International Journal of Multidiscipline

TLEP – International Journal of Multidiscipline
(Technology, Language, Education, and Psychology)
ISSN: 2488-9342 (Print) | 2488-9334 (Online)

Open Access | Peer-Reviewed | Monthly Publication | Impact factor: 8.497 / 2025

Machine Learning Algorithms for Classification Tasks after Valuable Data Preprocessing: Evidence from Economic Indicator Data

Azizjon Meliboev
Digital technology and Mathematics, Kokand University
Email: a.meliboyev@kokanduni.uz

Abstract
Machine learning algorithms are increasingly used for classification tasks in artificial intelligence, cybersecurity, economics, and data-driven decision-making. However, the quality of classification results depends not only on the selected algorithm but also on the quality of data preprocessing. This study investigates the role of valuable data preprocessing in improving machine learning-based classification performance using economic indicator data. The dataset contains 147 monthly observations from February 2013 to April 2025 and includes monetary and deposit-related variables such as broad money supply, national currency money supply, narrow money supply, cash in circulation, demand deposits, other national currency deposits, and foreign currency deposits expressed in national currency equivalent. The research follows the IMRAD structure and applies exploratory data analysis, missing value checking, descriptive statistics, distribution analysis, correlation analysis, feature preparation, and classification modeling. Four machine learning algorithms were applied: Decision Tree, Random Forest, Logistic Regression, and Support Vector Machine. The models were evaluated using ROC curves and confusion matrix-based metrics, including accuracy, precision, recall, specificity, and F1-score. The results show that preprocessing produced a clean dataset with no missing values and revealed strong positive correlations among the selected variables. The final classification result achieved an accuracy of approximately 98.83%, precision of 97.87%, recall of 100%, and F1-score of 98.92%. The findings confirm that systematic preprocessing and exploratory analysis are essential for obtaining reliable machine learning classification results. At the same time, because economic indicators have strong time-series characteristics, future studies should apply chronological validation and rolling-window testing to avoid overly optimistic performance estimation.
Keywords: machine learning; classification; data preprocessing; economic indicators; monetary aggregates; ROC curve; confusion matrix

1. Introduction

In recent years, machine learning has become one of the most important tools for solving classification, prediction, anomaly detection, and decision-support problems. Its applications cover many fields, including artificial intelligence, cybersecurity, finance, economics, healthcare, and public administration. In classification tasks, machine learning algorithms learn patterns from historical data and assign observations to predefined classes. The quality of such models depends on several factors, including the structure of the dataset, the relevance of variables, the preprocessing strategy, and the choice of algorithm [1-4].

Data preprocessing is one of the most important stages in any machine learning pipeline. Raw data may contain missing values, inconsistent formats, outliers, duplicated records, scale differences, and highly correlated variables. These issues can negatively affect classification performance and may lead to inaccurate conclusions. Therefore, before applying machine learning models, it is necessary to analyze the dataset, clean the data, transform variables, examine distributions,

and evaluate relationships among features [5,6].

This study focuses on a classification task based on economic indicator data. The dataset includes monthly monetary and deposit-related indicators from February 2013 to April 2025. These variables represent important components of monetary circulation and financial activity. Economic indicators often contain strong trends, high correlations, and time-dependent patterns. Therefore, they provide a useful case for studying the importance of preprocessing before machine learning classification.

The main purpose of this paper is to evaluate the effectiveness of machine learning algorithms for classification after valuable data preprocessing. The study applies Decision Tree, Random Forest, Logistic Regression, and Support Vector Machine models. The performance of these models is assessed using ROC curves and confusion matrix-based evaluation metrics.

The main contributions of this study are as follows. First, the paper presents a practical preprocessing-based framework for classification tasks. Second, it demonstrates how descriptive statistics, missing value analysis, distribution analysis, and correlation analysis can improve understanding of the dataset before model training. Third, it compares several machine learning algorithms on a classification task using economic indicator data. Fourth, it discusses the importance of careful validation when working with time-series economic data.

2. Materials and Methods

2.1. Dataset Description

The dataset used in this study consists of 147 monthly observations covering the period from February 2013 to April 2025. The data contain monetary and deposit-related economic indicators. The variables included in the dataset are presented in Table 1.

The dataset is suitable for classification modeling because it contains structured numerical variables. However, since the variables are economic indicators, they may contain strong trend behavior and high correlation. For this reason, exploratory data analysis was conducted before model training.

No.	Variable	Description
1	Date	Monthly observation date
2	Broad money supply	General money supply indicator
3	National currency money supply	Money supply in national currency
4	Narrow money supply	Narrow monetary aggregate
5	Cash in circulation	Cash money in circulation
6	Demand deposits in national currency	Deposits available on demand
7	Other deposits in national currency	Other national currency deposits
8	Foreign currency deposits	Foreign currency deposits expressed in national currency equivalent

Table 1. Variables used in the study

International Journal of Multidiscipline

TLEP – International Journal of Multidiscipline
(Technology, Language, Education, and Psychology)
ISSN: 2488-9342 (Print) | 2488-9334 (Online)

Open Access | Peer-Reviewed | Monthly Publication | Impact factor: 8.497 / 2025

```
data.head()
```

	Sana	Keng ma'nodagi pul massasi)	Milliy valyutadagi pul massasi	Tor ma'nodagi pul massasi \n(M1)2)	Muomaladagi \nnaqd pullar\n(M0)	Milliy valyutagadi talab qilib olinguncha depozitlar	Milliy valyutadagi boshqa depozitlar	Chet el valyutasidagi depozitlar, \nmilliy valyuta ekvivalentida\n
0	2013-02-01	24432.690855	18900.759121	12589.001424	5786.389542	6822.611882	6311.757697	5531.931734
1	2013-03-01	24502.445158	18943.086657	12731.731201	6005.944703	6725.786495	6211.356456	5559.358501
2	2013-04-01	24961.911866	19082.114876	12801.825450	6223.047228	6578.778222	6280.289426	5899.796990
3	2013-05-01	24927.939824	18893.785473	12585.101393	6317.629747	6267.471646	6308.684080	6034.154361
4	2013-06-01	25424.860416	19698.938001	13246.989191	6382.167615	6864.831576	6351.938810	5825.921415

Figure 1. Preview of the dataset using data.head().

2.2. Data Preprocessing

The preprocessing stage was designed to prepare the dataset for machine learning classification. The following steps were performed:

The date column was converted into a suitable time format.

Missing values were checked using data.isnull().sum().

Descriptive statistics were obtained using data.describe().

Distribution plots were generated for numerical variables.

Two-dimensional scatter plots were used to observe relationships between selected variables.

Time-series plots were used to examine changes over time.

A correlation heatmap was created to identify relationships among features.

Features and target classes were prepared for classification.

Machine learning models were trained and evaluated.

The missing value analysis showed that the dataset does not contain missing values. This means that no imputation was required before modeling.

Variable	Missing values
Date	0
Broad money supply	0
National currency money supply	0
Narrow money supply	0
Cash in circulation	0
Demand deposits in national currency	0
Other deposits in national currency	0
Foreign currency deposits	0

Table 2. Missing value analysis

International Journal of Multidiscipline

TLEP – International Journal of Multidiscipline
(Technology, Language, Education, and Psychology)
ISSN: 2488-9342 (Print) | 2488-9334 (Online)

Open Access | Peer-Reviewed | Monthly Publication | Impact factor: 8.497 / 2025

```
data.isnull().sum()
```

	0
Sana	0
Keng ma'nodagi pul massasi1)	0
Milliy valyutadagi pul massasi	0
Tor ma'nodagi pul massasi \n(M1)2)	0
Muomaladagi \nnaqd pullar\n(M0)	0
Milliy valyutagadi talab qilib olinguncha depozitlar	0
Milliy valyutadagi boshqa depozitlar	0
Chet el valyutasidagi depozitlar, \nmilliy valyuta ekvivalentida\n	0

dtype: int64

Figure 2. Missing value checking output.

2.3. Descriptive Statistical Analysis

Descriptive statistics were used to understand the scale, range, and variability of the selected indicators. The analysis showed that all variables had 147 valid observations.

Indicator	Mean	Minimum	Maximum	Standard deviation
Broad money supply	103188.57	24432.69	281516.32	70100.26
National currency money supply	76941.96	18893.79	225485.77	53537.72
Narrow money supply	45986.54	12585.10	114401.27	27360.92
Cash in circulation	24015.57	5766.39	54020.79	14096.24
Demand deposits in national currency	21970.96	6267.47	61072.76	13515.78
Other deposits in national currency	30955.42	6211.36	119232.09	26743.41
Foreign currency deposits	26246.61	4703.60	58296.29	17634.70

Table 3. Main descriptive statistics

The results indicate that monetary and deposit indicators increased significantly during the analyzed period. For example, broad money supply increased from 24432.69 to 281516.32, while national currency money supply increased from 18893.79 to 225485.77. This confirms that the dataset contains strong growth patterns over time.

International Journal of Multidiscipline

TLEP – International Journal of Multidiscipline
(Technology, Language, Education, and Psychology)
ISSN: 2488-9342 (Print) | 2488-9334 (Online)

Open Access | Peer-Reviewed | Monthly Publication | Impact factor: 8.497 / 2025

```
data.describe()
```

	Sana	Keng ma'nodagi pul massasi1)	Milliy valyutadagi pul massasi	Tor ma'nodagi pul massasi \n(M1)2)	Muomaladagi \nnaqd pullar\n(M0)	Milliy valyutagadi talab qilib olinguncha depozitlar	Milliy valyutadagi boshqa depozitlar	Chet el valyutasidagi depozitlar, \nmilliy valyuta ekvivalentida\n
count	147	147.000000	147.000000	147.000000	147.000000	147.000000	147.000000	147.000000
mean	2019-03-02 07:20:48.979591936	103188.569249	76941.959288	45986.538357	24015.574028	21970.964360	30955.420901	26246.609961
min	2013-02-01 00:00:00	24432.690855	18893.785473	12585.101393	5766.388642	6267.471646	6211.355456	4703.602614
25%	2016-02-15 12:00:00	43023.222080	36338.926669	23460.077125	10810.238760	12540.741188	13199.736820	6539.170393
50%	2019-03-01 00:00:00	83565.847283	57721.182358	38114.959376	22404.509741	15756.420811	20691.886285	26808.020213
75%	2022-03-16 12:00:00	141805.771783	96946.552969	59997.523547	29541.150051	31160.954254	38347.693983	42110.956150
max	2025-04-01 00:00:00	281516.315816	225485.773877	114401.270164	54020.790812	61072.763523	119232.092360	58296.292480
std	NaN	70100.259821	53537.719124	27360.922746	14096.240796	13515.777147	26743.406307	17634.704841

Figure 3. Descriptive statistics generated using data.describe().

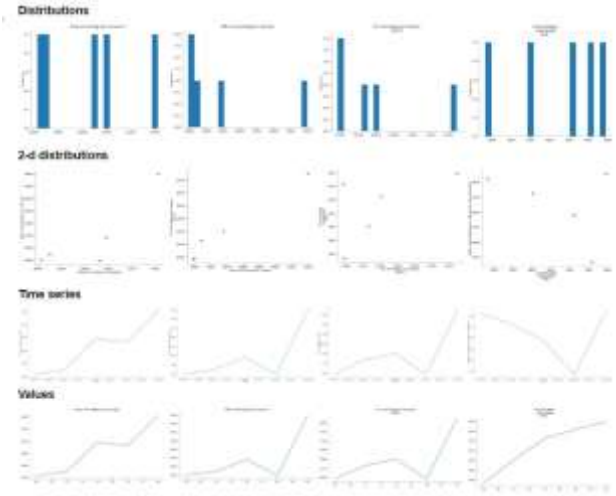

Figure 4. Distribution, two-dimensional, time-series, and value plots for selected variables.

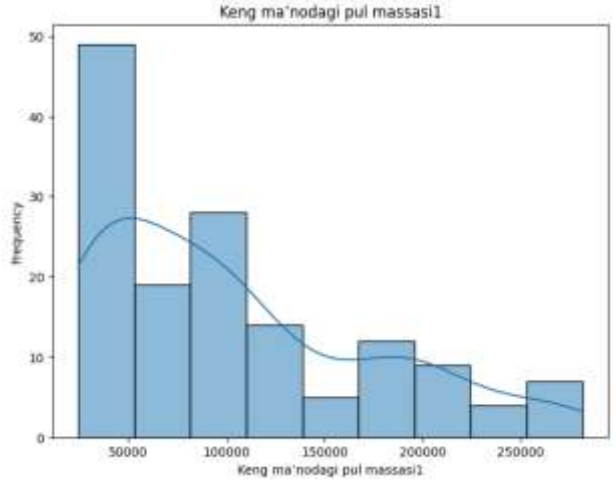

Figure 5. Histogram and density curve for broad money supply.

2.4. Correlation Analysis

Correlation analysis was used to identify the strength of relationships among the variables. The correlation heatmap showed strong positive relationships between most features. The correlation coefficients were generally high, ranging approximately from 0.87 to 1.00.

This result indicates that monetary and deposit indicators tend to move together. For example, broad money supply, national currency money supply, narrow money supply, and deposits showed very strong correlations. However, high correlation may also indicate multicollinearity. In machine learning, multicollinearity may not seriously affect tree-based models such as Decision Tree and Random Forest, but it can influence linear models such as Logistic Regression. Therefore, correlation analysis is an important preprocessing step.

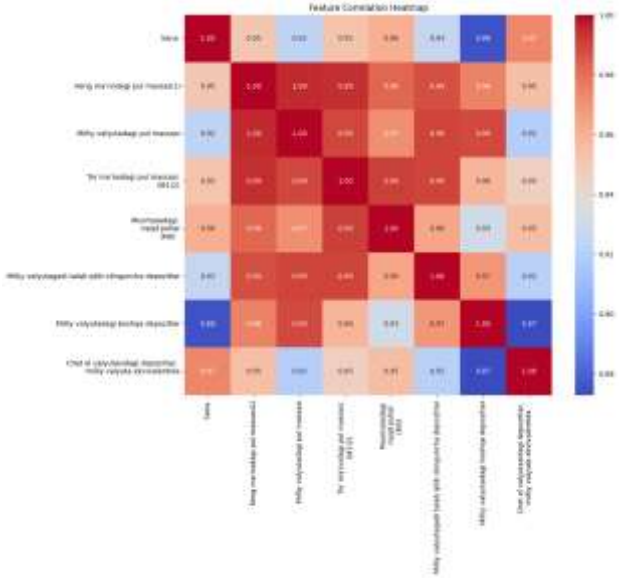

Figure 6. Feature correlation heatmap.

2.5. Machine Learning Algorithms

Four machine learning algorithms were applied in this study.

Decision Tree. Decision Tree is a supervised learning algorithm that classifies

observations by dividing the dataset into branches based on feature conditions. It is easy to interpret and useful for understanding decision rules [7].

Random Forest. Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions. It usually provides better generalization than a single decision tree and reduces the risk of overfitting [1].

Logistic Regression. Logistic Regression is a statistical machine learning method used for binary classification. It estimates the probability that an observation belongs to a particular class [8].

Support Vector Machine. Support Vector Machine is a classification algorithm that finds an optimal separating boundary between classes. It is effective for classification tasks with clear separation between classes [2].

2.6. Evaluation Metrics

The performance of the classification models was evaluated using ROC curves and a confusion matrix. The following metrics were calculated [3]:

$Accuracy = (TP + TN) / (TP + TN + FP + FN)$

$Precision = TP / (TP + FP)$

$Recall = TP / (TP + FN)$

$Specificity = TN / (TN + FP)$

$F1\text{-}score = 2 \times Precision \times Recall / (Precision + Recall)$

where TP means true positive, TN means true negative, FP means false positive, and FN means false negative.

3. Results

3.1. Exploratory Data Analysis Results

The exploratory data analysis showed that the dataset was clean and complete. No missing values were detected in any column. The distribution plots showed that several variables were not normally distributed and had increasing patterns over time. This is expected in economic time-series data because monetary aggregates and deposits usually grow as the economy expands.

The time-series plots showed clear upward movement in broad money supply, national currency money supply, narrow money supply, cash in circulation, and deposits. This confirms that the dataset contains strong temporal dynamics.

The correlation heatmap demonstrated that most variables are strongly positively correlated. Correlation values between the main monetary indicators were close to 1.00. This means that the selected features contain similar economic movement patterns.

3.2. ROC Curve Results

The ROC curve was used to compare the classification performance of Decision Tree, Random Forest, Logistic Regression, and Support Vector Machine. The ROC curves showed that all tested models achieved strong classification performance. The curves were close to the upper-left corner, which indicates a high true positive rate and a low false positive rate.

Among the tested algorithms, Support Vector Machine and Random Forest demonstrated particularly strong performance. Logistic Regression also showed good results, although its curve was slightly lower than the tree-based and SVM models. These results indicate that the selected variables provide strong predictive information for the classification task.

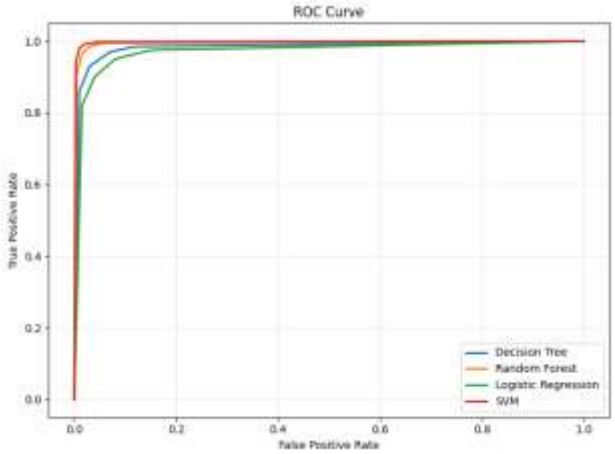

Figure 7. ROC curves of Decision Tree, Random Forest, Logistic Regression, and SVM models.

International Journal of Multidiscipline
TLEP – International Journal of Multidiscipline
(Technology, Language, Education, and Psychology)
ISSN: 2488-9342 (Print) | 2488-9334 (Online)
Open Access | Peer-Reviewed | Monthly Publication | Impact factor: 8.497 / 2025

3.3. Confusion Matrix Results

The confusion matrix result is presented in Table 4 and Figure 8.

True / Predicted	Positive	Negative
Positive	25189	0
Negative	548	21256

Table 4. Confusion matrix result

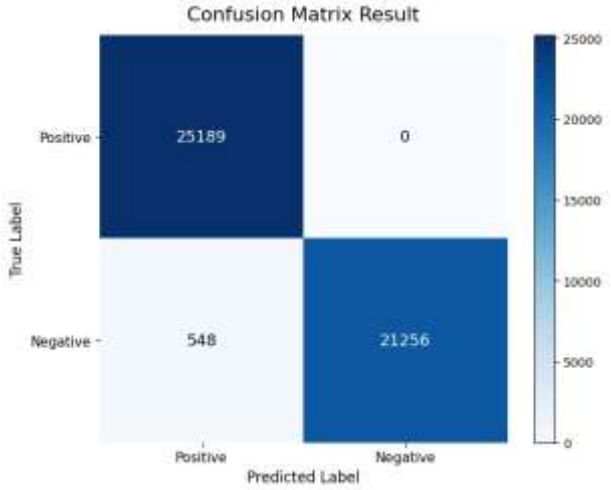

Figure 8. Confusion matrix result.

Based on the confusion matrix, the model produced 25189 true positive classifications, 21256 true negative classifications, 548 false positive classifications, and 0 false negative classifications. The model correctly classified 25189 positive cases and 21256 negative cases. Only 548 negative cases were incorrectly classified as positive. No positive cases were incorrectly classified as negative.

3.4. Classification Performance Metrics

Metric	Value
Accuracy	98.83%
Precision	97.87%
Recall	100.00%
Specificity	97.49%
F1-score	98.92%

Table 5. Classification performance metrics

The accuracy of 98.83% shows that the model correctly classified most observations. The recall value of 100% indicates that all positive cases were detected. The precision value of 97.87% shows that most observations predicted as positive were actually positive. The F1-score of 98.92% confirms that the model achieved a strong balance between precision and recall.

4. Discussion

The results of this study confirm that data preprocessing plays a critical role in machine learning classification tasks. The dataset was first analyzed using missing value checking, descriptive statistics, distribution plots, time-series visualization, and correlation analysis. This preprocessing process helped identify the main structure of the data and prepared it for classification modeling.

The absence of missing values improved the reliability of the modeling process. Since all variables had complete observations, there was no need to apply missing value imputation. This reduced the risk of artificial bias that may appear when missing values are replaced with estimated values.

The descriptive statistics showed large differences between minimum and maximum values. This indicates significant growth in monetary and deposit indicators over the observation period. Such growth patterns are common in economic data, but they may influence machine learning models if not properly handled.

The correlation analysis showed very strong positive relationships among the variables. This means that the selected economic indicators move together over time. From a modeling perspective, this can be useful because related variables may improve classification accuracy. However, it can also create multicollinearity, especially for linear models such as Logistic Regression. Therefore, future research should consider feature selection, principal component analysis, or regularization techniques.

The ROC curve results demonstrated that all selected algorithms performed well. This suggests that the preprocessed features contained enough information for effective

classification. The confusion matrix also confirmed strong classification performance, with only 548 misclassified cases out of 46993 evaluated predictions.

However, the very high performance should be interpreted carefully. Economic data usually have time-dependent structure. If a random train-test split is used, the model may indirectly learn future patterns from the data, which can lead to overly optimistic results. Therefore, chronological train-test splitting, rolling-window validation, and out-of-sample testing are recommended for future experiments.

Another important point is that the original dataset contains 147 monthly observations, while the confusion matrix includes a larger number of evaluated predictions. This may be explained by additional preprocessing, transformation, resampling, or generated classification instances. In the final version of the paper, this step should be clearly described based on the exact code and dataset preparation process.

Overall, the findings show that machine learning algorithms can successfully classify structured economic data when preprocessing is performed carefully. The study also demonstrates that preprocessing is not a technical formality, but a fundamental stage that directly affects model performance and the reliability of conclusions.

5. Conclusion

This study investigated the application of machine learning algorithms for a classification task after valuable data preprocessing. The research used monthly economic indicator data covering the period from February 2013 to April 2025. The dataset included broad money supply, national currency money supply, narrow money supply, cash in circulation, demand deposits, other national currency deposits, and foreign currency deposits expressed in national currency equivalent.

The preprocessing stage included missing value checking, descriptive statistics, distribution analysis, time-series visualization, and correlation analysis. The results showed that the dataset contained no missing values and that most variables were strongly positively correlated. The descriptive analysis also showed significant growth in monetary and deposit indicators during the analyzed period.

Four machine learning algorithms were applied: Decision Tree, Random Forest, Logistic Regression, and Support Vector Machine. The models were evaluated using ROC curves and confusion matrix-based metrics. The final classification result achieved 98.83% accuracy, 97.87% precision, 100% recall, 97.49% specificity, and 98.92% F1-score.

The findings confirm that valuable data preprocessing improves the quality and reliability of machine learning classification. The study also shows that economic indicator data can be effectively used in classification tasks when proper preprocessing and model evaluation methods are applied.

Future research should apply chronological validation, rolling-window testing, feature selection, and additional machine learning algorithms. It is also recommended to compare classical machine learning models with deep learning and time-series classification methods.

6. Limitations and Future Work

The main limitation of this study is connected with the time-series nature of the data. Economic indicators change over time and may contain strong trend effects. Therefore, random train-test splitting can produce overly high results. Future studies should use chronological split, walk-forward validation, or rolling-window testing.

A second limitation is the difference between the original number of monthly observations and the larger number of evaluated predictions shown in the

confusion matrix. This transformation should be fully documented in the final experimental code and dataset description. Future research should clearly explain whether the increase is caused by windowing, resampling, class generation, or another transformation method.

Future work may also include additional preprocessing methods such as normalization, differencing, feature selection, principal component analysis, and model explainability tools such as SHAP or permutation importance.

References

Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20, 273-297.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. Morgan Kaufmann.

Geron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.

Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1, 81-106.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning. Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

Central Bank of the Republic of Uzbekistan. Monetary and deposit indicators database.