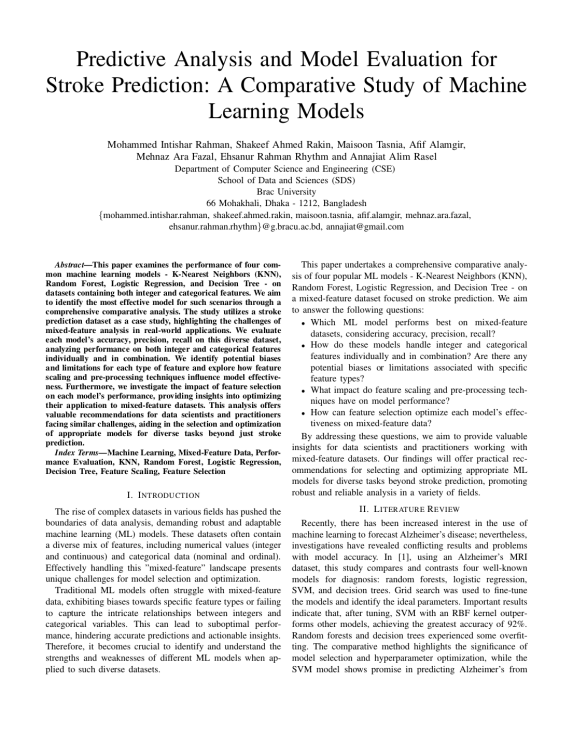
Digital Receipt

This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

The first page of your submissions is displayed below.

Submission author:	AFIF ALAMGIR
Assignment title:	part6
Submission title:	Task 6 438
File name:	output_35.pdf
File size:	161.87K
Page count:	4
Word count:	2,009
Character count:	12,285
Submission date:	14-Dec-2023 09:29AM (UTC+0530)
Submission ID:	2258541702



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