THE IMPACT OF FEATURE SELECTION AND DATA PRE-PROCESSING ON ML MODELS

Alexandru-Carlos Vlad [1*]

Emilian Marian Iovanovici [1]

[1] Petroleum-Gas University of Ploiesti, Romania
email: alexandru.vlad@student.upg-ploiesti.ro

DOI: 10.51865/JPGT.2025.01.11

ABSTRACT

This research aims to identify issues with datasets that affect the training of machine learning models. The paper demonstrates the impact of feature processing on the model evaluation metrics. Initially, the paper evaluates unprocessed categorical features based on customized rules. The results of the seven evaluated algorithms show very low metric values. The demonstration is conducted on a dataset with 16 features for predicting depression among students. The dataset consists of 7,022 records. The issue of depression was chosen because it involves many features, which facilitate the analysis of the correlations between them. Moreover, the large number of features and records allowed for the analysis of generalization capacity by training on different dataset scenarios for all seven algorithms. The demonstration shows that data pre-processing generates better results when the inputs are exclusively numerical. Subsequently, the research demonstrates the importance of individually analyzing the contribution of each parameter within the model. The study encompasses three categories of tests, which are implemented using the ML.NET framework in the C# programming language.

Keywords: machine learning; depression; dataset; feature; algorithms.

INTRODUCTION

According to the World Health Organization, there are nearly 450 million people affected by depression. Since this disorder is closely related to genetics, biochemistry, environment, and psychology, the problem can be studied using machine learning (ML) techniques [1]. Due to the necessity of handling a large volume of data and a high number of features, the authors investigated the implications of developing ML models for addressing the issue of depression. Writing software tests for the comparative evaluation of ML algorithm performance requires infrastructure compatibility, as discussed in the paper [2].

The article [3] examines the need for using ML algorithms to identify symptoms of depression accurately. The article uses persistent sadness, cognitive difficulties, and the influence of social, psychological, and depressive factors as features. The evaluated ML methods were Random Forest (RF), Recurrent Neural Network (RNN), and Support Vector Machine (SVM). The study [4] proposes an ML model based on visual stimuli that uses 10 features. The study used 219 participants, both depressed and non-depressed, and the obtained accuracy was 85.6%. Other research focuses on developing a triage algorithm for symptom-based healthcare decision-making [5].

The paper [6] proposes an ML model for identifying depression in individuals aged between 6 and 17 years. The comparative analysis employed three algorithms: RF, SVM, and Logistic Regression (LR). Following the tests, it was found that the RF algorithm performed best, achieving 94% accuracy and 95% precision for non-depressed individuals and 88% for depressed individuals. The LR algorithm had an accuracy of 89% and a precision of 91%. Also, for identifying depression among adolescents aged 12 to 17, research was conducted in the paper [7]. The study was based on the analysis of calls, conversations, location, and heart rates of 55 adolescents. The data were collected through mobile phones and Fitbit device sensors. After the 24-week data collection period, the results showed that the main features influencing the outcomes were screen usage, calls, and location data [7].

Numerous studies have focused on ML models for depression detection through semantic text analysis. The tests in [8] examine online posts, and the degree of depression identification is reported with superior accuracy compared to existing ML models [9]. The paper [10] aims to implement a method for detecting and treating depression using mobile devices that generate real-time data about the individual. The proposed system analyzes nine characteristics of depression. The tests were conducted on 106 psychiatric patients and demonstrated the possibility of using it as a tool for predicting depression. Similarly, a study [11] conducted on 219 adults over 16 weeks collected data from phones, performing sentiment analysis on texts from the online environment. ML models achieved an AUC of 0.72 when only text messages were used and 0.76 when combined with other factors. The research [12] also developed a model based on Long-Short Term Memory (LSTM) for identifying depression from text messages. The dataset consists of young people's questions on an online Norwegian informational channel. Instead of word frequency, the model analyzes features proposed by expert doctors, differentiating between texts that are depressive and those that are not. Sentiment analysis can be performed automatically using pre-implemented tools to extract the user's mood from the text. The results presented in the papers [13, 14] demonstrate the feasibility of integrating them into the issue of depression detection. The study [15] suggests using a model that utilizes LSTM for detecting depressive traits, also through text analysis. The dataset contained 233,000 records. The identified features were processed using one-hot encoding. The method uses LSTM for classifying depressive and non-depressive sentiments from text. The performance reported by the authors of the paper [15] was 99%.

Article [16] considers the severe consequences that the pandemic has generated on the mental health of students [17]. The study proposes an ML model to predict depression among students. The experimental results showed a performance of 76%. The study [18] analyzes postpartum depression (PPD). A dataset of 214,359 births from 2008 to 2015 in Israel was used to train the model. The sociodemographic, clinical, and obstetric characteristics were analyzed by an algorithm based on decision trees with gradient boosting. As a reference value, the number of women who developed PPD was 1.9% or 4104. Upon validation, the proposed model demonstrated an AUC of 0.712, with a sensitivity of 0.349 and a specificity of 0.905 at a 90% risk threshold. A study was conducted during the pandemic using a dataset of 5,001 people from Norway [19]. The model aims to identify the factors most strongly correlated with depression. The study [19] also aimed to determine the best predictive models for this issue. The results showed that RF had the highest accuracy. The main features for predicting depression were self-perceived risk of exposure, financial situation, work-life balance, and social contact. The study found that although the epidemiological factor was predominant in the initial phase of the pandemic, the primary influence shifted towards socio-economic factors in the later stages. Thus, the importance of processing demographic, socioeconomic, and behavioral data for the possibility of timely intervention in preventing depression was highlighted. On the other hand, industrial evolution through the implementation of AI tools [20, 21] becomes a factor that favors the onset of depression. For this reason, the study of depression through ML models should be urgently researched to intervene preventively.

Article [22] has a different approach to analyzing depression. It analyzes audio features in identifying depression. The proposed model analyzes the Gaussian Naive Bayes (GNB), SVM, K-Nearest Neighbor (KNN), LR, and RF algorithms. From the tests, an accuracy of 82% was obtained for 15 selected features. A similar approach is reported in the paper [23], highlighting negative speech features to achieve a model with promising results [24]. The paper [25] also proposes using an ML model based on vocal audio features for classifying major depression. In the study, 120 subjects participated, of which 64 had severe depression, and 56 were healthy, aged between 16 and 25 years. As a result of the study, the ML model generated 1200 audio features per patient. The results showed that the classification model achieved an accuracy of 84.16% and a sensitivity of 95.38%. Also, for the identification of depression from audio content, the emotional state at the vocal data level was analyzed. The results showed a performance 67% higher than that of previous systems [26]. The analyzed audio characteristics were frequency and speech patterns. The model is superior due to the reduction of external noise. In this way, the margin of error that appeared in previous systems is reduced. The study [27] proposes a multimodal model that analyzes audio and text data. The results obtained from the tests reach an F1 score of 95.80% in depression detection.

The study [28] proposes an approach to identifying depression in elderly individuals through motion sensors. The research was conducted on a small group of people aged 65 and older. The monitoring period was about 6 months. The study employed Wi-Fi sensor monitoring. The ML model achieved an accuracy of 87.5%. The features on which the model was trained were sleep duration and the frequency of sleep interruptions.

In the oil and gas industry, ML algorithms are used to detect anomalies and defects in pipelines [29], [30] and to predict their corrosion [31], [32]. Models based on RF or XGBoost demonstrate the possibility of integrating these methods into the oil and gas industry, achieving performance metrics that validate their applicability in practice, with values such as 97.4% [29]. These approaches are based on processing operational data through feature selection and noise elimination. In this context, the software tools used influence the performance of ML algorithms through pre-implemented optimization procedures in data processing [33]. In this way, integrating ML into the transportation infrastructure and monitoring of natural gas represents a future strategy for maintenance optimization processes.

Analyzing the specialized literature, the importance of dataset quality is noted [34]. For this, the research focuses on the following research questions (RQs):

RQ1: What is the impact of the volume and diversity of data on the performance of ML models?

RQ2: How does data preprocessing (categorical variable encoding, normalization, handling missing values) affect the performance of ML models?

RQ3: Are there significant differences between the performance of models trained on raw data and those trained on pre-processed data?

RQ4: What are the characteristics whose contribution allows for the most accurate prediction?

RQ5: Which ML algorithm offers the best accuracy for the type of problem that includes many features?

The research examines the impact of dataset quality and processing method on the performance of ML models. This approach enables an understanding of how factors such as dataset volume, feature diversity, pre-processing techniques, and parameter correlation impact the generalization capability of ML models. Therefore, the research objectives are:

- Evaluation of the impact of raw data quality on model performance;
- Analysis of the contribution of features to model performance;
- Data optimization to facilitate model performance improvement;
- Identifying an algorithm suitable for the problem modeled through the training data.

A voluminous dataset comprising 7,022 records and 16 features was utilized to achieve these objectives. The features demonstrate the importance of each objective through models capable of making accurate predictions. The features correspond to the synthesis of the parameters analyzed in the previously presented specialized literature, except those related to text and voice.

ML Analyzed Dataset

The dataset used for comparative training of the models is sourced from the Kaagle platform [35]. The original dataset was processed to contain only data relevant to the research. Additionally, the values were pre-processed to be suitable for training an ML model. Therefore, normalization and outlier exclusion operations, which are used to identify anomalies, were applied. This research's dependent variable is the depression level (DL). This represents a student's depression level and is classified as an output variable (label). The independent variables (features) include information about age (A), university course (C), biological gender (G), overall average for study years (CGPA), stress level (SL), anxiety score (AS), sleep quality (SQ), physical activity (PA), diet quality (DQ), social support (SS), relationship status (RS), substance use (SU), counseling service use (CSU), family history of mental health issues (FH), presence of chronic illnesses (CI), financial stress (FS), and extracurricular involvement (EI). Figure 1 presents the conceptual diagram of the relation between features and labels.

Dividing the features into categories facilitates the understanding of the dataset by conducting tests on the contributions of each category in correctly identifying the value associated with DS. The categories were established based on the following reasoning:

- Demographic characteristics reflect the degree of cognitive deterioration introduced by the A-G-FH correlation. These variables influence the overall context of students' lives through the predisposition of age and gender to specific forms of degradation from academic pressures. For example, younger students may be more vulnerable to adapting to the educational environment. On the other hand, upperclassmen cope more easily with career-related pressures due to the experience accumulated over the years. The range of values for age varies between 18 and 28 years, and gender is coded as "Male" and "Female". Students with a family history of mental health issues are more likely to develop depression. This is due to genetic or environmental factors influencing their mental state. The range of values is quantified by "Yes" or "No";
- Academic Performance includes C and CGPA. Feature C reflects the difficulty of the studied specialization, with the value range being {"Engineering", "Business", "Computer Science", "Medical", "Law", and "Others"}. CGPA is an indicator of students' academic performance. Academic performance influences the level of stress, anxiety, or personal satisfaction, all of which have an impact on DL. CGPA ranges between 2.44 and 4, according to the American grading scale. These values indicate that the students included in the dataset have obtained overall averages ranging from 2.44 (relatively low) to 4.00 (the highest possible, representing excellent performance);

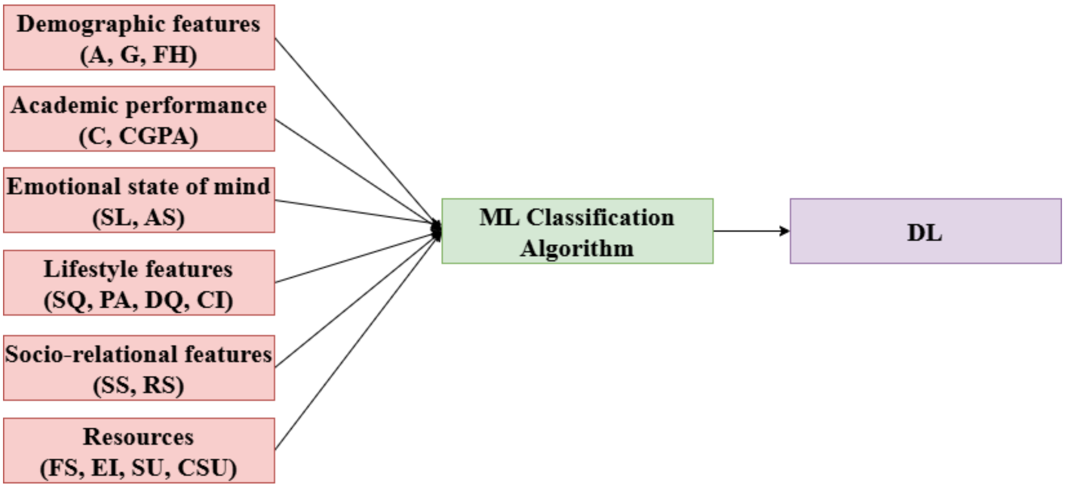

Demographic features (A, G, FH)
Academic performance (C, CGPA)
Emotional state of mind (SL, AS)
Lifestyle features (SQ, PA, DQ, CI)
Socio-relational features (SS, RS)
Resources (FS, EI, SU, CSU)
ML Classification Algorithm
DL

Figure 1. *The conceptual diagram of the ML Classification algorithms*

- The emotional state of mind contains SL and AS. These features are related to the emotional state of the students. These factors influence the level of depression because stress and anxiety are associated with the development of depression. SL ranges from 0 to 5, where higher values indicate a higher stress level. The same domain of representativeness is also present in the AS factor.
- Lifestyle features include the main aspects of students' lifestyles. These parameters can influence a student's level of depression. The SQ factor tracks sleep quality. Thus, a reduced quality of sleep, such as insomnia or sleep fragmentation, is associated with a state of mental fog that can lead to long-term depression. The range of values for this parameter is {"Good", "Average", or "Poor"}. Similarly, the PA and DQ parameters are introduced, as physical activity improves mood, and proper nutrition

contributes to overall body well-being. The range of values for PA is {"High", "Moderate", or "Low"}, and for DQ, it is {"Good", "Average", or "Poor"}. Additionally, students who suffer from chronic illnesses face additional difficulties compared to other students. The problems can be physical or psychological and can contribute to depression. The range of values for this parameter, CI, is {"Yes", "No"};

- Socio-relational features reflect social support and relational status. Students who receive social support from friends, family, or colleagues may cope more easily with depressive states. On the other hand, difficulties in relationships with others, especially at a young age, can promote the onset of depression. The range of values for SS is {"Good", "Average", or "Poor"}, and for RS {"Single", "Married", "In a Relationship"};
- In the end, access to financial resources, the behaviors associated with their use, and extracurricular involvement influence students' mental state, which is reflected in the value of the DS parameter. Students who cope with stress may feel frustration and helplessness, which can lead to depression. On the other hand, students who participate in extracurricular activities are engaged in fun activities, which can reduce the level of depression. The range of values for FS and EI is {"High", "Moderate", or "Low"}. Regarding the state of freedom, students who use psychotropic substances are exposed to an imminent risk of depression. The range of values for SU is {"Never", "Occasionally", "Frequently"}. On the other hand, students who use counseling services manage their stress or anxieties more easily, which reduces their level of depression. The range of values for CSU is {"Yes", "No"}.

The division of features into these categories enables the study of factors that influence the level of depression among students within the developed ML models. Each category designates a distinct domain of students' lives. The interactions between these domains can provide insight into the progression of depression development. For these reasons, the methodology for developing the best depression prediction model will include multiple analyses, through which the classes that significantly contribute to the correct identification of depression among students will be examined.

A total of 7,022 records were used for training and validation, comprising 80% for training (5,617 records) and 20% for validation. The validation stage generates a series of metrics that evaluate the model's ability to generalize, meaning to identify the degree of depression for value combinations of features it has never seen before. The research will answer the RQs by exemplifying the detection of depression among students due to the availability of data and the issue of feature selection, which makes the problem type under analysis particularly relevant.

MATERIALS AND METHODS

Predicting the level of depression involves identifying the non-linear relationships between the 16 features, which categorizes this problem as an ML classification issue. Modeling this problem consists of understanding fundamental concepts, which will be illustrated using the issue of depression among students. To answer the four RQs, three types of tests were conducted:

Test 1: Identifying the best algorithm for modeling the problem of identifying depression in students. The methodology involves evaluating seven classification algorithms. These algorithms were selected based on their ability to handle different complex datasets [36]:

- SdcaMaximumEntropy is based on the Stochastic Dual Coordinate Ascent (SDCA) method and is optimized for multinomial classification problems. This algorithm was chosen due to its ability to process large datasets, even when they contain redundant features;
- LbfgsLogisticRegression This is an improved version of the Logistic Regression algorithm. This is one of the fastest classification methods. The One-vs-All (OvA) implementation can be extended to multinomial classification problems. This extension makes it suitable for the issue of depression classification;
- FastForestOva originates from FastForest and represents an optimized implementation variant of the RF algorithm. This algorithm was analyzed because it is considered a benchmark for ML models;
- FastTreeOva is a classification algorithm based on decision trees, which is part of the Gradient Boosting Decision Trees (GBDT) model family;
- LightGbmMulti is a boosting algorithm based on decision trees. The Multi variant refers to the application of this algorithm to multi-class classification problems;
- LbfgsMaximumEntropyMulti combines the Maximum Entropy model with the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) optimization method for multi-class classification problems;
- SdcaLogisticRegression OvA combines LR with the SDCA method applied in a One-vs-All (OvA) approach for multi-class classification.

Each algorithm was configured individually and subjected to rigorous tests to evaluate its performance in different scenarios. Thus, SdcaMaximumEntropy stands out for handling large data volumes, while LbfgsLogisticRegressionOva and FastTreeOva excel in processing speed. FastForestOva, on the other hand, boasts a good reputation compared to other ML models.

Test 2: In the second test set, the issue of data pre-processing was raised to achieve better accuracy. This pre-processing reflects aspects of data quality. For this, the following modifications were made to the data:

- The records that contained NULL values for one of the features were removed;
- The issue was raised that ML algorithms work better with numerical data than with categorical data. Therefore, the feature values were converted to a numeric format using one-hot encoding or label encoding techniques. For example, the Gender variable was transformed into a binary variable (0 for male, 1 for female), and variables with multiple categories (such as Course) were encoded using one-hot encoding;
- Subsequently, the issue of binarizing all features was raised to determine whether the algorithms perform better with this approach.

Two specific metrics for classification algorithms were used to compare model performances: macro-accuracy and micro-accuracy. Macro-accuracy calculates the average accuracy for each class, giving equal importance to all classes. This metric is useful when the classes are imbalanced (for example, some levels of depression occur more frequently than others). Its calculation formula is presented in Eq. (1). Micro-accuracy treats all instances equally, regardless of their class, and is calculated using Eq. (2). This metric helps evaluate the model's overall performance [37].

Test 3: The third test aimed to identify the contribution of each parameter within the model. For this, the code was run for all 16 features, then they were excluded one by one, and then combinations of 2, 3, and partial 4 were generated. The authors resorted to this approach because obtaining all combinations would have generated 65535, which would have required too much computational effort.

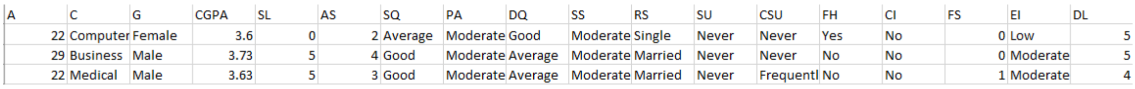
This research aims to provide a detailed understanding of how data quality and processing impact the performance of ML models. Evaluating a large dataset and several ML algorithms will identify best practices for building prediction models. The results obtained from the tests will address the research questions and provide concrete recommendations for optimizing the ML model development process in the context of students' mental health.

RESULTS

The algorithms for performing the tests are pre-implemented through the ML.NET tool, but the code is customized for each type of test. The code for each test is written in the C# programming language within the Visual Studio development environment.

Test 1

Figure 2 presents an example of data from the initial dataset. This figure shows that the data were pre-processed in an initial version, as the users' responses correspond to restricted value ranges. However, they are considered raw in this research because they have not been processed based on additional rules proposed by the authors. This test will conduct a comparative analysis between the ML algorithms, and the results obtained for micro-accuracy and macro-accuracy will be interpreted.



A	C	G	CGPA	SL	AS	SQ	PA	DQ	SS	RS	SU	CSU	FH	CI	FS	EI	DL
22	Computer	Female	3.6	0	2	Average	Moderate	Good	Moderate	Single	Never	Never	Yes	No	0	Low	5
29	Business	Male	3.73	5	4	Good	Moderate	Average	Moderate	Married	Never	Never	No	No	0	Moderate	5
22	Medical	Male	3.63	5	3	Good	Moderate	Average	Moderate	Married	Never	Frequentl	No	No	1	Moderate	4

***Figure 2**. Initial dataset example*

The results for the two metrics are presented in Figure 3. LightGBMMulti and L-BFGS Maximum Entropy Multi achieved the best results, with values of approximately 21% for macro-accuracy. These figures are low for the model to be used in prediction. These results reflect the difficulties in managing raw data. The LightGbmMulti and FastTreeOva algorithms perform worse than the others because they directly depend on the data's quality. Therefore, the next step is to optimize the data and reevaluate the model to achieve predictions that can be generalized.

In the following, the issue of improving the data used in training is addressed. A first idea considers categorical features, such as C or RS. A problem arises with their numerical encoding. Two pre-processing types will be performed in Test 2 to improve the results.

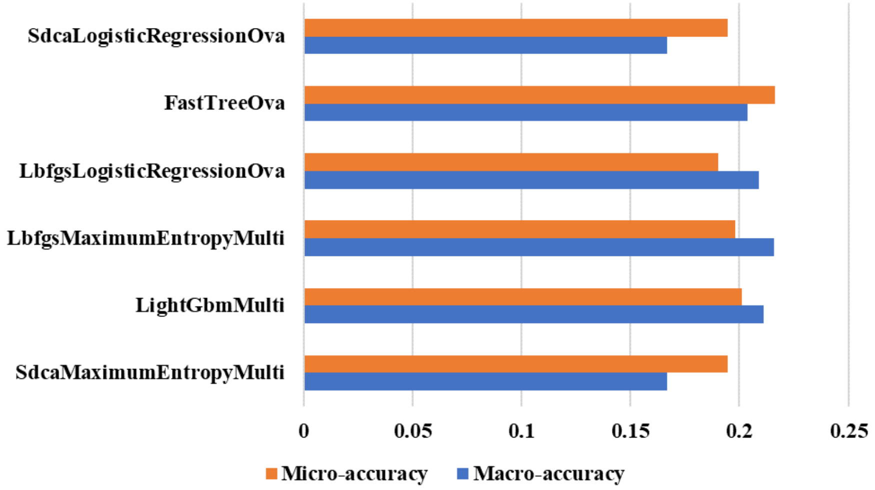
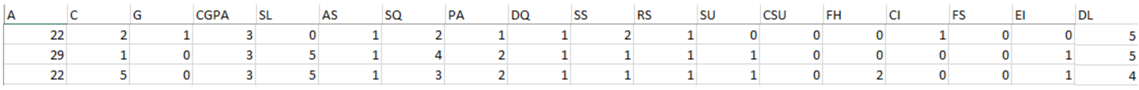


Figure 3. *Test 1 - comparative results of multiple ML algorithms*

Test 2

In Test 2, the issue of data preprocessing and the strategies used to improve prediction are addressed. Test 2 contains two subtasks labeled Test 2A and Test 2B.

In Test 2A, categorical features were numerically encoded using a systematic approach. For feature C, numerical encoding was used: Business = 1, Computer Science = 2, etc. Other categorical variables, such as G, SQ, PA, etc., are transformed into corresponding numerical values, as shown in Figure 4. This approach enables the analysis of evaluation metrics for ML models on numerical data. In this way, it can be determined whether this pre-processing proposal improves the metrics compared to the dataset used in Test 1.



A	C	G	CGPA	SL	AS	SQ	PA	DQ	SS	RS	SU	CSU	FH	CI	FS	EI	DL
22	2	1	3	0	1	2	1	1	2	1	0	0	0	1	0	0	5
29	1	0	3	5	1	4	2	1	1	1	1	0	0	0	0	1	5
22	5	0	3	5	1	3	2	1	1	1	1	0	2	0	0	1	4

Figure 4. *The second processed dataset example*

The results of Test 2A demonstrate an improvement in model performance compared to those of Test 1. The numerical encoding of categorical features had a positive impact on macro-accuracy. This suggests that this intermediate data processing contributes to the optimization of the models. Figure 5 presents the comparative results of the evaluation metrics of the ML algorithms. In this figure, it can be observed that FastForestOva had the best results for both micro-accuracy and macro-accuracy. Also, in Figure 5, it can be observed that all algorithms achieved results that were approximately 30% higher than in Test 1. This performance indicates that ML algorithms handle data complexity more effectively after numerical encoding.

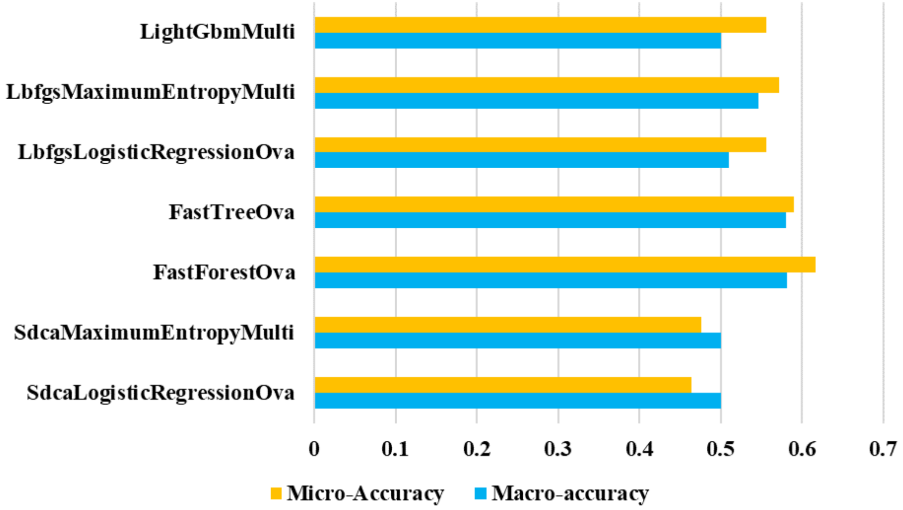
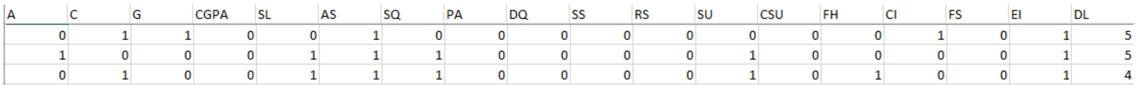


Figure 5. *Test 2A - comparative results of multiple ML algorithms for the second processed dataset*

Test 2B binarizes all features according to the rules in the SQL query. In Figure 6, both categorical and numerical features can be observed as binarized. Binarization should simplify the data. In this way, the model can interpret the data more effectively. At the same time, the model may lose detailed information from the initial numerical features. For these reasons, it is necessary to test the datasets multiple times in different scenarios to identify the sensitivity of the algorithms to data size, representation, and generalization capability.



A	C	G	CGPA	SL	AS	SQ	PA	DQ	SS	RS	SU	CSU	FH	CI	FS	EI	DL
0	1	1	0	0	1	0	0	0	0	0	0	0	0	1	0	1	5
1	0	0	0	1	1	1	0	0	0	0	1	0	0	0	0	1	5
0	1	0	0	1	1	1	0	0	0	0	1	0	1	0	0	1	4

Figure 6. *Test 2B - comparative results of multiple ML algorithms for the third processed dataset*

In test 2B, all features were binarized, transforming them into values of 0 or 1. The performance of the algorithms decreased compared to Test 2A, as shown in Figure 7. For all algorithms, the macro-accuracy was approximately 0.50, with insignificant variations.

The FastTreeOva algorithm, the leader in Test 2A, scored only 0.5056 compared to the previous test, which reported a macro-accuracy of 0.5805. The explanation is that binarization reduces data variability and eliminates subtle class differences. Consequently, the models can no longer use complex relationships between numerical variables, which were available in test 2A. Moreover, if the classes are imbalanced, binarization can exacerbate the discrepancies, which affects macro-accuracy.

Test 2 shows that numerical encoding of categorical features is more efficient than complete binarization. This is explained by numerical encoding, which retains more helpful information and leads to better training.

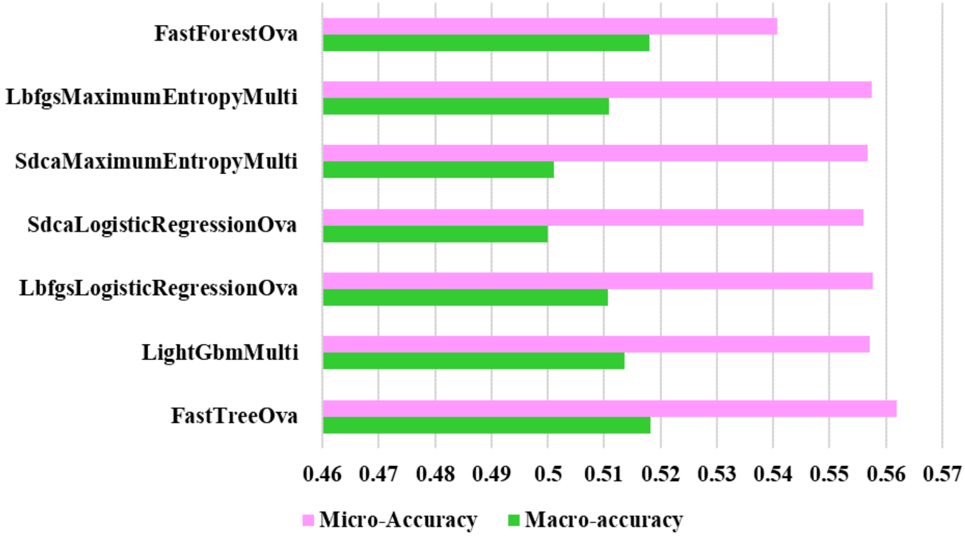


Figure 7. *Test 2B - comparative results of multiple ML algorithms for the third processed dataset*

Test 3

The third test aims to evaluate the contribution of each parameter within the model. For this purpose, a systematic evaluation was conducted, initially running the model with all 16 features included. Subsequently, the features were excluded individually, and various combinations of 2, 3, and partially 4 features were tested. The limitation on the number of partial tests for combinations of 4 is motivated by the computational effort required to evaluate all possible feature combinations. Moreover, generating 65,535 distinct configurations would have required high computational effort and execution time.

To evaluate the impact of different feature subsets on model performance, macro-accuracy, and micro-accuracy metrics were calculated for each configuration. The results presented in Figure 8 demonstrate that as the number of features increases from 1 to 4, both macro-accuracy and micro-accuracy exhibit subtle variations. For example, when only a single feature was used, the macro-accuracy was 0.6648, and the micro-accuracy reached a value of 0.6901. The macro-accuracy improved as additional features were added, reaching 0.6745 with 4 features. However, the micro-accuracy decreased to 0.6886 under the same conditions.

This behavior suggests that although increasing the number of features generally improves the model's ability to generalize across all classes (as reflected by macro-accuracy), it may also introduce complexities that slightly reduce the overall prediction accuracy (as indicated by micro-accuracy). The results highlight the importance of carefully selecting feature subsets to strike a balance between model complexity.

Test 3 highlighted the importance of identifying contributing features and graphically illustrated the contribution of feature selection. This perspective has significant implications for optimizing ML models.

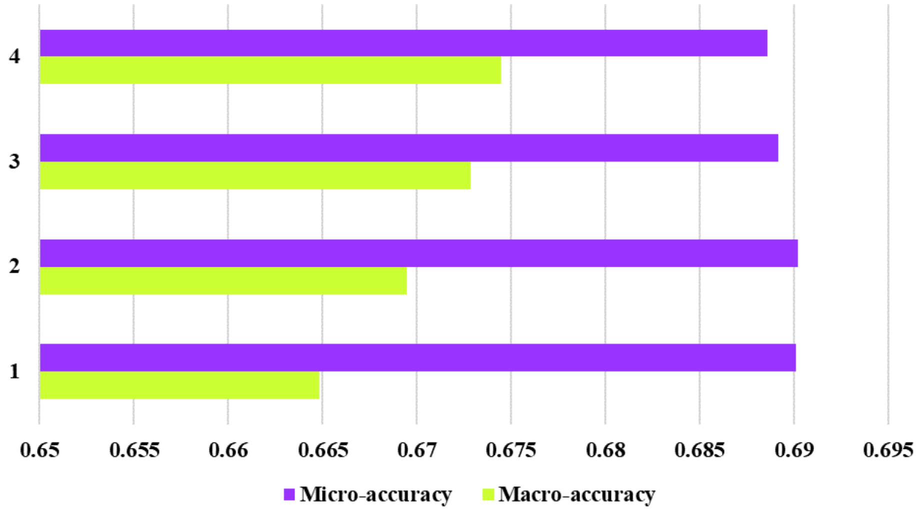


Figure 8. *The impact of feature selection combination on macro-Accuracy and micro-Accuracy analysis*

CONCLUSIONS

From the tests conducted in this research, the following answers to the RQs result:

RQ1: The volume of data and the number of features influence the generalization capacity of ML models. Thus, the models will have better accuracy if the modeled problem pre-processes the data. If the data is not sufficiently processed, it leads to poor performance for multiple reasons. For example, algorithms like FastForestOva and FastTreeOva achieved better results on numerically encoded data than raw data.

RQ2: Numeric encoding of categorical variables improves macro-accuracy (e.g., FastForestOva increased from 0.581 to 0.6745). Complete binarization reduced accuracy. From these tests, the importance of aligning the data with the modeled issue is deduced.

RQ3: Models trained on raw data performed approximately 20%, while data preprocessing (numerical encoding or partial binarization) increased macro-accuracy to 67%.

RQ4: The analysis conducted on subsets of features showed that gradually adding relevant features (up to 4) improved the macro-accuracy from 0.6648 to 0.6745. Additionally, the micro-accuracy improved from 0.6901 to 0.6886.

RQ5: The FastForestOva and FastTreeOva algorithms generated the best performance on both numerically encoded data and feature subsets. FastForestOva achieved a macro-accuracy of 67.45%.

The results obtained in this study can be extended to applications in the oil and gas sector by integrating software products that perform predictive analysis of operational safety. For example, ML models can be trained on historical data to anticipate the formation of cracks in the welded joints of pipelines. The direct consequence aims to prevent accidents and optimize processes related to maintenance programs.

The study highlights the importance of data preprocessing and feature selection in the model training stage, regardless of the nature of the fields in which they are applied. Future research will contribute to the development of applications in the oil and gas sector, and an initial article will review the works in the specialized literature that have contributed to this field through proposals of algorithms that integrate ML.

In conclusion, the research underscores the importance of data preprocessing and feature selection in improving the performance of ML models. The authors emphasize the necessity of correlating the pre-processing with the specifics of the problem. Additionally, identifying the algorithm with the best performance is also achieved by explicitly writing tests that allow for a comparative evaluation of their quality metrics. Although the results obtained for the models' performance are low, which renders the model unusable in practice, the dataset enabled the implementation of all tests that demonstrate the objectives set by the authors in this research.

REFERENCES

[1] Shaha T. R., Begum M., Uddin J., Torres V.Y., Iturriaga J. A., Ashraf I., Samad Md. A. Feature group partitioning: an approach for depression severity prediction with class balancing using machine learning algorithms, BMC Medical Research Methodology, vol. 24, issue 1, pp 1–18, 2024. https://doi.org/10.1186/s12874-024-02249-8

[2] Roșca C.M., Convergence Catalysts: Exploring the Fusion of Embedded Systems, IoT, and Artificial Intelligence. In Engineering Applications of AI and Swarm Intelligence, Yang, X.-S., Ed., Springer Nature, Singapore, pp 69-87, 2025. https://doi.org/10.1007/978-981-97-5979-8_4

[3] da Silva A.C.B., Santana R.C., de Lima, T.H.N., Teodoro M.L.M., Song M.A., Zárate L.E., Nobre C.N. A Review of the Main Factors, Computational Methods, and Databases Used in Depression Studies, In Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies, Vol, 5, pp 413-420, 2022. https://doi.org/10.5220/0010815800003123

[4] Mamidisetti S, Reddy A.M., Enhancing Depression Prediction Accuracy Using Filter and Wrapper-Based Visual Feature Extraction, Journal of Advances in Information Technology, Vol. 14, Issue 6, pp 1425-1435, 2023. https://doi.org/10.12720/jait.14.6.1425-1435

[5] Roșca C.-M., Bold R.-A., Gerea A.-E., A Comprehensive Patient Triage Algorithm Incorporating ChatGPT API for Symptom-Based Healthcare Decision-Making, In Proceedings of the Emerging Trends and Technologies on Intelligent Systems. ETTIS 2024. Lecture Notes in Networks and Systems, Noida, India, Vol. 1073, pp. 167-178, 2025. https://doi.org/10.1007/978-981-97-5703-9

[6] Haque U.M., Kabir E., Khanam, R., Insights into depression prediction, likelihood, and associations in children and adolescents: evidence from a 12-years study, Health Information Science and Systems, Vol. 13, Issue 1, pp 1–17, 2025. https://doi.org/10.1007/s13755-025-00335-9

[7] Tahsin M., Radovic A., Shaaban S., Doryab A., Predicting Depression in Adolescents Using Mobile and Wearable Sensors: Multimodal Machine Learning–Based Exploratory Study, JMIR Formative Research, Vol. 6, Issue 6, e35807, 2022. https://doi.org/10.2196/35807

[8] Zogan H., Razzak I., Wang X., Jameel S., Xu, G., Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media, World Wide Web, Vol. 25, Issue 1, pp 281–304, 2022. https://doi.org/10.1007/s11280-021-00992-2

[9] Roșca C.M. Comparative Analysis of Object Classification Algorithms: Traditional Image Processing Versus Artificial Intelligence – Based Approach, Romanian Journal of Petroleum & Gas Technology, Vol. IV (LXXV), Issue 2, pp 169-180, 2023. https://doi.org/10.51865/JPGT.2023.02.17

[10] Hong J., Kim J., Kim S., Oh J., Lee D., Lee S., Uh J., Yoon J., Choi, Y., Depressive Symptoms Feature-Based Machine Learning Approach to Predicting Depression Using Smartphone, Healthcare, Vol. 10, Issue 7, 1189, 2022. https://doi.org/10.3390/healthcare10071189

[11] Liu T., Meyerhoff J., Eichstaedt J.C., Karr C.J., Kaiser S.M., Kording K.P., Mohr D.C., Ungar L.H., The relationship between text message sentiment and self-reported depression, Journal of Affective Disorders, Vol. 302, pp 7–14, 2022. https://doi.org/10.1016/j.jad.2021.12.048

[12] Uddin M.Z., Dysthe K.K., Følstad A., Brandtzaeg, P.B., Deep learning for prediction of depressive symptoms in a large textual dataset, Neural Computing and Applications, Vol. 34, Issue 1, pp 721–744, 2021. https://doi.org/10.1007/s00521-021-06426-4

[13] Roșca, C.M., Ariciu, A.V. Unlocking Customer Sentiment Insights with Azure Sentiment Analysis: A Comprehensive Review and Analysis, Romanian Journal of Petroleum & Gas Technology, Vol. IV (LXXV), Issue 1, pp 173-182, 2023. https://doi.org/10.51865/JPGT.2023.01.15

[14] Roșca C.M., Stancu A.; Ariciu A.V. Algorithm for child adoption process using artificial intelligence and monitoring system for children, Internet of Things, Vol. 26, 101170, 2024. https://doi.org/10.1016/j.iot.2024.101170

[15] Amanat A., Rizwan M., Javed A.R., Abdelhaq M., Alsaqour R., Pandya S, Uddin M., Deep Learning for Depression Detection from Textual Data, Electronics, Vol. 11, Issue 5, 676, 2022. https://doi.org/10.3390/electronics11050676

[16] Steinmetz L.C.L., Sison M., Zhumagambetov R., Godoy J.C., Haufe S., Machine learning models predict the emergence of depression in Argentinean college students during periods of COVID-19 quarantine, Frontiers in Psychiatry, Vol. 15, 1376784, 2024. https://doi.org/10.3389/fpsyt.2024.1376784

[17] Roșca, C.-M., New Algorithm to Prevent Online Test Fraud Based on Cognitive Services and Input Devices Events, In Proceedings of Third Emerging Trends and Technologies on Intelligent Systems. ETTIS 2023. Lecture Notes in Networks and Systems, Noor, A., Saroha, K., Pricop, E., Sen, A., Trivedi, G., Eds., Springer Nature, Singapore, Vol. 730, pp 207-219, 2023. https://doi.org/10.1007/978-981-99-3963-3

[18] Hochman E., Feldman B., Weizman A., Krivoy A., Gur S., Barzilay E., Gabay H., Levy J., LevinkronO., Lawrence G., Development and validation of a machine learning-based postpartum depression prediction model: A nationwide cohort study, Depression and Anxiety, Vol. 38, Issue 4, pp 400–411, 2020. https://doi.org/10.1002/da.23123

[19] Bakkeli, N.Z. Predicting Psychological Distress During the COVID-19 Pandemic: Do Socioeconomic Factors Matter?, Social Science Computer Review, Vol. 44, Issue 4, pp 1227-1251, 2022. https://doi.org/10.1177_08944393211069622

[20] Roșca C.-M., Stancu A., Fusing Machine Learning and AI to Create a Framework for Employee Well-Being in the Era of Industry 5.0., Applied Sciences, Vol. 14, issue 23, 10835, 2024. https://doi.org/10.3390/app142310835

[21] Roșca C.-M., Rădulescu, G., Stancu A., Artificial Intelligence of Things Infrastructure for Quality Control in Cast Manufacturing Environments Shedding Light on Industry Changes, Applied Sciences, Vol. 15, issue 4, 2068, 2025. https://doi.org/10.3390/app15042068

[22] Janardhan N., Kumaresh, N., Improving Depression Prediction Accuracy Using Fisher Score-Based Feature Selection and Dynamic Ensemble Selection Approach Based on Acoustic Features of Speech, Traitement du Signal, Vol. 39, Issue. 1, pp 87-107, 2022. https://doi.org/10.18280/ts.390109

[23] Lu X., Shi D., Liu Y., Yuan, J., Speech depression recognition based on attentional residual network, Frontiers in Bioscience-Landmark, Vol. 26, Issue 12, pp 1746–1759, 2021. https://doi.org/10.52586/5066

[24] Roșca C.M., Gortoescu I.A., Tanase, M.R., Artificial Intelligence – Powered Video Content Generation Tools, Romanian Journal of Petroleum & Gas Technology, Vol. V (LXXVI), Issue 1, pp 131-144, 2024. https://doi.org/10.51865/JPGT.2024.01.10

[25] Liang L., Wang Y., Ma H., Zhang R., Liu R., Zhu R., Zheng Z., Zhang X., Wang, F., Enhanced classification and severity prediction of major depressive disorder using acoustic features and machine learning, Frontiers in Psychiatry, Vol. 15, 1422020, 2024. https://doi.org/10.3389/fpsyt.2024.1422020

[26] Hasanin T., Kshirsagar P.R., Manoharan H., Sengar S.S., Selvarajan S., Satapathy S.C. Exploration of Despair Eccentricities Based on Scale Metrics with Feature Sampling Using a Deep Learning Algorithm, Diagnostics, Vol. 12, Issue 11, 2844, 2022. https://doi.org/10.3390/diagnostics12112844

[27] Mao K., Zhang W., Wang D.B., Li A., Jiao R., Zhu, Y., Wu B., Zheng T., Qian L., Lyu W., Ye M., Chen, J., Prediction of depression severity based on the prosodic and semantic features with bidirectional LSTM and time distributed CNN, IEEE transactions on affective computing, Vol. 14, issue 3, pp 2251-2265, 2022. https://doi.org/10.1109/TAFFC.2022.3154332

[28] Nejadshamsi S., Karami V., Ghourchian N., Armanfard N., Bergman H., Grad R., Wilchesky M., Khanassov V., Vedel I., Abbasgholizadeh Rahimi S. (2025). Development and Feasibility Study of HOPE Model for Prediction of Depression Among Older Adults Using Wi-Fi-based Motion Sensor Data: Machine Learning Study, JMIR Aging, Vol. 8, e67715, 2025. https://doi.org/10.2196/67715

[29] Aljameel S. S., Alomari D. M., Alismail S., Khawaher F., Alkhudhair A. A., Aljubran F., Alzannan R. M. (2022). An Anomaly Detection Model for Oil and Gas Pipelines Using Machine Learning. Computation, Vol. 10, Issue 8, 138. https://doi.org/10.3390/computation10080138

[30] Wang Q., Song Y., Zhang X., Dong L., Xi Y., Zeng D., Liu Q., Zhang H., Zhang Z., Yan R., Luo H. (2023). Evolution of corrosion prediction models for oil and gas pipelines: From empirical-driven to data-driven. Engineering Failure Analysis, Vol. 146, 107097. https://doi.org/10.1016/j.engfailanal.2023.107097

[31] Kanoun Y., Mohammadi Aghbash A., Belem T., Zouari B., Mrad H. (2024). Failure prediction in the refinery piping system using machine learning algorithms: classification and comparison. Procedia Computer Science, Vol. 232, pp. 1663–1672. https://doi.org/10.1016/j.procs.2024.01.164

[32] Xu L., Wang Y., Mo L., Tang Y., Wang F., Li C. (2023). The research progress and prospect of data mining methods on corrosion prediction of oil and gas pipelines. Engineering Failure Analysis, Vol. 144, 106951. https://doi.org/10.1016/j.engfailanal.2022.106951

[33] Roșca C.-M., Cărbureanu M. (2025). A Comparative Analysis of Sorting Algorithms for Large-Scale Data: Performance Metrics and Language Efficiency. In: Noor A., Saroha K., Pricop E., Sen A., Trivedi G. (eds) Emerging Trends and Technologies on Intelligent Systems. ETTIS 2024. Lecture Notes in Networks and Systems, Vol. 1073. Springer, Singapore. https://doi.org/10.1007/978-981-97-5703-9_8

[34] Roșca C.-M., Stancu A. (2025). A Comprehensive Review of Machine Learning Models for Optimizing Wind Power Processes. Applied Sciences, Vol. 15, Issue 7, 3758. https://doi.org/10.3390/app15073758

[35] Kaagle Depression Professional Dataset with CC0 Public Domain license: https://www.kaggle.com/datasets/ikynahidwin/depression-professional-dataset/data

[36] Roșca C.-M., Cărbureanu M.A., A Comparative Analysis of Sorting Algorithms for Large-Scale Data: Performance Metrics and Language Efficiency, In Proceedings of the Emerging Trends and Technologies on Intelligent Systems. ETTIS 2024. Lecture Notes in Networks and Systems, Noida, India, Vol 1073, pp 99-113, 2025. https://doi.org/10.1007/978-981-97-5703-9_8

[37] Roșca C.-M., Stancu A., Fusing Machine Learning and AI to Create a Framework for Employee Well-Being in the Era of Industry 5.0, Applied Sciences, Vol. 14, Issue 23, 10835, 2024. https://doi.org/10.3390/app142310835

Received: April 2025; Revised: May 2025; Accepted: May 2025; Published: May 2025