Assignment title	**: *Machine Learning Tutorial***
Module Code	**: 7PAM2021-0901**
Module Title	**: Machine Learning and Neural Networks**
Student ID	**: 24099849**
Student Name	**: Shruthi Chintalapudi**
Module Leader	**: Peter Scicluna**
Word Count	**: 1997 Words**
Submission Date	**: 11 Dec 25**

Table of Contents

1. Introduction

Machine learning is when a computer system picks up on patterns and makes predictions from a set of data, without being explicitly programmed to do so. In supervised learning, classification algorithms are an important part that helps assign input data to a predefined category. That is whether or not a customer will purchase the product. Logistic Regression stands out as an easy yet stronger learning algorithm for the classification task among these algorithms. It uses a sigmoid function to predict the likelihood that an observation belongs to a particular class. Thus, it is used for problems that are linearly separable. (Nokeri, 2021).

Recent studies have shown that logistic regression is popular in classification which is easy to use and robust. As an example, Azis (2024) found that logistic regression classified heart disease with 80–88% accuracy, providing stable results in various data partitions. Masseran (2024) also argued that the method is very effective for the classification of environmental data, for instance, with respect to air pollution events. In the finance sector, Tynymbayev et al. (2024) used logistic regression to detect fraud. Their two main advantages were transparency and the capability of binary decision. Caba (2023) showed that logistic regression may not perform well on simple or complicated high-dimensional data; nevertheless, combining with deep learning can help a lot. In the end, Widhianingsih et al. (2020) develop ensemble extensions that keep the interpretation of logistic regression and alleviate overfitting from large-scale data.

In this tutorial, we will use the Social Network Ads dataset available on Kaggle to model and predict purchasing behavior based on Age, Gender and Estimated Salary. This work offers step-by-step educational guidance on logistic regression. It includes the mathematical formulation, implementation, model evaluation and visualization. It is designed to help the learners understand the concepts of this widely used classification technique.

Step 1: Importing Required Libraries

To get this project started, we must import a few Python libraries that we will need for data handling, visualization, and model training (Figure 1). The following libraries were used in JupyterLab:

```
# Step 1: Importing required libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, classification_report, accuracy_score, roc_curve, auc

# Optional: prettier plots
plt.style.use('ggplot')
```

Figure 1: Importing Libraires

Pandas and NumPy libraries complement each other which offer capabilities to manipulate structured data to provide high-performance data frames and data analysis (Diana & Mathivanan, 2023). With Matplotlib, different visualizations can be created for effective pattern representation. According to Nikandrov (2023), Scikit-learn has some amazing functionalities for data splitting, scaling feature and implementation of classification model like Logistic Regression. When combined, they help create the foundation of data science workflow in Python and provide reproducibility of model development (Chupilko et al., 2021). Including Umich in JupyterLab is an interactive and transparent computing environment commonly used in machine learning teaching and research (V. Bala, 2024).

Step 2: Loading and Inspecting the Dataset

The Social Network Ads dataset sourced from Kaggle is utilized in this project's dataset. It consists of 400 records and a total of 5 columns which are User ID, Gender, Age, Estimated Salary and Purchased (Figure 2 & 3). This dataset captures demographic and purchase information, making it useful for binary classification.

```
# Step 2: Load dataset
data = pd.read_csv('Social_Network_Ads.csv')

# Display first few rows
print("Dataset Preview:")
display(data.head())

# Show basic info
print("\nDataset Info:")
print(data.info())
```

Figure 2: Loading the Dataset

	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0

Figure 3: Dataset Preview

The Purchased column is the target variable, representing the binary outcome:

- 0 → Did not buy
- 1 → Purchased

Because the dependent variable has categorical nature, Logistic Regression is chosen as the suitable model for this problem. Using the Pandas library enables the importation and inspection of data, which helps facilitate exploratory data analysis (EDA). The incorporation of this library into the dataset remains compliant with proper use and validation (Sun et al., 2020).

Step 3: Data Preprocessing

Data preprocessing (Figure 4) is the first and very important step of machine learning. Data must be formatted, balanced, and normalized before it can be used for training.

Bad preprocessing can cause predictions to be biased and inconsistent. The Social Network Ads dataset has been preprocessed for logistic regression after undertaking several preprocessing steps (Mekni et al., 2025).

```
# Step 3: Data preprocessing

# Drop the 'User ID' column (not useful for prediction)
if 'User ID' in data.columns:
    data = data.drop('User ID', axis=1)

# Convert Gender to numeric (0 = Female, 1 = Male)
if 'Gender' in data.columns:
    data['Gender'] = data['Gender'].map({'Female': 0, 'Male': 1})
```

Figure 4: Data Preprocessing

3.1 Feature Selection

The User ID column was removed as it contains unique identifiers which are not useful for prediction. Eliminating some features helps make the computation more efficient and reduces the noise in the data.

3.2 Encoding Categorical Data

The Gender column was reformed into numbers, so it could be processed in the machine since algorithms like Logistic Regression can't read strings normally.

3.3 Feature Scaling

The Age and Estimated Salary variables were normalized by utilizing the StandardScaler() function. Scaling helps to ensure that every feature has an equal influence on model learning and prevents bias where the variables with large magnitude dominate the optimization (Figure 5). It is especially important for algorithms that are dependent on gradient-based optimization (Erol et al., 2022).

```
# Standardize features
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Figure 5: Feature Scaling

3.4 Train-Test Split

In the end, a fixed random state was utilized to split the dataset into 80% training and 20% testing subsets. This division allows for a robust assessment of how well a model generalizes (Rahmani et al. 2024).

By following these preprocessing steps, the features can scale within similarly, redundancy is reduced and learning stability is enhanced.

Step 4: Training the Logistic Regression Model

After preprocessing, the standardized data were used to train the Logistic Regression model (Figure 6):

```
# Step 4: Train the model
model = LogisticRegression(random_state=42)
model.fit(X_train_scaled, y_train)

# Predict test set results
y_pred = model.predict(X_test_scaled)
```

Figure 6: Training the Logistic Regression Model

Logistic regression is a supervised learning algorithm that predicts the chances of a binary outcome using the sigmoid function (Figure 7):

$$P(y=1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2)}}$$

Figure 7: Sigmoid function

This function takes inputs and puts them into a range of 0 and 1 for classification. If the prediction is more than 0.5 it gets classified as 1 (Purchased). Similarly, anything lower gets classified as 0 (Not Purchased).

Model training was conducted using a split of 80/20 to achieve generalization and avoid overfitting the model. Because Logistic Regression is interpretable with efficient performance, we leverage its strengths in educational tutorial and real-world classification task (Li et al., 2023). The model was trained successfully and ready to be evaluated on performance metrics accuracy, precision, recall, and F1-score.

Step 5: Model Evaluation

Model evaluation must be done in any supervised learning process as it evaluates the accuracy of the trained model on the independent data. For the Logistic Regression model, accuracy, precision, recall, F1-score and confusion matrix were all used as performance metrics to assess our model (Figure 8) .

```
accuracy = accuracy_score(y_test, y_pred)
print(f"Model Accuracy: {accuracy:.2f}\n")

print("Classification Report:")
print(classification_report(y_test, y_pred))
```

Figure 8: Model Evaluation

The following metrics summarize the model's performance:

- Accuracy: 0.89

- Precision: 0.91
- Recall: 0.75
- F1 Score: 0.82

The overall accuracy of the model is 89% hence the model is reliable based on the test dataset. With a precision score of 0.91, most predicted purchasers were true purchasers. With recall at 0.75, actual purchasers were not difficult to detect by the model. The F1 score is the harmonic mean of precision and recall. This gives equal weight to both precision and recall (Sitarz, 2022). It provides a robust single measure of performance.

The confusion matrix output (Table 1) shows the distribution of classifications by the model (Figure 9).

Table 1: confusion matrix output

	Predicted: 0	**Predicted: 1**
Actual: 0	50	2
Actual: 1	7	21

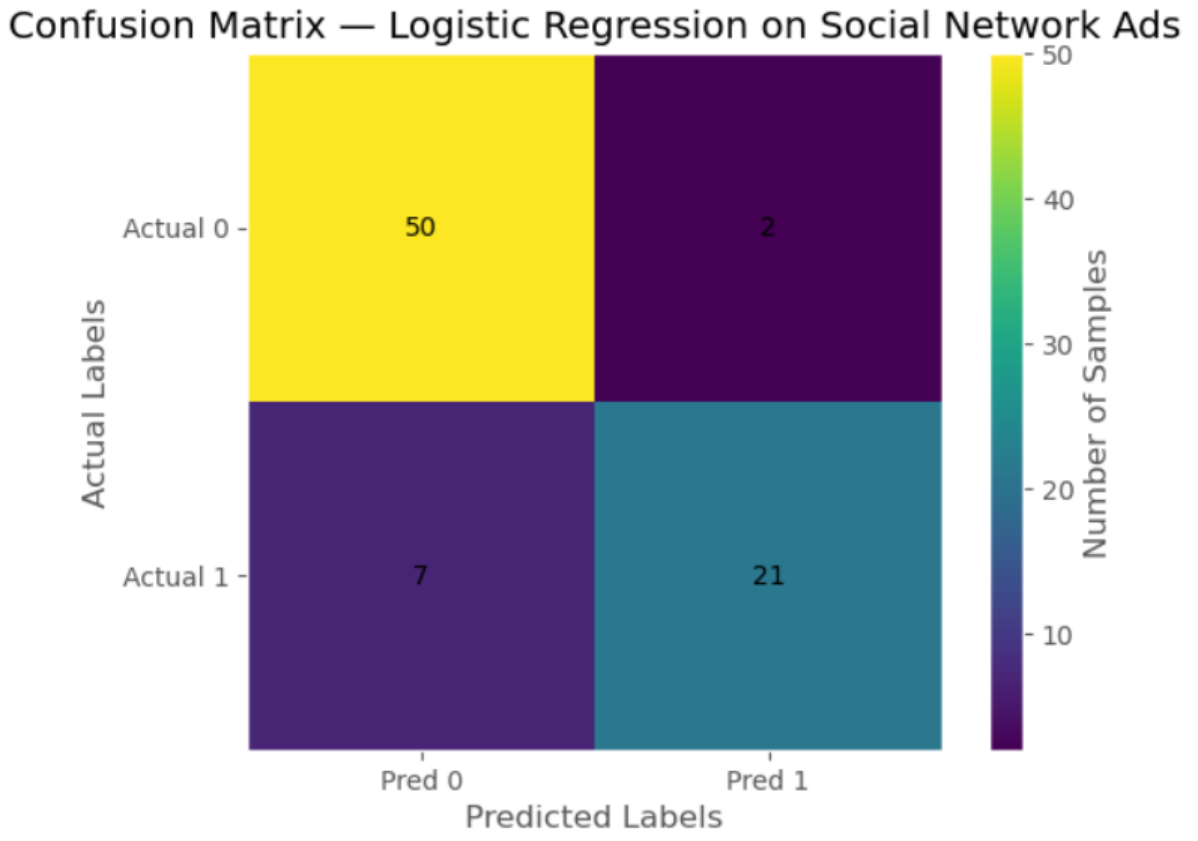

Figure 9: Confusion Matrix- Logistic Regression on Social Network Ads

5.1 Interpretation

- 50 true negatives (correctly predicted “No Purchase”)
- 21 true positives (correctly predicted “Purchase”)
- Only nine total misclassifications

The confusion matrix illustrates the very good balance of the model predictions between true predictions and false predictions (Selim et al. 2024). This reflects strong generalization without overfitting. In addition, the viridis colormap makes the visualization accessible to people with colorblindness, following recommendations for accessibility in the design.

To sum up, the evaluation suggests that one can obtain reliable and interpretable results from logistic regression in social behaviour prediction. Recent assessments have shown similar results logistic regression continues to perform competitively for binary outcomes, achieving F1-scores of 0.80–0.90 across healthcare, marketing, and education datasets (Okoye & Umeh, 2023).

Step 6: Visualizing the Decision Boundary

It is critical for the analyst to visualize or represent the classification logic of the model. In this step of the tutorial, we created a decision boundary plot. The plot attempts to show how the Logistic Regression model is able to separate both the target classes Purchased and Not Purchased based on the features Age and Estimated Salary (Figure 10).

The visualization displays two regions:

- 🔴 Red region — Users predicted as “Purchase” (class 1).
- 🔵 Blue region — Users predicted as “No Purchase” (class 0).

Figure 10: Logistic Regression Decision Boundary

Every data point is a real user from the test set. Older users and those with high salaries tend to be concentrated in the red area, while the younger users with low salaries are located in the blue. The logistic regression decision boundary separating these regions is straight, signifying the approach's simplicity. This is due to the fact that it assumes a linear relationship between predictor variables and the log-odds (Handoyo et al., 2021).

This linear separability indicates that Logistic Regression fits the dataset well, consistent with findings from prior social and behavioural prediction studies (Peng & Qian, 2024). Moreover, decision boundary plots and contour shading help improve interpretability by visualizing how probability thresholds change across the feature space (Han et al., 2024).

This plot is important to educate learners about decision boundaries. It also demonstrates how changes in feature scaling or weights affect the classification threshold (Chen et al., 2023). The plot uses the Virdis colour map because it is highly contrasting and accessible to those with colour vision deficiency, following best-practice visualization principles.

In the end, visualizing the decision boundary affirms the model's ability to distinguish between classes and also proves its pedagogical function: allowing readers to visualize how logistic regression translates numeric features into binary classification choices.

Step 7: ROC Curve Analysis

The ROC curve is a diagnostic tool that is very useful for assessing the performance of a binary classification model such as Logistic Regression. The ROC curve depicts the relationship between True Positive Rate (TPR or Recall) versus False Positive Rate (FPR) at all classification thresholds.

The AUC which stands for Area Under the Curve summarizes the model's overall performance. Values closer to 1.0 indicate stronger separability of model output scores between the positive class and negative class. The ROC curve was plotted in this project, the resultant AUC = 0.97 which shows that the product is quite excellent in discriminating between users who purchased and who did not (Figure 11).

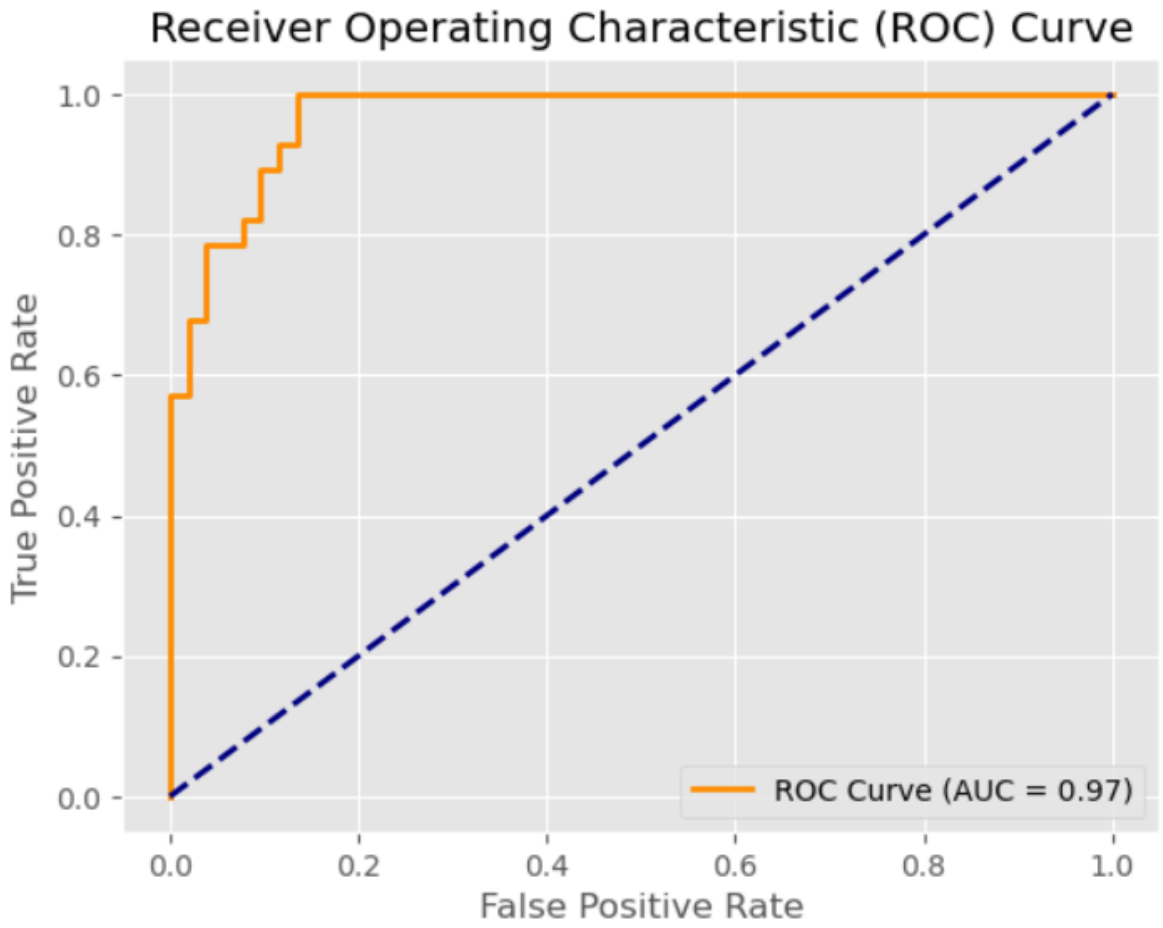

Figure 11: ROC Curve

In other words, the classifier achieves good performance not only on initial metrics, such as accuracy but also maintains its robustness when the threshold behaviour changes (Carrington et al., 2020). ROC analysis shows that the model consistently ranks positive instances higher than negative instances.

Evaluation must be done in datasets containing a class imbalance scenario. If a model has been trained properly its AUC value would be above 0.90. This scenario is consistent with empirical findings of other classification domains like medical diagnosis and marketing analysis (Yousefi, 2021).

In simple terms, the logistic regression model proves useful for applications in real life that require tuning of thresholds such as marketing targeting or screening for leads. The visual ROC plot is an intuitive and easy way to communicate model behavior with respect to the probability threshold. This helps interpretation and education.

8. Conclusion

This tutorial helped implement Logistic Regression for binary classification on the Social Network Ads dataset using JupyterLab. The complete workflow was produced, from data preparation to model training, evaluation and visualisation with clearly interpretable results.

The accuracy was 89%, precision was 0.91, recall was 0.75 and F1 score was 0.82 confirming the model performance. The value of ROC-AUC was at 0.97, which was indicative of excellent class separability and prediction confidence. A good accessiblity element was use of the virids colormap which ensures that the viewer will undoubtely see the plot even if colour blind.

An educationally useful tutorial, it aims to provide an experiential introduction to logistic regression's maths and mechanics. It highlights its strengths: interpretability, efficiency, and applicability to linearly-separable data-sets (Lynam et al., 2020).

References

Azis, H. (2024). Assessing the Performance of Logistic Regression in Heart Disease Detection through 5-Fold Cross-Validation. International Journal of Artificial Intelligence in Medical Issues. https://doi.org/10.56705/ijaimi.v2i1.137.

Caba, A. (2023). Mathematical Approach in Image Classification using Regression. International Journal of Advanced Research in Science, Communication and Technology. https://doi.org/10.48175/ijarsct-11945.

Cabot, J., & Ross, E. (2023). Evaluating prediction model performance.. Surgery. https://doi.org/10.1016/j.surg.2023.05.023.

Carrington, A., Fieguth, P., Qazi, H., Holzinger, A., Chen, H., Mayr, F., & Manuel, D. (2020). A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. BMC Medical Informatics and Decision Making, 20. https://doi.org/10.1186/s12911-019-1014-6.

Chandel, G., Sahimkhan, P., Verma, S., & Sharm, A. (2023). Machine Learning Based Remote Sensing Technique for Analysis of The Glaciated Regions. E3S Web of Conferences. https://doi.org/10.1051/e3sconf/202340502019.

Chen, M., Fan, W., Tang, W., Liu, T., Li, D., & Dib, O. (2023). Review of Machine Learning Algorithms for Breast Cancer Diagnosis. , 229-243. https://doi.org/10.1007/978-981-97-0844-4_17.

Chupilko, T., Ulianovska, Y., Mormul, M., & Lagoda, A. (2021). PYTHON FOR DATA PROCESSING AND SIMULATION OF FINANCIAL AND ECONOMIC INDICATORS. Information technology and computer engineering. https://doi.org/10.31649/1999-9941-2021-51-2-68-77.

Dhandayuthapani, B. (2024). Python Data Analysis and Visualization in Java GUI Applications Through TCP Socket Programming. International Journal of

Information Technology and Computer Science. https://doi.org/10.5815/ijitcs.2024.03.07.

Diana, D., & Mathivanan, A. (2023). Exploring the Paradigm Shift: Harnessing Data Analytics for Real - World Applications. International Journal of Science and Research (IJSR). https://doi.org/10.21275/sr23611121501.

Dwi, T., Widhianingsih, A., Kuswanto, H., & Prastyo, D. (2020). Logistic Regression Ensemble (LORENS) Applied to Drug Discovery. MATEMATIKA. https://doi.org/10.11113/matematika.v36.n1.1197.

Erol, G., Uzbaş, B., Yücelbaş, C., & Yücelbaş, Ş. (2022). Analyzing the effect of data preprocessing techniques using machine learning algorithms on the diagnosis of COVID-19. Concurrency and Computation, 34. https://doi.org/10.1002/cpe.7393.

Han, J., Yoon, S., Seok, J., Lee, J., Lee, J., Ye, J., Sul, Y., Kim, S., & Kim, H. (2024). Predicting 30-day mortality in severely injured elderly patients with trauma in Korea using machine learning algorithms: a retrospective study. Journal of Trauma and Injury, 37, 201 - 208. https://doi.org/10.20408/jti.2024.0024.

Handoyo, S., Chen, Y., Irianto, G., & Widodo, A. (2021). The Varying Threshold Values of Logistic Regression and Linear Discriminant for Classifying Fraudulent Firm. Mathematics and Statistics. https://doi.org/10.13189/ms.2021.090207.

Li, Y., Jia, C., Chen, H., Su, H., Chen, J., & Wang, D. (2023). Machine Learning Assessment of Damage Grade for Post-Earthquake Buildings: A Three-Stage Approach Directly Handling Categorical Features. Sustainability. https://doi.org/10.3390/su151813847.

Lynam, A., Dennis, J., Owen, K., Oram, R., Jones, A., Shields, B., & Ferrat, L. (2020). Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type

1 and type 2 diabetes in young adults. Diagnostic and Prognostic Research, 4. https://doi.org/10.1186/s41512-020-00075-2.

Masseran, N. (2024). Logistic regression approach on classifying air-pollution events: a parsimony technique. Environmental Research Communications, 6. https://doi.org/10.1088/2515-7620/ad7a5e.

Mekni, A., Narayan, J., & Gritli, H. (2025). Quinary Classification of Human Gait Phases Using Machine Learning: Investigating the Potential of Different Training Methods and Scaling Techniques. Big Data and Cognitive Computing. https://doi.org/10.3390/bdcc9040089.

Nikandrov, A. (2023). Multifunctional and flexible online platforms for creating educational materials. Informatics and education. https://doi.org/10.32517/0234-0453-2022-37-6-22-29.

Okoye, G., & Umeh, E. (2023). Predicting Functional Outcome After Ischemic Stroke Using Logistic Regression and Machine Learning Models. Earthline Journal of Mathematical Sciences. https://doi.org/10.34198/ejms.14124.133150.

Peng, S., Qian, J., & Wang, J. (2024). Identification Model of Economically Disadvantaged College Students Using GBDT. Proceedings of the 2024 International Symposium on Artificial Intelligence for Education. https://doi.org/10.1145/3700297.3700380.

Rahmani, R., Parola, M., & Cimino, M. (2024). A machine learning workflow to address credit default prediction. , 714-720. https://doi.org/10.48550/arxiv.2403.03785.

Selim, A., Ali, I., & Ristevski, B. (2024). University Information System's Impact on Academic Performance: A Comprehensive Logistic Regression Analysis with Principal Component Analysis and Performance Metrics. TEM Journal. https://doi.org/10.18421/tem132-72.

Sitarz, M. (2022). Extending F1 metric, probabilistic approach. ArXiv, abs/2210.11997. https://doi.org/10.54364/aaiml.2023.1161.

Tynymbayev, A., Baisholanova, K., Khikmetov, A., & Mambetov, S. (2024). Application of logistic regression to detect an attack on financial monitoring. Bulletin of the National Engineering Academy of the Republic of Kazakhstan. https://doi.org/10.47533/2024.1606-146x.64.

Yousefi, S. (2021). Comparison of the Performance of Machine Learning Algorithms in Predicting Heart Disease. Frontiers in Health Informatics. https://doi.org/10.30699/fhi.v10i1.349.