Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Privacy-Aware Income Prediction Using Deep Neural Networks on the UCI Adult Dataset

Omar Nassim Adel Benyamina[1*], Zohra Slama[2]
[1,2]Department of Computer Science, Djillali Liabes University, Sidi bel abbes, Algeria
[1,2]EEDIS Laboratory, Sidi bel abbes, Algeria
** Corresponding Author: benyamina.adel98@gmail.com*

ARTICLE INFO

Received: 14 May 2025

Revised: 10 Jul 2025

Accepted: 20 Jul 2025

ABSTRACT

Income prediction is essential for applications in financial planning, credit scoring, and socioeconomic policy. In this study, we evaluate the effectiveness of deep learning neural networks—specifically Feedforward Neural Networks (FFN), Recurrent Neural Networks (RNN), and TabNet—for predicting income levels using the Adult dataset. We benchmark these models against traditional machine learning algorithms and investigate the impact of various preprocessing techniques, including missing value imputation, label encoding, categorical grouping, and normalization. Furthermore, we assess multiple feature selection methods—such as Chi-square, ANOVA, Random Forest importance, and L1 regularization—to determine their influence on predictive performance. Our experiments show that FFN achieves the highest baseline accuracy of 85.14% using optimized preprocessing, and further improves to 85.76\% when combined with feature selection techniques like Chi-square and ANOVA. These results confirm the advantage of deep learning approaches over classical models and underscore the value of well-designed data preprocessing pipelines. This study provides a comprehensive comparative analysis and highlights the importance of combining preprocessing with feature selection to optimize model accuracy in tabular data contexts. Future work will explore privacy-preserving training methods to enhance data protection when working with sensitive personal information.

Keywords: Income prediction, Deep learning, TabNet, Preprocessing, Feature selection, Data privacy.

INTRODUCTION

In recent years, deep neural networks have emerged as a powerful tool for various machine learning tasks, including income prediction. Unlike traditional machine learning models, deep learning algorithms can automatically learn hierarchical representations from raw data, potentially capturing complex patterns and relationships within the dataset. This capacity for automatic feature extraction has led to the growing popularity of deep learning techniques in numerous domains, including computer vision, natural language processing, and, more recently, income prediction.

Income prediction is a versatile and indispensable tool that influences decision-making processes in numerous domains ([1],[2],[3]). From the banking, sector, where it aids in credit assessment, to economic policy making, where it informs social welfare programs, and even in trading and marketing, where it guides investment choices and customer targeting, income prediction has a profound impact on our daily lives and the functioning of various industries. This interconnectedness of income prediction across different domains highlights its vital role in shaping decision-making processes and underscores the need for careful handling of sensitive data, particularly in research involving the Adult dataset.

The type of data used in the adult dataset poses specific challenges that demand more economic and careful processing due to its sensitive nature. The dataset contains a collection of personal information, including age, education, occupation, and marital status, among others, making it imperative to handle this data with utmost privacy and security concerns in mind. As income is a sensitive attribute, ensuring data privacy becomes a paramount concern when conducting research on income prediction.

In recent years, machine learning algorithms [4] and deep learning neural networks [5] have shown promising results in income prediction tasks ([6],[7],[8],[9]). However, the selection of appropriate preprocessing techniques and feature selection methods can greatly impact the accuracy and efficiency of these models. Furthermore, the complexity of deep learning architectures demands careful hyperparameter tuning and regularization techniques to prevent overfitting.

Through our experiments and analysis, we aim to provide insights into the effectiveness of deep learning neural networks for income prediction, as well as the impact of preprocessing techniques and feature selection methods. The findings from this study can inform the development of accurate and efficient income prediction models and contribute to the growing field of predictive analytics in finance and socio-economic research. Our research delves into the realm of income prediction, using deep learning neural networks and traditional machine learning algorithms. We emphasize the importance of sensitive data handling in the adult dataset, discuss the impact of deep learning techniques on income prediction accuracy, and emphasize the significance of proper preprocessing and feature selection methods. Through our comprehensive investigation, we strive to provide a comprehensive analysis that advances the field of income prediction and data privacy protection.

DATASET

The Adult dataset [10], also known as the "Census Income" dataset, is a widely used benchmark dataset in machine learning and data mining research. It provides a comprehensive set of attributes related to individuals, allowing researchers to explore various aspects of socio-economic factors and predict income levels based on those attributes. Understanding the details and context of the Adult dataset is essential for conducting accurate income prediction using deep learning neural networks.

A- Dataset Description

- The Adult dataset contains anonymized information about individuals collected from the 1994 United States Census.
- It consists of around 48,000 instances and can be used for machine learning and income classification. The exact size of the dataset may vary slightly depending on the version and source of the dataset. The version we'll be using in this article has 32,561 instances.
- Each instance represents a person and is associated with attributes that provide information on the relationship, education, profession, and other information.
- This information is represented by 14 attributes of two types (9 categorical and 6 integers).

- The target variable in the dataset is the income level, which is classified as either "$<=$50K" (indicating an income less than or equal to $50,000 per year) or "$>$50K" (indicating an income greater than $50,000 per year). The prediction task is to determine whether a person makes over 50K a year.

TABLE 1. LISTING OF ATTRIBUTES

Income	<=50K, >50K
Age	Continuous
Workclass	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
Fnlwgt	Continuous
Education	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
Education-num	Continuous
Marital-status	Marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
Occupation	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Armed-Forces, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Priv-house-serv, Transport-moving, Protective-serv.
Relationship	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
Race	Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black, White.
Sex	Female, Male.
Capital-gain	Continuous
Capital-loss	Continuous
Hours-per-week	Continuous
Native-country	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, South, China, Cuba, Iran, Outlying-US (Guam-USVI-etc), India, Japan, Greece, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad & Tobago, Peru, Hong, Holand-Netherlands.

Table 1 summarizes each attribute with its value. The longest categorical feature contains 41 different countries.

B- Sensitivity of Personal Financial Data

One crucial aspect of the Adult dataset is the inclusion of personal financial information, making it sensitive and subject to privacy concerns [15]. The dataset contains information related to an individual's income, capital gains, and financial contributions, which require careful handling to protect the privacy and confidentiality of the individuals involved.

Machine learning is playing a vital role in avoiding financial losses in the banking industry [16]. Respecting data privacy and ensuring the ethical use of personal information is crucial throughout the entire research process, from preprocessing to modeling and evaluation.

DEEP LEARNING FOR INCOME PREDICTION

Deep learning has emerged as a powerful technique for solving complex problems in various domains, including image recognition [17], natural language processing [18], and now, income prediction. This section explores the advantages of deep learning neural networks for predicting income levels and highlights their ability to learn intricate patterns from the Adult dataset.

A- Traditional Machine Learning Algorithms

Income prediction can be performed using various machine learning algorithms without relying on deep learning neural networks. Multiple traditional machine learning algorithms have been successfully applied to income prediction tasks as Logistic Regression [19], Random Forest [20], Gradient Boosting [21], Support Vector Machines (SVM) [22], and Naive Bayes. It is important to experiment with different algorithms and compare their performance to determine which method is best for income prediction via Adult dataset.

B- Advantages of Deep Neural Networks and Comparison with Traditional Machine Learning Models

- The Adult dataset contains anonymized information about individuals collected from the 1994 United States Census.
- Deep learning neural networks (DNNs) offer advantages such as non-linear modeling [26], automatic feature learning, and scalability. They can capture complex interactions and dependencies in the Adult dataset efficiently.
- DNNs automatically learn features from raw data [27], reducing the need for extensive feature engineering. Traditional ML algorithms require manual feature engineering, making DNNs more versatile for complex tasks.
- DNNs excel in handling large-scale datasets efficiently, accommodating diverse data types in the Adult dataset. Their complexity and scalability distinguish them from traditional ML algorithms.
- Training DNNs involves backpropagation and requires more computational resources and larger datasets compared to traditional ML algorithms. However, DNNs can handle noisy and incomplete data more effectively through techniques like dropout and regularization.

By leveraging the advantages of DNNs and understanding their differences from traditional ML algorithms, we can enhance income prediction accuracy and foster insightful decision-making in various domains.

METHODOLOGY

We aim to explore the application of deep learning neural networks for income prediction using the Adult dataset. We compare the performance of traditional machine learning algorithms [11], including logistic regression, random forest, gradient boosting, support vector machine, and naive Bayes, with deep learning neural networks, specifically feedforward networks, tabular neural networks, and recurrent neural networks. Furthermore, we investigate the impact of different preprocessing techniques on model performance [12]. These techniques include handling missing values, label encoding, grouping categories, and normalization. By applying these preprocessing techniques, we aim to enhance the quality of the dataset and improve the accuracy of income prediction models. Additionally, we explore various feature selection methods, such as correlation, chi-square [13], ANOVA [14], and feature importance from models, to evaluate their influence on model performance. Feature selection plays a crucial role in identifying the most relevant features for income prediction, allowing for more efficient and interpretable models.

A- Data Cleaning

- *Handle Missing Values:* The columns that carry these values are native-country, occupation, and workclass. In our case, the missing values are categorical and our aim is to preserve the data leakage, that is why we opted for approximation and not for deleting missing values. A form of imputation is applied [28]. It consists of replacing the missing values with the most frequent value. **Table 2** shows the number/percentage of missing values for the various attributes and their most frequent value, i.e. 2,399 records.

TABLE 2. MISSING VALUES IDENTIFICATION

Attribute	Missing Value	Pourcentage	MFV
Workclass	1836	5.64 %	Private
Occupation	1843	5.66 %	Prof-specialty
Native_country	583	1.79 %	United-States

MFV: Most frequent value

- *Handling Categorical Variables:* Label encoding assigns a unique numeric label to each category. **Table 3** represents the conversion of some categorical attributes.

TABLE 3. LABEL ENCODED

Feature	Value	Encoded Label
Workclass	Private	1
	Self-em-not-inc	2
	Self-emp-inc	3
	Federal-gov	4
	State-gov	5
	State-gov	6
	Without-pay	7
	Never-worked	8
Relationship	Wife	1
	Own-child	2
	Husband	3
	Not-in-family	4
	Other-relative	5
Sex	Female	1
	Male	2
Income	<=50K	0
	>=50K	1

- *Grouped Label:* After understanding the meaning of the different categories, we were able to group them into 2 or more. **Table 4** illustrates this method, applied only on two attributes.

Table 4. Label Encoded -Grouped value-

Feature	Group Value	Encoded Value
education	Preschool, 1st-4th, 5th-6th, 7th8th, 9th, 10th, 11th, 12th	0
	Hs-grad	1
	Assoc-voc, Assoc-acdm, Profschool, Some-college	2
	Bachelors	3
	Masters	4
	Doctorate	5
marital_status	Married-civ-spouse, Married-spouse-absent, Married-Af-spouse	0
	Divorced, Separated, Never-married, Widowed	1

B- Feature Engineering

As can be seen, there are 14 features to predict the income target and in order to improve the performance of machine learning algorithms and deep learning neural networks, we need to select the most relevant features from the dataset.

In this section, we focus on feature selection techniques for the Adult dataset.

1. ***Feature Selection:*** Aims to identify the subset of features that have the most predictive power while reducing dimensionality and eliminating irrelevant or redundant features. **Figure 1,2,3** show the results of Correlation analysis, Statistical Tests (Chi-square, ANOVA) respectively.

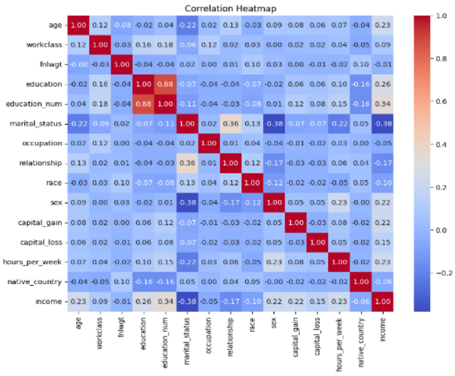

Figure 1. Correlation Matrix

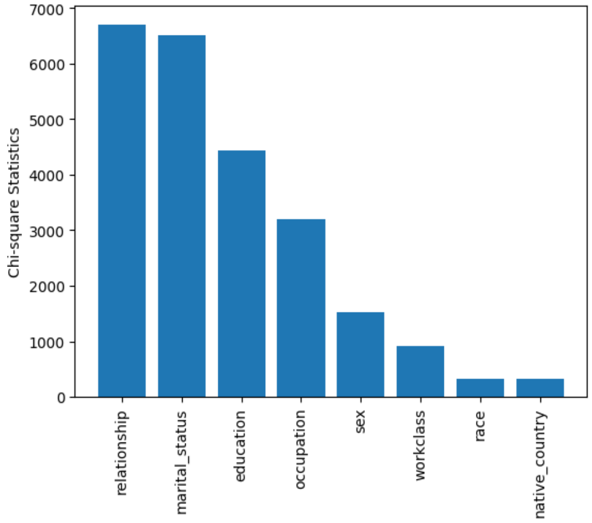

Figure 2. Chi-square Test Results for Categorical Features

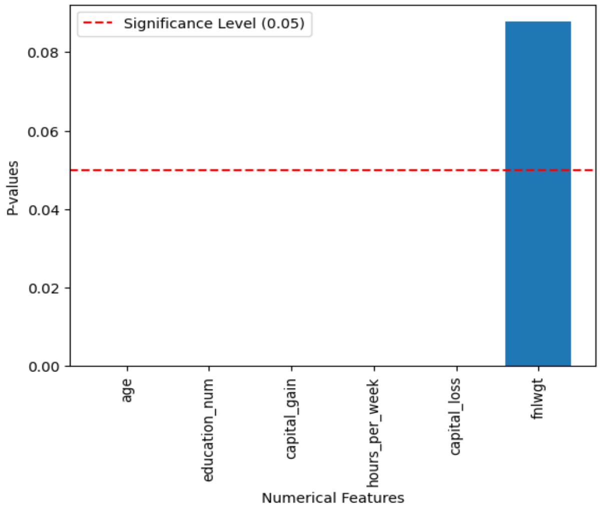

Figure 3. ANOVA Test Results for Numerical Features

2. Feature importance from models

- *Random Forest Feature Importance:* Random Forest models can estimate feature importance by calculating the average impurity reduction (or Gini importance) caused by each feature when it is used in different trees of the forest [23]. Features that lead to a higher impurity reduction are considered more important as shown in **Figure 4**.

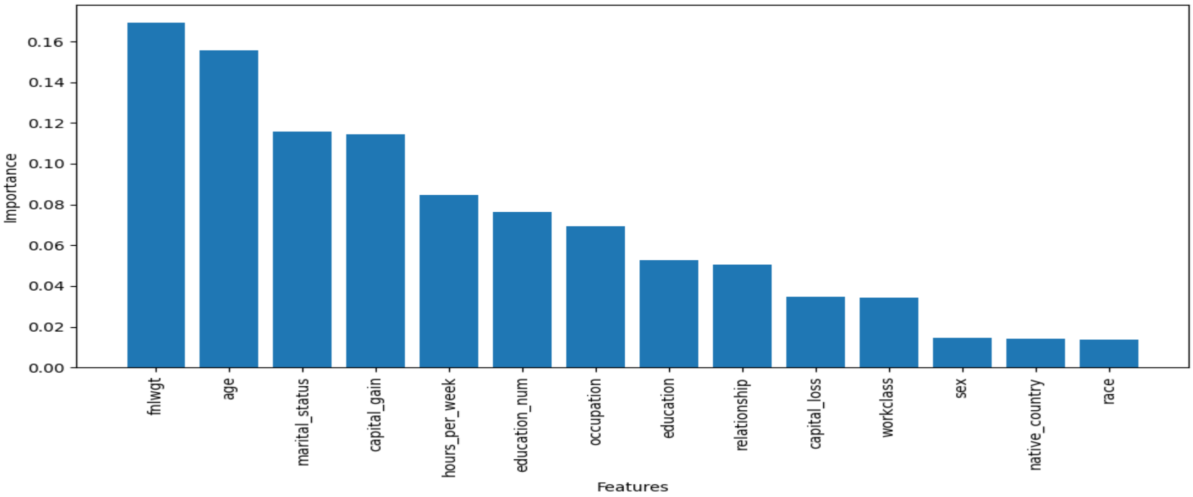

Figure 4. Random Forest Feature Importance

- *Gradient Boosting Feature Importance:* Similar to Random Forest, Gradient Boosting models can also estimate feature importance by measuring the improvement in the loss function caused by each feature during the boosting process [24]. Features that result in a larger decrease in the loss function are considered more important **Figure 5**.

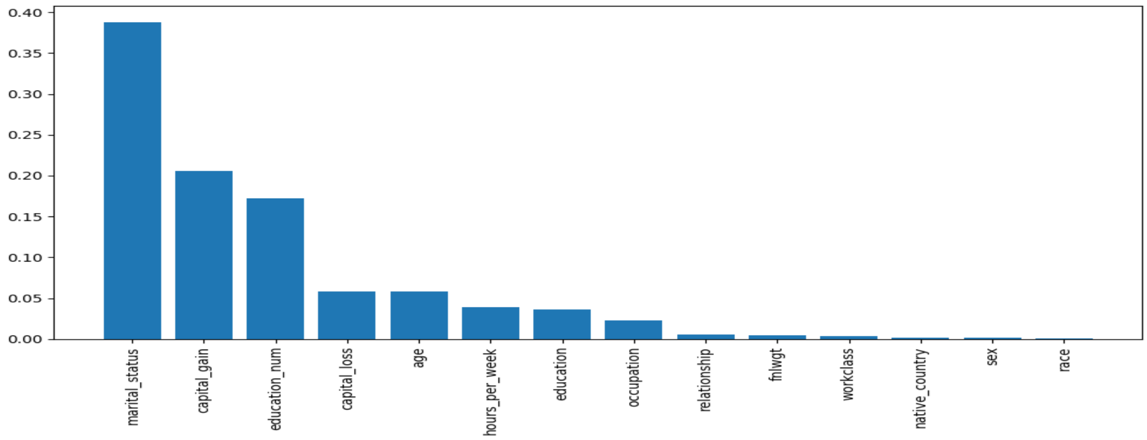

Figure 5. Gradient Boosting Feature Importance

- *Permutation Importance:* is a model-agnostic method that measures the impact of permuting (randomly shuffling) the values of a feature on the model's performance [25]. By permuting a feature and observing the drop in performance, we can quantify the importance of that feature in **Figure 6**.

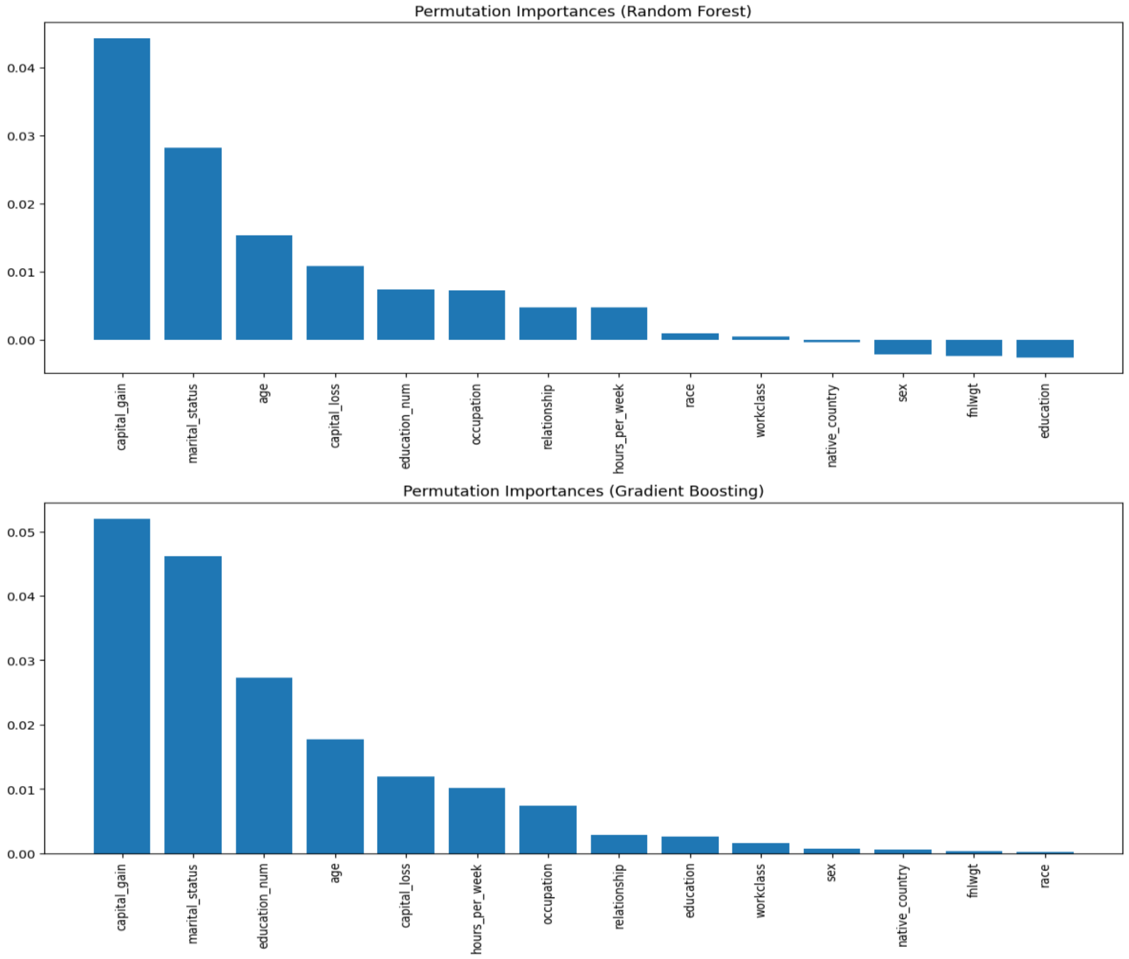

Figure 6. Permutation Importance

- *L1 Regularization (Lasso):* We apply L1 regularization, also known as Lasso regularization, to penalize less important features. By adding a penalty term to the loss function based on the absolute values of the feature coefficients. As a result, L1 regularization encourages sparsity in the feature coefficients, effectively selecting the most relevant features while setting the coefficients of irrelevant features to zero. **Figure 7** illustrates the chosen features with larger positive or negative coefficients, i.e. have a stronger influence on the target variable.

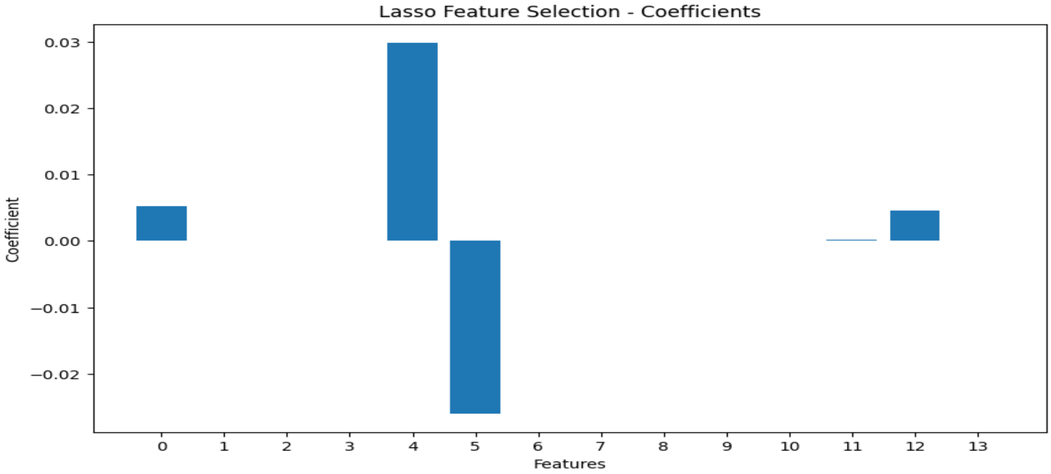

Figure 7. L1 Regularization

- *Recursive Feature Elimination:* Based on **Figure 8**, 10 features were assigned a rank of 1, considered important by the RFE algorithm. On the other hand, the features of education, race, sex, and native country were not selected, with ranks ranging from 2 to 5.

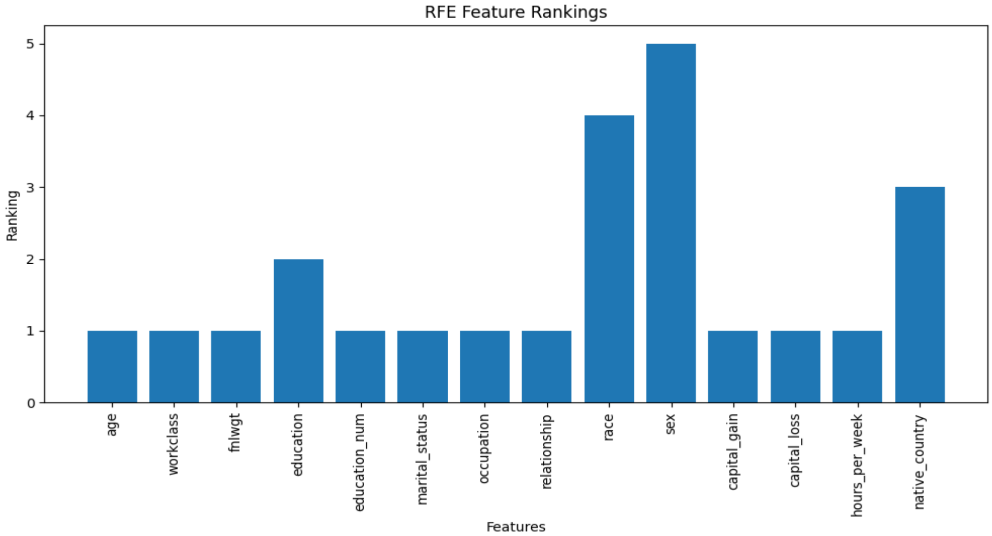

Figure 8. RFE(Decision)

- The goal of *RFECV* is to identify the optimal number of features that maximize the cross-validation score, and the selected features may vary slightly based on the specific data splits and model fitting during the cross-validation process. **Figure 9** shows that Random Forest selects more features (12) than Decision Tree (9), suggesting that Random Forest relies on a broader set of contributing features, while Decision Tree focuses on a smaller, more discriminative subset.

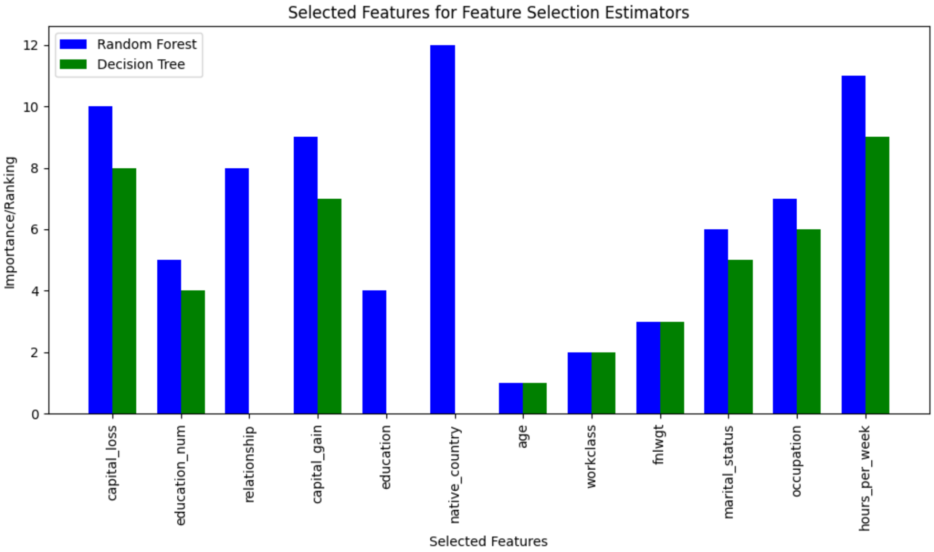

Figure 9. Selected Features using RFECV

C- Model Implementation and Experimental Setup

The experiments were conducted using Python on Google Colab, leveraging both Keras (for Feedforward Neural Network and Recurrent Neural Network) and PyTorch (via the pytorch tabnet library for TabNet). The computing environment was Google Colab with optional GPU acceleration enabled.

- **Data Splitting and Standardization:** For all models, the dataset was divided using a 70/30 train-test split implemented via train_test_split from Scikit-learn with a fixed random seed (random state=42) to ensure reproducibility. Feature scaling was applied using StandardScaler for Keras-based models. TabNet does not require external normalization, as it handles input scaling internally.

TABLE 5. SUMMARY OF IMPLEMENTATION PARAMETERS FOR EACH DL MODEL

Parameter	FFN & RNN	TabNet
Framework	Keras	PyTorch
Epochs	10	100
Batch Size	64	Default
Optimizer	Adam	Adam (default)
Early Stopping	No	Yes (patience=10)

Loss Function	Binary Crossentropy	Binary classification loss
Input Scaling	StandardScaler	Internal

These implementation details ensure consistent and reproducible experimentation across all models. No additional hyperparameter tuning was applied, as the focus was on evaluating the impact of preprocessing and feature selection strategies on model performance.

RESULTS AND EXPERIMENT

This section presents the experimental results comparing traditional machine learning (ML) models and deep learning (DL) neural networks for income prediction using the Adult dataset. The evaluation focuses on accuracy and mean squared error (MSE), as well as training and prediction times across various pre-processing and feature selection strategies.

A- Performance of Traditional Machine Learning Models

TABLE 6. MACHINE LEARNING ESTIMATORS/MODELS

ML models/algorithms	Accuracy	Mean Squared Error
Logistic Regression	0.786164	(0.213836)
Linear Discriminant Analysis	**0.822632**	(0.177368)
KNeighbors	0.760194	(0.239806)
Decision Tree	0.807272	(0.192728)
Gaussian Naive Bayes	0.783733	(0.216267)
Support Vector Machine	0.779313	(0.220687)
Linear SVC	0.776660	(0.223340)
SGD	0.777544	(0.222456)

Table 6 summarizes the performance of traditional ML algorithms, including Logistic Regression, Linear Discriminant Analysis, Decision Tree, and others. Among them, Linear Discriminant Analysis (LDA) achieved the highest accuracy of 82.26%, with a corresponding MSE of 0.177, outperforming other conventional models. Decision Tree followed closely with an accuracy of 80.72% and MSE of 0.193, indicating its robustness in handling categorical data.

B- Performance of Deep Learning Models with Preprocessing Techniques

The evaluation of deep learning architectures (FFN, TabNet, and RNN) is detailed in **Table 7.** Each model was trained using multiple combinations of preprocessing techniques, including: **M**: Imputation of missing values using the most frequent value, **D**: Deletion of missing values, **L**: Label encoding of categorical features, **G**: Grouping of specific categorical features, **N**: Normalization of numerical features.

TABLE 7. EVALUATION OF DNN PERFORMANCE WITH VARIOUS PREPROCESSING METHODS

DNN	Preprocessing					Accuracy	Train Time	Prediction Time
	M	D	L	G	N	(%)	(s)	(s)
FFN	*		*			78.59	8.70	0.84
	*		*	*		78.95	8.36	1.55
	*		*		*	**85.14**	9.57	0.70
	*		*	*	*	85.08	8.90	0.72
		*	*			77.14	5.74	0.95
		*	*	*		78.40	6.27	0.51
		*	*		*	**84.80**	21.70	0.52
		*	*	*	*	**84.72**	13.49	0.81
TabNet	*		*			**85.00**	48.86	52.86
	*		*	*		**85.07**	72.86	30.06
	*		*		*	**85.14**	43.52	30.18
	*		*	*	*	**85.14**	58.07	29.03
		*	*			**84.30**	50.36	18.35
		*	*	*		**84.30**	39.74	24.89
		*	*		*	84.71	63.63	21.81
		*	*	*	*	84.53	45.04	23.12
RNN	*		*			78.42	35.99	1.60
	*		*	*		31.38	42.81	1.64
	*		*		*	83.93	43.47	1.62
	*		*	*	*	84.15	36.07	1.61
		*	*			76.82	27.40	1.56
		*	*	*		78.06	28.34	1.56
		*	*		*	83.47	28.31	1.58
		*	*	*	*	83.38	31.79	1.59

Among all configurations, TabNet and FFN achieved the highest accuracy of 85.14% using the preprocessing combinations (M+L+N) and (M+L+G+N). RNN models also performed reasonably well, although slightly behind FFN and TabNet.

C- Impact of Feature Selection Techniques

To assess the benefit of feature selection, we applied several techniques—including Correlation, Chi-square, ANOVA, L1 Regularization, and Recursive Feature Elimination (RFE)—to the best-performing DL model (FFN with M+L+N). **Table 8** presents the accuracy obtained after removing less informative features, i.e. number of features removed (NFR).

TABLE 8. EVALUATION OF DNN PERFORMANCE WITH VARIOUS PREPROCESSING METHODS

Techniques	NFR	Accuracy (%)
Correlation	5	84.94 – 85.26
Chi-square	2	85.42 - 85.59
Anova	1	85.31
Chi-square + Anova	3	85.31 – 85.66
Random Forest (RF)	3	85.15 – 85.45
Gradient Boosting (GB)	6	85.52 – 85.76
Permutation Importance (RF)	6	85.63 – 85.72
Permutation Importance (GB)	4	85.26 – 85.83
L1 Regularization (Lasso)	9	83.17 – 83.51
RFE (Decision)	4	84.00 – 85.51
RFECV (Random Forest)	2	85.17 – 85.31
RFECV (Decision Tree)	5	85.18 – 85.48

D- Summary and Comparison

Deep learning models significantly outperform traditional ML approaches for income prediction using the Adult dataset. While the best-performing traditional ML model **Table 6**, Linear Discriminant Analysis, achieves an accuracy of 82.26%, DL models, particularly TabNet and FFN **Table 7**, improve this to 85.14% with optimal preprocessing techniques (M+L+N/G). Furthermore, applying feature selection further refines the model **Table 8**, boosting accuracy to 85.76%, demonstrating its effectiveness in enhancing predictive performance. This comparison highlights the superiority of DL over traditional ML and the added value of feature selection in optimizing model accuracy.

This comparison highlights the superiority of deep learning over traditional machine learning methods, and the added value of feature selection in optimizing model accuracy while maintaining interpretability and computational efficiency.

CONCLUSION

This study investigated the application of deep learning models for income prediction using the Adult dataset, with a strong emphasis on preprocessing and feature selection. We evaluated multiple deep learning architectures—including Feedforward Neural Network (FFN), Recurrent Neural Network (RNN), and Tabular Neural Network (TabNet)—against traditional machine learning algorithms. Our findings demonstrate that deep learning models, particularly FFN and TabNet, significantly outperform classical approaches, with FFN achieving the highest accuracy of 85.14% when combined with the optimal preprocessing pipeline (missing value imputation, label encoding, and normalization).

We also examined the impact of feature selection techniques, showing that methods such as Chi-square combined with ANOVA or Gradient Boosting-based importance can further enhance model performance, pushing accuracy to 85.76%. These results highlight the effectiveness of careful preprocessing and targeted feature reduction in improving deep learning outcomes for structured tabular data.

Overall, our contributions include a comprehensive comparative analysis of DL and ML models, an empirical evaluation of preprocessing strategies, and a systematic study of feature selection impacts. While promising, the models presented here still depend on sensitive demographic attributes, and further research is required to improve their privacy-awareness, generalizability, and fairness.

In future work, we aim to incorporate differential privacy mechanisms and fairness-aware learning algorithms to ensure that income prediction models are not only accurate but also ethically and legally responsible in real-world deployments.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFRENCES

[1] Z. Dai, Z. Yuchen, A. Li and G. Qian, "The application of machine learning in bank credit rating prediction and risk assessment," 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Nanchang, China, 2021, pp. 986-989, doi: 10.1109/ICBAIE52039.2021.9389901.

[2] Andrea R. V., Sergio M. L., Sandra C., Guillermo T., Silvia R., Ignacio S., Yuldor Caballero-Díaz, Luis J. H., John M. G., Leonardo S., Rachid L., Giancarlo B., Fernando H., Martha V. F., Elkin O., Diana S. R., and Eduardo B., Prediction of SARS-CoV-2 infection with a

Symptoms-Based model to aid public health decision making in Latin America and other low and middle income settings, Preventive Medicine Reports, Volume 27, 2022, 101798, ISSN 2211-3355, https://doi.org/10.1016/j.pmedr.2022.101798.

[3] Rahmat, R. F., Syaputra, A. H., Faza, S., and Arisandi, D. (2020, May). Prediction of Regional Revenue and Expenditure Budget using Autoregressive Integrated Moving Average (ARIMA). In IOP Conference Series: Materials Science and Engineering (Vol. 851, No. 1, p. 012064). IOP Publishing.

[4] Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI. 2, 160 (2021). https://doi.org/10.1007/s42979-021-00592-x

[5] Wang, H., and Raj, B. (2017). On the origin of deep learning. arXiv preprint arXiv:1702.07800.

[6] N. Chakrabarty, and S. Biswas, "A Statistical Approach to Adult Census Income Level Prediction," 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 2018, pp. 207-212, doi: 10.1109/ICACCCN.2018.8748528.

[7] J. Vemulapati, A. Bayyana, S. H. Bathula, S. Tokala, K. Hajarathaiah, and M. K. Enduri, "Empirical Analysis of Income Prediction Using Deep Learning Techniques," 2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, India, 2023, pp. 1-6, doi: 10.1109/SCEECS57921.2023.10062992.

[8] Sundsøy, P., Bjelland, J., Reme, B. A., Iqbal, A. M., and Jahani, E. (2016, January). Deep learning applied to mobile phone data for individual income classification. In 2016 International Conference on Artificial Intelligence: Technologies and Applications (pp. 96-99). Atlantis Press.

[9] Wang, Z., Sugaya, S., and Nguyen, D. P. (2019). Salary prediction using bidirectional-gru-cnn model. Assoc. Nat. Lang. Process.

[10] https://archive.ics.uci.edu/dataset/2/adult

[11] Ahsan MM, Mahmud MAP, Saha PK, Gupta KD, and Siddique Z. Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. Technologies. 2021; 9(3):52. https://doi.org/10.3390/technologies9030052

[12] Alshdaifat E, Alshdaifat D, Alsarhan A, Hussein F, and El-Salhi SMFS. The Effect of Preprocessing Techniques, Applied to Numeric Features, on Classification Algorithms' Performance. Data. 2021; 6(2):11. https://doi.org/10.3390/data6020011

[13] Peters, C. C., and Van Voorhis, W. R. (1940). Chi square.

[14] St, L., and Wold, S. (1989). Analysis of variance (ANOVA). Chemometrics and intelligent laboratory systems, 6(4), 259-272.

[15] Yasunori, F., Kiyoshi, M., Andrew, A. A., Yohko, O., and María, L. P. A. (2017). Personal data sensitivity in Japan: An exploratory study. The ORBIT Journal, 1(2), 1-13.

[16] Jillian M. Clements, Di Xu, Nooshin Yousefi, and Dmitry Efimov, 2020. "Sequential Deep Learning for Credit Risk Monitoring with Tabular Financial Data," Papers 2012.15330, arXiv.org.

[17] Pak, M., and Kim, S. (2017, August). A review of deep learning in image recognition. In 2017 4th international conference on computer applications and information processing technology (CAIPT) (pp. 1-3). IEEE.

[18] Otter, D. W., Medina, J. R., and Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. IEEE transactions on neural networks and learning systems, 32(2), 604-624.

[19] M. A. Sheikh, A. K. Goel, and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 490-494, doi: 10.1109/ICESC48915.2020.9155614

[20] Eichinger, F., and Mayer, M. (2022). Predicting Salaries with Random-Forest Regression. In Machine Learning and Data Analytics for Solving Business Problems: Methods, Applications, and Case Studies (pp. 1-21). Cham: Springer International Publishing.

[21] Wisesa, O., Adriansyah, A., and Khalaf, O. I. (2020, September). Prediction analysis sales for corporate services telecommunications company using gradient boost algorithm. In 2020 2nd International Conference on Broadband Communications, Wireless Sensors and Powering (BCWSP) (pp. 101-106). IEEE.

[22] Baranes, A., and Palas, R. (2019). Earning movement prediction using machine learning-support vector machines (SVM). Journal of Management Information and Decision Sciences, 22(2), 36-53.

[23] Yifan Zhao, Weiwei Zhu, Panpan Wei, Peng Fang, Xiwang Zhang, Nana Yan, Wenjun Liu, Hao Zhao, and Qirui Wu, (2022). Classification of Zambian grasslands using random forest feature importance selection during the optimal phenological period. Ecological Indicators, 135, 108529.

[24] Adler, A. I., and Painsky, A. (2022). Feature importance in gradient boosting trees with cross-validation feature selection. Entropy, 24(5), 687.

[25] Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. Bioinformatics, 26(10), 1340-1347.

[26] Almeida, J. S. (2002). Predictive non-linear modeling of complex data by artificial neural networks. Current opinion in biotechnology, 13(1), 72-76.

[27] Chen, X., Xu, Y., Yan, S., Wong, D. W. K., Wong, T. Y., and Liu, J. (2015). Automatic feature learning for glaucoma detection based on deep learning. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18 (pp. 669-677). Springer International Publishing.

[28] Osborne, J. W. (2013). Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data. Thousand Oaks, CA: Sage Publications.