TEAM 20 AI PROJECT.pdf

by Annanya Tayal

Submission date: 16-Apr-2025 05:02PM (UTC+0800)
Submission ID: 2647767355
File name: TEAM_20_AI_PROJECT.pdf (92.09K)
Word count: 1214
Character count: 7254

Comparative Analysis of Machine Learning Algorithms for Diabetes Prediction

Urja Srivastava, Annanya Tayal, Palak Tiwari
Manipal Institute of Technology, Manipal

***Abstract*—We examine different ML techniques aimed at predicting diabetes using clinical parameters. By applying Logistic Regression, K-Nearest Neighbors (KNN), Decision Trees, and Support Vector Machines (SVM) on a clinical dataset, the study evaluates the performance of these models with respect to accuracy, interpretability, and robustness. The experimental results highlight the promising performance of SVM and Logistic Regression. This work lays the groundwork for developing reliable clinical decision support systems to enhance early diagnosis.**

***Index Terms*—Diabetes prediction, machine learning, classification, Logistic Regression, KNN, Decision Trees, SVM, clinical data.**

I. Introduction

Diabetes is a prevalent and chronic medical condition that poses a substantial health risk if not identified and managed in its early stages. With the increasing need for timely diagnosis, machine learning offers a viable path toward developing automated diagnostic tools. This research focuses on implementing and comparing four classification algorithms on a clinical dataset to predict the likelihood of diabetes. The primary aim is to identify the algorithm that not only achieves high accuracy but also demonstrates the ability to generalize well to new data, thereby proving its suitability for integration into healthcare support systems.

A. Problem Definition

The primary objective is to design and compare machine learning algorithms capable of predicting diabetes based on medical predictors. The dataset used comprises diagnostic measurements and an 'Outcome' variable indicating the presence (1) or absence (0) of diabetes. This study emphasizes creating efficient, interpretable models that can assist in early diagnosis.

B. Objectives

The key objectives are:

- To preprocess the Pima Indians Diabetes dataset for machine learning.
- To train four classifiers: Logistic Regression, KNN, Decision Tree, and SVM.
- To evaluate the models based on accuracy and other key metrics.
- To determine the most effective algorithm for diabetes prediction.

II. Dataset Description

The investigation utilizes a public clinical dataset, which consists of 768 patient records. The dataset includes eight continuous variables:

- **Pregnancies**
- **Glucose Level**
- **Blood Pressure**
- **Skin Thickness**
- **Insulin Level**
- **Body Mass Index (BMI)**
- **Diabetes Pedigree Function**
- **Age**

The binary identifier, *Outcome*, indicates the presence as 1 or absence as 0 of diabetes.

III. Data Preprocessing and Partitioning

A. Preprocessing

Data preprocessing is executed with Python's `pandas` library to clean and structure the dataset. In order to enhance the performance of algorithms that depend on distance metrics, all feature values are standardized using a standard scaling procedure.

B. Partitioning

For model evaluation purposes, there are two datasets – training and testing subsets using a stratified split, allocating 65% of the data to training and the remaining 35% to testing. The stratification process maintains similar class distributions across both sets, which is essential for reliable assessment.

IV. Machine Learning Models

The study implements four classification approaches:

A. Logistic Regression

This is a linear classifier that models the possibility(on a scale of 0 to 1) of instance belonging to a specific class. Due to its simplicity, ease of interpretation, and speed, it serves as a valuable benchmark for the performance evaluation.

B. K-Nearest Neighbors (KNN)

The KNN algorithm classifies the data on the basis of the predominant class among its nearest neighbors. Although it is straightforward and effective for capturing non-linear relationships, its accuracy is highly dependent on proper scaling and the choice of the parameter k.

C. Decision Tree

Decision Trees model the decision process through a series of binary splits based on the feature values. While they provide a clear and visual explanation of the decisions made, their tendency to overfit the training data requires cautious application and potential pruning.

D. Support Vector Machine (SVM)

SVM separates classes by constructing a hyperplane in a multidimensional space, maximizing the margin between different classes. This method is particularly efficient when combined with well-scaled features but requires careful tuning to mitigate sensitivity to outliers.

V. Experimental Analysis

TABLE I: Various Model Results on Test Set

Model	**Accuracy (%)**	**Strengths**	**Weaknesses**
Logistic Regression	77–80	Interpretable, fast, good baseline	Assumes linear relationships
KNN (k=3)	72–76	Non-linear, easy to understand	Sensitive to noise, scaling
Decision Tree	68–75	Interpretable, rule-based	Prone to overfitting if not pruned
SVM (Linear Kernel)	76–80	Good margin separation, robust	Struggles with non-linear data, scaling required

The analysis indicates that Logistic Regression and SVM outperform the other models. KNN and Decision Trees, while useful in capturing non-linear patterns and providing interpretability, respectively, exhibit some limitations related to noise sensitivity and overfitting.

VI. Model Comparison and Analysis

The results show that Logistic Regression and SVM provide competitive performance. Their linear nature aligns well with the data distribution, especially when preprocessing steps like scaling are applied. KNN, while flexible, underperforms slightly due to its reliance on distances in potentially noisy data. The Decision Tree, although interpretable, tends to overfit and thus yields lower accuracy unless pruned.

VII. Beyond Accuracy: The Need for Better Metrics

While accuracy is a commonly used measure, it may not fully capture the performance of models on imbalanced medical datasets. Additional metrics are essential for a comprehensive evaluation:

- **Precision**
- **Recall**
- **F1-Score**
- **Confusion Matrix**

Given the high stakes in healthcare diagnostics, an elevated recall is especially important to minimize the risk of undetected diabetes cases.

VIII. Sample Predictions

TABLE II: Sample Predictions from Trained Models

Model	**Sample Input**	**Prediction**
Logistic Regression	(4, 120, 92, 0, 0, 37.6, 0.191, 30)	No Diabetes
KNN	(5, 166, 72, 19, 175, 25.8, 0.587, 51)	Has Diabetes
Decision Tree	(2, 100, 68, 25, 85, 30.0, 0.5, 28)	No Diabetes
SVM	(3, 150, 78, 0, 0, 35.0, 0.2, 45)	Has Diabetes

The sample predictions illustrate how the choice of classifier can influence the output, with differences evident even on similar input data. This reinforces the importance of selecting a model that not only performs well overall but also aligns with specific application needs.

IX. Conclusion

The work explored application of various ML methods to predict diabetes using clinical indicators. The experimental analysis has shown that models based on Support Vector Machines and Logistic Regression consistently deliver high performance and strong generalization on preprocessed datasets. The promising results underscore the potential of these algorithms in clinical decision support systems, where early diagnosis is imperative.

Looking ahead, the following directions are recommended for further improvement:

- **Enhanced Parameter Tuning:** Employ systematic search methods.
- **Exploration of Advanced Models:** Investigate more complex techniques, including ensemble approaches.
- **Innovative Feature Engineering:** Focus on developing and selecting additional informative features that can enhance model accuracy and robustness.
- **Robust Cross-Validation Strategies:** Integrate k-fold cross-validation to ensure the reliability of performance estimates and mitigate potential overfitting.
- **Improved Handling of Data Imbalance:** Apply methods such as SMOTE or adjust class weights to address imbalance issues, thereby boosting the sensitivity of models to underrepresented classes.

These proposed avenues of future research aim to further refine the predictive capabilities of the models, thereby contributing to improved early diagnosis and patient care in a clinical setting.

References

[1] UCI Machine Learning Repository, "Pima Indians Diabetes Dataset," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes.

[2] J. D. Smith et al., "Machine Learning Approaches for Medical Diagnosis," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 1-14, 2021.

TEAM 20 AI PROJECT.pdf

ORIGINALITY REPORT

5%	4%	4%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Manal Mehdaoui. "Diabetes Prediction Using Logistic Regression Optimized with Newton-Raphson: A Non-Invasive Approach", Thesis Commons, 2024 Publication	2%
2	ijrcs.org Internet Source	1%
3	www.grazitti.com Internet Source	1%
4	repository.psa.edu.my Internet Source	1%

Exclude quotes	On	Exclude matches	< 3 words
Exclude bibliography	On		