Predictive Analysis and Model Evaluation for Stroke Prediction: A Comparative Study of Machine Learning Models

Mohammed Intishar Rahman, Shakeef Ahmed Rakin, Maisoon Tasnia, Afif Alamgir,
Mehnaz Ara Fazal, Ehsanur Rahman Rhythm and Annajiat Alim Rasel
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
Brac University
66 Mohakhali, Dhaka - 1212, Bangladesh
{mohammed.intishar.rahman, shakeef.ahmed.rakin, maisoon.tasnia, afif.alamgir, mehnaz.ara.fazal,
ehsanur.rahman.rhythm}@g.bracu.ac.bd, annajiat@gmail.com

***Abstract*—This paper examines the performance of four common machine learning models - K-Nearest Neighbors (KNN), Random Forest, Logistic Regression, and Decision Tree - on datasets containing both integer and categorical features. We aim to identify the most effective model for such scenarios through a comprehensive comparative analysis. The study utilizes a stroke prediction dataset as a case study, highlighting the challenges of mixed-feature analysis in real-world applications. We evaluate each model's accuracy, precision, recall on this diverse dataset, analyzing performance on both integer and categorical features individually and in combination. We identify potential biases and limitations for each type of feature and explore how feature scaling and pre-processing techniques influence model effectiveness. Furthermore, we investigate the impact of feature selection on each model's performance, providing insights into optimizing their application to mixed-feature datasets. This analysis offers valuable recommendations for data scientists and practitioners facing similar challenges, aiding in the selection and optimization of appropriate models for diverse tasks beyond just stroke prediction.**

***Index Terms*—Machine Learning, Mixed-Feature Data, Performance Evaluation, KNN, Random Forest, Logistic Regression, Decision Tree, Feature Scaling, Feature Selection**

I. Introduction

The rise of complex datasets in various fields has pushed the boundaries of data analysis, demanding robust and adaptable machine learning (ML) models. These datasets often contain a diverse mix of features, including numerical values (integer and continuous) and categorical data (nominal and ordinal). Effectively handling this "mixed-feature" landscape presents unique challenges for model selection and optimization.

Traditional ML models often struggle with mixed-feature data, exhibiting biases towards specific feature types or failing to capture the intricate relationships between integers and categorical variables. This can lead to suboptimal performance, hindering accurate predictions and actionable insights. Therefore, it becomes crucial to identify and understand the strengths and weaknesses of different ML models when applied to such diverse datasets.

This paper undertakes a comprehensive comparative analysis of four popular ML models - K-Nearest Neighbors (KNN), Random Forest, Logistic Regression, and Decision Tree - on a mixed-feature dataset focused on stroke prediction. We aim to answer the following questions:

- Which ML model performs best on mixed-feature datasets, considering accuracy, precision, recall?
- How do these models handle integer and categorical features individually and in combination? Are there any potential biases or limitations associated with specific feature types?
- What impact do feature scaling and pre-processing techniques have on model performance?
- How can feature selection optimize each model's effectiveness on mixed-feature data?

By addressing these questions, we aim to provide valuable insights for data scientists and practitioners working with mixed-feature datasets. Our findings will offer practical recommendations for selecting and optimizing appropriate ML models for diverse tasks beyond stroke prediction, promoting robust and reliable analysis in a variety of fields.

II. Literature Review

Recently, there has been increased interest in the use of machine learning to forecast Alzheimer's disease; nevertheless, investigations have revealed conflicting results and problems with model accuracy. In [1], using an Alzheimer's MRI dataset, this study compares and contrasts four well-known models for diagnosis: random forests, logistic regression, SVM, and decision trees. Grid search was used to fine-tune the models and identify the ideal parameters. Important results indicate that, after tuning, SVM with an RBF kernel outperforms other models, achieving the greatest accuracy of 92%. Random forests and decision trees experienced some overfitting. The comparative method highlights the significance of model selection and hyperparameter optimization, while the SVM model shows promise in predicting Alzheimer's from

scan data so patients can seek early treatment. Future research ought to examine more datasets and ensemble methods to improve reliability even more. This is a solid reference for applying machine learning in Alzheimer's prediction using medical imaging data.

ANCOVA, logistic regression, support vector regression, decision trees, random forests, log-linear regression, and partial least squares regression are the seven supervised machine learning models that are evaluated in [2] using a dataset of student performance. Comparing algorithms and determining what influences academic success are the goals. Before and after hyperparameter adjustment, models are evaluated using MSE, RMSE, and R-squared metrics. The main conclusions indicate that log-linear regression works best, having the lowest error and highest accuracy. The data also shows that behavioral traits that have a substantial impact on results include participating in class and finishing assignments. The comparison method shows that accurate predictions depend on carefully choosing and fine-tuning the model. When it comes to applying machine learning in the field of education, this is a useful resource.

The authors of [3] compares several supervised machine learning techniques—such as support vector machines, neural networks, and logistic regression—for identifying fraudulent transactions in blockchain networks. Time between transactions and account balances have been proven to be relevant features in prior publications that employed methods such as clustering, LSTM networks, XGBoost classifiers, etc. for fraud detection. Eight machine learning models are tested for classification performance in this paper utilizing bootstrap sampling and accuracy. The accuracy of the top three models, Random Forests, SVM, and AdaBoost, is 97% for each. For the final fraud categorization, the paper suggests utilizing an ensemble of these best models. The results set an accuracy baseline for blockchain fraud detection using supervised learning, even though unsupervised and deep learning methods are not tested. Incorporating unsupervised approaches and testing on private blockchains are among the upcoming tasks. All things considered, the research offers a thorough comparison of the main machine learning techniques for fraud detection in blockchain transaction networks.

III. Methodology

A. Dataset and Features

The stroke prediction dataset from Kaggle [4] was used, containing 5110 patient entries with 11 features. Features include both integer data (age, average glucose level, BMI) and categorical data (gender, hypertension, work type, etc.). The challenge lies in handling the mixed-feature nature of the data and ensuring model compatibility.

B. Preprocessing Techniques

- Missing values were imputed using the Simple Imputing technique from Sklearn, considering the median value for numerical features and appropriate strategies for categorical data.

```
[7] from sklearn.impute import SimpleImputer
    imputer = SimpleImputer(missing_values=np.nan, strategy= 'median')
    imputer.fit(dataset[['bmi'] ])
    dataset[['bmi']]= imputer.transform(dataset[['bmi']])
```

Fig. 1. Imputing Missing Values.

- OneHotEncoder was applied to encode categorical features (columns 5, 6, and 9) into numerical representations for model compatibility.

```
[15] from sklearn.compose import ColumnTransformer
     from sklearn.preprocessing import OneHotEncoder

     ct = ColumnTransformer(transformers=[('encoder',OneHotEncoder(),[5,6,9])] , remainder='passthrough' )
     x= np.array(ct.fit_transform(x))
```

Fig. 2. Encoding Categorical Features.

- Label encoding was used for binary features (columns B and F) to convert them into numerical values.

```
[13] from sklearn.preprocessing import LabelEncoder
     le = LabelEncoder()
     dataset.gender= le.fit_transform(dataset.gender.values)
     dataset.ever_married= le.fit_transform(dataset.ever_married.values)
```

Fig. 3. Encoding For Binary Features.

- Features and labels were split into separate training (70%) and test (30%) sets for model evaluation.

```
16] from sklearn.model_selection import train_test_split
    x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.3, random_state = 0)
```

Fig. 4. Dataset Splitting.

C. Model Descriptions

- KNN: K-Nearest Neighbors, a non-parametric algorithm, classifies new data points based on their similarity to existing neighbors. It was chosen for its simplicity and ability to handle mixed features without assumptions about the data.
- Logistic Regression: This probabilistic model predicts binary outcomes based on the relationships between features and the label. Its suitability for mixed features lies in its ability to handle both linear and non-linear relationships.

- Decision Tree: This tree-based model makes decisions based on a series of feature comparisons, leading to a leaf node with a predicted class. Its strength lies in its ability to handle categorical features naturally and identify complex interactions between features.
- Random Forest: This ensemble method combines multiple decision trees to improve overall accuracy and reduce overfitting. Its advantage in mixed-feature scenarios is its robustness to outliers and its ability to handle both numerical and categorical data.

D. Performance Evaluation

Model performance will be evaluated using standard metrics for binary classification, including accuracy, precision, recall, and training time. The metrics will be compared across models to identify the one with the best overall performance on the mixed-feature dataset. Statistical tests (e.g., t-test) may be used to assess the significance of differences in performance between models.

E. Experimental Setup

The training/test split ratio of 70/30 ensures sufficient data for training while providing a representative test set for unbiased evaluation. Cross-validation techniques may be used to further assess model generalizability and robustness.

IV. Data Analysis

The dataset contains information on 5110 patients with 11 features including demographic details, health parameters, and stroke outcome. The features are a mix of numerical and categorical data. There are no missing values. The data was divided into 70% training set and 30% test set for modeling.

Key observations from exploratory data analysis:

- Imbalanced dataset with 28.9% patients who had a stroke and 71.1% who did not.
- There is a good mixture of integer and categorical data types.

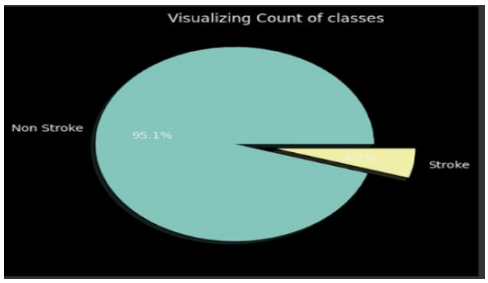

Fig. 5. Dataset Visualizations.

V. Prototype Implementation

Four machine learning models were trained on the training data and evaluated on the test data. The models were implemented using scikit-learn library in Python. Key hyperparameters like number of neighbors for KNN and number of trees in random forest were tuned through grid search cross-validation to maximize accuracy. We'll be using the metrics such as accuracy, precision, recall, F1-score and confusion matrix for evaluating the models.

VI. Results

The table below summarizes model performance:

Model	Accuracy
KNN	93.74%
Logistic regression	95.04%
Decision Tree	91.65%
Random forest	94.95%

TABLE I
Model Accuracies

Logistic regression achieved the highest accuracy of 95.04% on test data, proving to be an efficient linear model for stroke prediction.

- Non-linear models KNN and Random forest also performed well with over 93% accuracy owing to their capability to capture complex relationships.
- Decision trees were found to be less accurate than ensemble models like random forest highlighting the importance of combining multiple decision trees.

The study indicates traditional machine learning algorithms can predict stroke effectively given clinical data. Further optimizations like handling class imbalance can improve model performance. The prediction system can identify patients prone to stroke and enable preventive interventions.

VII. Limitations

While this analysis demonstrates promising results for using machine learning to predict stroke risk, there are some limitations:

- The dataset only comprises 11 parameters. Including detailed medical history, scans and risk factor data could improve predictions.
- Class imbalance with greater samples of non-stroke patients can bias the models.
- Only traditional machine learning models were explored. Advanced Deep Learning approaches were not implemented.
- Model interpretability is limited. It is unclear which parameters are most indicative of stroke prediction for each algorithm.
- Evaluation uses accuracy metric. Using precision and recall would allow assessing real world implications better.
- Predictions need prospective validation in a clinical workflow before full scale deployment.

VIII. Future Plans

Some recommendations to address the limitations and further improve stroke prediction performance:

- Collect more patient data including temporal history to account for greater risk factors.

- Use resampling techniques like SMOTE to handle class imbalance.
- Implement Deep Neural Network architectures like LSTM and CNN that can learn data representations.
- Perform feature importance analysis to understand prediction attributes.

Addressing these limitations can enhance model performance and move towards deployable AI solutions for stroke prediction that enable improved patient outcomes.

IX. Conclusion

This study presented a comprehensive comparative analysis of four popular machine learning models - K-Nearest Neighbors, Random Forest, Logistic Regression, and Decision Tree - for stroke prediction using a mixed-feature dataset. The key findings indicate that logistic regression achieved the highest accuracy of 95.04%, proving to be an efficient linear model for this task. Non-linear models like KNN and Random Forest also performed well, demonstrating the capability to capture complex relationships in the data. The relatively lower accuracy of decision trees highlights the value of ensemble approaches that combine multiple models.

Overall, the analysis shows the potential of traditional machine learning algorithms for effective stroke risk prediction based on clinical parameters. The prediction system can identify patients prone to stroke and enable preventive medical interventions. However, there are some limitations regarding data size, class imbalance, model interpretability, and evaluation metrics which need to be addressed. Future work should focus on collecting more patient data, using techniques to handle skew, implementing advanced deep learning architectures, performing feature importance analysis, and validating predictions prospectively before full-scale clinical deployment.

Addressing these limitations can further optimize model performance and pave the path towards deployable AI solutions for stroke prediction that enable improved patient outcomes. The study provides a useful framework for selecting and evaluating machine learning models on mixed-feature medical datasets. The methodology and findings can guide healthcare practitioners and data scientists working with diverse clinical data to predict various diseases. Overall, this comparative analysis offers valuable practical recommendations beyond stroke, promoting the adoption of robust analytics in healthcare.

References

[1] Kang, J., Bari Antor, M., Jamil, A. H. M. S., Mamtaz, M., Monirujjaman Khan, M., Aljahdali, S., Kaur, M., Singh, P., & Masud, M. (2021). A Comparative Analysis of Machine Learning Algorithms to Predict Alzheimer's Disease. Journal of Healthcare Engineering, 2021, 9917919. https://doi.org/10.1155/2021/9917919

[2] El Guabassi, I., Bousalem, Z., Marah, R., & Qazdar, A. (2021, February). Comparative Analysis of Supervised Machine Learning Algorithms to Build a Predictive Model for Evaluating Students' Performance. International Association of Online Engineering. Retrieved from https://www.learntechlib.org/p/218992

[3] Pandey, S., & Sharma, S. (2023). A comparative study of retrieval-based and generative-based chatbots using Deep Learning and Machine Learning. Healthcare Analytics, 3, 100198. https://doi.org/10.1016/j.health.2023.100198

[4] Federoriano, F. (2020). Stroke prediction dataset. Retrieved from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/