes

The dataset is split into 80% train and 20% test. This proved to be more robust than using a 70% train and 30% test, as trial showed that models for this particular dataset perform better on the former split. For the numerical data, median imputation was used to replace missing values. Furthermore Yeo and Johnson (2000) as a power transform was utilized to reduce skewness in numerical inputs for developing better models. Additionally, standard scaling (Mean = 0, Sd = 1) was employed to ensure there are no bias and that all features contribute equally to the model.

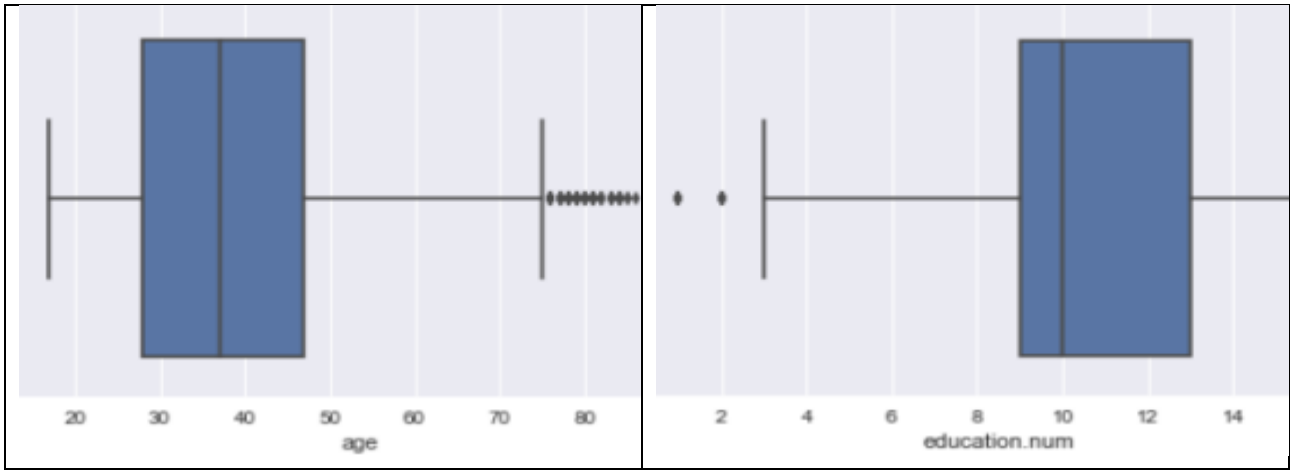

Figure 3: Numeric inputs with skewed distributions

As for the categorical features, the most-frequent imputation method was implemented to impute missing values according to the most frequent instance in the column. Additionally, one-hot encoding converted the categorical variables into binary, which is suitable for most machine-learning models. Furthermore, dimensionality reduction techniques are applied, such as SVD (Singular Value

Decomposition) to reduce the dimensionality of the dataset while still preserving the variance of the overall model. Finally, all the above-mentioned preprocessing steps are integrated into a pipeline using Column Transformer. It makes sure that no data leakage is present to reduce overfitting.

2.8 Modeling

Four machine learning models are evaluated after preprocessing: Logistic Regression, Linear Support Vector Classifier, Random Forest, and Histogram-Based Gradient Boosting. All these models used stratified k-fold cross validation.

3. Result

The results are measured using multiple metrics such as accuracy, recall, precision, F1 score, ROC-AUC, Log-loss. The results were as follows:

Table 2: Validation scores for different machine learning models after preprocessing

Model	CV Accuracy	CV F1	CV ROC-AUC
HistGB	0.8721	0.7135	0.9261
Logistic Regression	0.8511	0.6594	0.9063
Random Forest	0.8566	0.6791	0.9068
Linear SVC	0.8506	0.6518	0.9055

Table 3: Test scores for different machine learning models after preprocessing

Model	Test Accuracy	Test F1	Test Precision	Test Recall	Test ROC-AUC	Test Log Loss
HistGB	0.8680	0.7012	0.7712	0.6429	0.9206	0.2910
L.R	0.8502	0.6558	0.7344	0.5925	0.9004	0.3267
Random Forest	0.8496	0.6625	0.7209	0.6129	0.8946	0.3628
Linear SVC	0.8483	0.6453	0.7391	0.5727	0.8978	N/A

These tables summarize the performance of different machine learning models using all the above-mentioned preprocessing steps. Validation scores and test scores are measured to assess the generalization of the models and check for overfitting. As we can see from Figures 4 and 5, the models did not significantly perform better in the validation scores from the test scores, meaning that there was no overfitting.

Overall Performance Comparison

Histogram-based Gradient Boosting had the highest test accuracy with 86.8%.

Additionally, it had the highest ROC-AUC score of 92%. All the other models performed adequately with test scores ranging from 84.6% - 85.1% test accuracy. Overall, the model that achieved the best test score, F1 score, ROC-AUC score was the HB-Gradient Boosting. With its ability to capture non-linear relationships, Histogram-based Gradient boosting outperformed other models and being the best fit for the adult-income dataset.

Figure 4: Evaluation metrics for the machine learning models

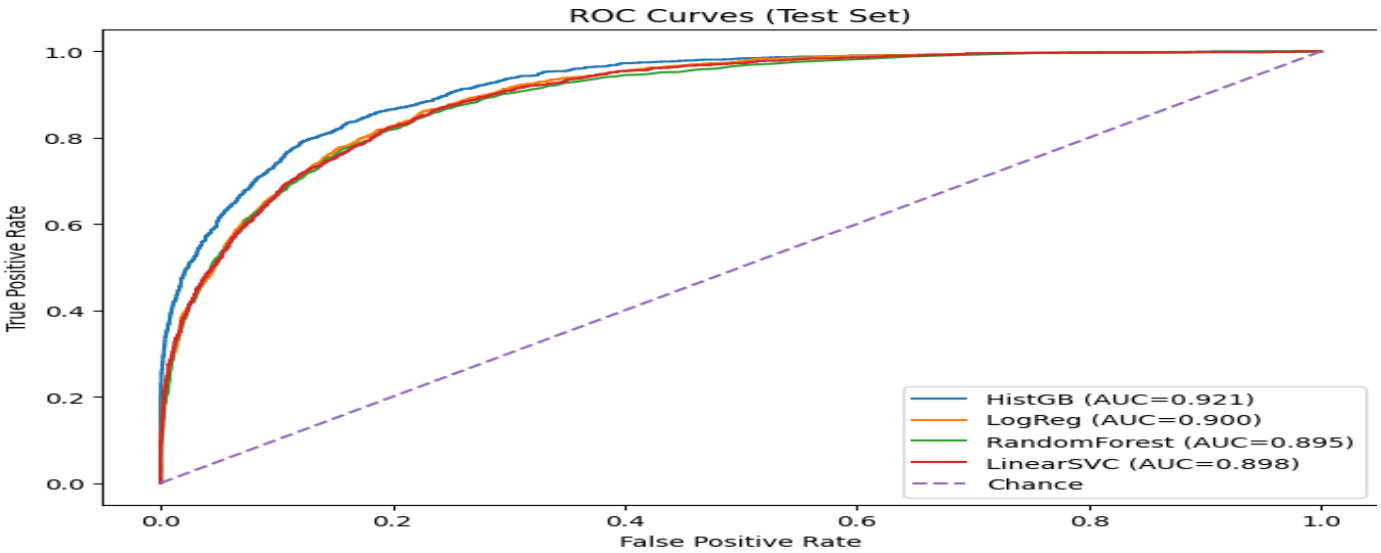

Figure 5: Evaluating machine learning models using ROC-AUC

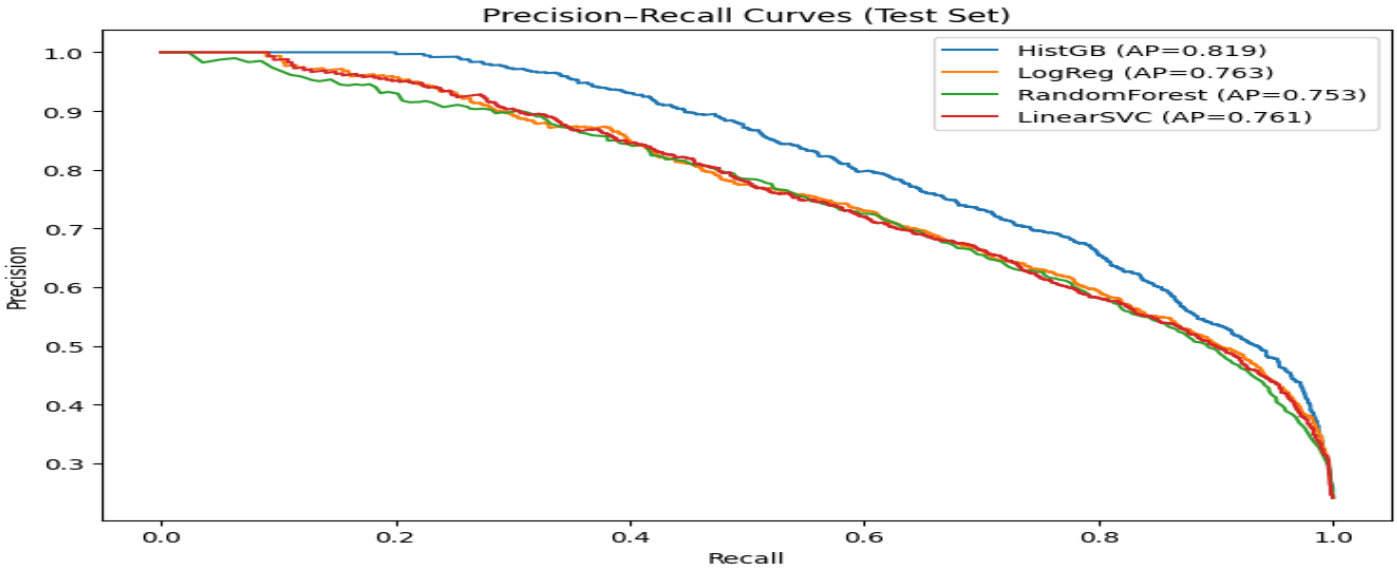

Figure 6: Plot of Precision and Recall for different machine learning models

According to Figures 6-8, the models are a good fit. The best model (HB-GB) achieved the highest score of 86.8%, which is a great benchmark compared to other models created by other authors with poor preprocessing steps.

4. Conclusion

This paper made a comparison between different preprocessing steps using different models. This was achieved by developing a robust pipeline with advance data preprocessing techniques and comparing the best results for each approach. The adult-income dataset is an example of real-world data that contains messy data, redundant features, missing values, and skewed distributions amongst different variables. The highest achieving model was the histogram-based Gradient Boosting, which nearly achieved the theoretical limit of the highest accuracy with a score of 86.8%. Comparing these results with other literature, we can see that creating a robust preprocessing pipeline gives quality results while making predictions, and for decision-making tasks. Future work may reinforce these findings with better fine-tuning and integration of domain-specific constraints.

References

Becker, B. & Kohavi, R. (1996). *Adult [Dataset]. UCI Machine Learning Repository*.

Chakrabarty, N., & Biswas, S. (2018, October). A statistical approach to adult census income level prediction. In *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)* (pp. 207-212). IEEE.

Islam, M. A., Nag, A., Roy, N., Dey, A. R., Fahim, S. F. A., & Ghosh, A. (2023, November). An investigation into the prediction of annual income levels through the utilization of demographic features employing the modified UCI adult dataset. In *2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)* (pp. 1080-1086). IEEE.

Jo, K. (2024). *Income prediction using machine learning techniques* [Master's thesis, University of California, Los Angeles]. eScholarship. https://escholarship.org/uc/item/6d01c9v7

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Thapa, S. (2023). Adult income prediction using various ML algorithms. *Available at SSRN 4325813*.

Yeo, I. K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*(4), 954-959.