Kuwait Journal of Computer Science
ISSN: 2984-7605
Vol 3 Issue 1 (2025)

Diabetes Prediction Using Machine Learning

[1]Nitin Mukesh Joshi, [2] Jai Bhaskar

[1]Researcher, Bikaner Technical University, Bikaner.

nikunjjoshi8@gmail.com

[2]Assistant Professor, Bikaner Technical University, Bikaner.

jaibhaskar@cet-gov.ac.in

Abstract:

Diabetes is a chronic condition with elevated blood sugar levels, leading to severe complications if undiagnosed or untreated. The diagnosis process often involves patient referrals and multiple consultations. Predictive analytics in healthcare offers potential for timely decision-making based on patient data.This study aims to develop an accurate model for predicting diabetes using machine learning. The dataset is split into training, validation, and testing sets, with each set serving a distinct purpose. Various classification algorithms are employed, including Logistic Regression, K Nearest Neighbors, Gaussian Naive Bayes, Support Vector Machines, and more.The study uses the Pima Indians Diabetes Database (PIDD), which contains diagnostic measures for diabetes detection. Performance metrics such as Precision, Accuracy, Specificity, and Recall are calculated using confusion matrix analysis.By comparing algorithm performance, the study identifies the best approach for early diabetes detection. The goal is to assist healthcare professionals in diagnosing diabetes sooner, improving patient outcomes, and reducing complications.

Keywords: reducing, algorithm, Regression

Introduction: Diabetes mellitus is a global health concern, with millions of individuals affected annually. Its prevalence has escalated due to lifestyle changes, making early diagnosis a critical step in preventing associated complications such as cardiovascular disease, neuropathy, and retinopathy. Traditional diagnostic methods often require laboratory testing, which can be time-consuming and expensive. Machine learning (ML) offers a promising alternative by leveraging patient data to predict the likelihood of diabetes efficiently.

This research focuses on using ML techniques to build a predictive model that aids in early diabetes detection. The study leverages the Pima Indians Diabetes Database (PIDD), a widely used dataset in healthcare predictive modeling. The research objectives include evaluating multiple ML algorithms, comparing their performance, and identifying the optimal approach for diabetes prediction.

Methodology :

The research methodology involves the following detailed steps:

1. **Data Collection :**

In order to gather relevant and reliable data, the project utilizes reputable sources such as pima Indians diabetes database (PIDD).

2. **Data Preprocessing:**

○ **Data Cleaning:**

▪ Data integrity can be preserved by handling missing values with methods like median replacement or mean imputation.

▪ Removing outliers based on statistical analysis or domain knowledge to ensure data quality.

- **Feature Scaling:**
  - Normalizing continuous variables to a standard range (e.g., 0 to 1) for ensuring uniformity across features.
  - Standardizing data for attaining 0 mean

    along with 1standard deviation for algorithms sensitive to feature scaling.
- **Feature Engineering:**
  - New features have been created or existing ones have been modified for representing relationships in dataset better.

3. **Dataset Splitting:**
   - Dividing dataset into validation (10%), training (80%) as well as testing (10%) subsets for ensuring unbiased performance evaluation.
   - Utilizing stratified splitting to maintain the proportion of positive and negative cases across subsets.
4. **Model Development:**
   - Implementing and fine-tuning various machine learning classifiers:
     - **Logistic Regression:** For binary classification, a linear model is appropriate.
     - **Gaussian Naive Bayes:** a Bayes theorem-based probabilistic model.
     - **K Nearest Neighbors (KNN):** A distance-based algorithm for classification.
     - **Support Vector Machines (SVM):** a classification model that determines optimal hyperplane.
     - **Decision Trees:** a model based on trees that divides data according to feature thresholds.
     - **Random Forests:** Decision tree ensemble for improving robustness and accuracy.
     - **Bagging Classifier:** Ensemble method that reduces variance by averaging predictions.
     - **AdaBoost Classifier:** A boosting algorithm that combines weak learners sequentially.
     - **Gradient Boosting Classifier:** An advanced boosting technique that optimizes model performance by reducing loss incrementally.
5. **Hyperparameter Tuning:**
   - Utilizing grid search or randomized search for systematic exploration of hyperparameter combinations for each algorithm.
   - Evaluating models on the validation set to select the best hyperparameter configuration.
   - Tuned hyperparameters examples include learning rate in Gradient Boosting, number of trees in Random Forests, along with regularization parameters in SVM.
6. **Model Evaluation:**

   Employing the testing dataset to evaluate the final model's performance.

   Metrics derived from the Confusion Matrix include:

   - **Accuracy:** “Proportion of correct predictions out of all total predictions.
   - **Precision:** Proportion of true positive predictions out of all positive predictions.
   - **Recall (Sensitivity):** Proportion of actual positives correctly identified.
   - **Specificity:** Proportion of actual negatives correctly identified.

   Comparing algorithms based on these metrics to determine the best-performing model” for diabetes prediction.

7. **Cross-Validation:**

Implementing k-fold cross-validation to assess model generalization by splitting the data into k subsets and iteratively training and testing the model.

System Development :

1. Data Processing

The initial step involves preparing the data for analysis. This includes:

- **Data Cleaning:** Missing values are imputed using median replacement, while outliers are identified and managed through statistical techniques.
- **Feature Scaling:** Continuous variables are normalized to a standard range for ensuring consistency across features.
- **Feature Engineering:** New attributes are derived to enhance the representation of underlying patterns in the dataset, and irrelevant features are eliminated.

2. Model Training :

The training process focuses on building and optimizing machine learning models:

- **Algorithm Selection:** Multiple classifiers including Random Forest, Logistic Regression, as well as Gradient Boosting are implemented.
- **Training Process:** Dataset has been divided into validation as well as training subsets, with former used for fitting model parameters.
- **Hyperparameter Tuning:** Grid search as well as random search methods have been employed for finding optimal hyperparameters that maximize model performance.
- **Cross-Validation:** k-fold cross-validation makes sure that the trained models generalize well for unseen data.

3. Result Visualization

Post-training, performance of the models has been visualized for interpreting their effectiveness:

- **Confusion Matrix Analysis:** Key metrics including Specificity, Precision, Accuracy, along with Recall have been derived.
- **Graphical Representations:** Performance metrics are depicted using bar charts, ROC curves, and precision-recall plots for easy comparison.
- **Insights:** The visual analysis guides the selection of the best-performing algorithm for deployment.

4. System Integration

The prediction system integrates multiple components for seamless operation:

- **Data Input Interface:** A module that allows users to upload patient data in a predefined format.
- **Prediction Module:** The selected model generates real-time diabetes predictions based on input data.
- **Data Flow:** A pipeline connects preprocessing, feature engineering, and prediction steps, ensuring efficient data movement.

5. Testing and Refinement

Comprehensive testing ensures the reliability of the system:

- **Unit Testing:** Each module (e.g., preprocessing, model training) is individually tested for functionality.
- **Integration Testing:** Interaction between modules is validated to ensure consistent data flow.

- **Model Evaluation:** The testing subset is used to assess the system's predictive power, refining the model based on accuracy, precision, recall, and specificity.

6. Deployment and Maintenance

The final stage involves making the system operational and ensuring its sustainability:

- **Deployment:** The model is deployed in a real-world setting, with a user-friendly interface for healthcare professionals to input patient data.
- **Monitoring and Updates:** System performance is continuously monitored, and the model is retrained periodically with updated data to maintain accuracy.
- **Scalability:** Provisions are made to accommodate additional datasets and users as the system expands.

Technology and Tools Used :

The "**Diabetes Prediction Using Machine Learning**" uses diverse technologies along with tools for supporting data processing, model training, results visualization, and process development These technologies and tools contribute to the efficient and effective implementation of the project. Some of the major technologies and tools used in this field are:

1. Programming Language: Python
2. Data Manipulation and Analysis: Pandas, NumPy
3. Machine Learning and Deep Learning: scikit- learn, Keras
4. Data Visualization: Matplotlib
5. Data Sources: Pima Indians diabetes database (PIDD)
6. Web Development: Flask
7. Web Technologies: HTML, CSS

Python serves as primary programming language for implementing various components of the project. NumPy along with Pandas are used for optimal data manipulation as well as analysis, while scikit-learn and Keras are employed for ML tasks, specifically for implementing the prediction models,

Matplotlib is utilized for data visualization, allowing the creation of charts and graphs to analyze historical data and visualize the predictions.

The Flask web framework is used for developing the interactive web application, enabling users to input stock symbols and visualize the predicted prices. HTML along with CSS have been utilized to structure as well as style web interfaces, to ensure user-friendly and visually appealing experience.

Results and Discussion :

Our models have been trained, and a table summarizing the metrics of the different models is presented. Objectives include,

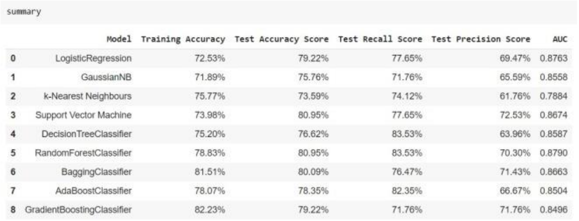
1) To check if any information can further be extracted from data for establishing correlation among diabetes as well as parameters.

2) To use a variety of supervised learning machine learning algorithms in an effort to obtain the highest accuracy score.

Although we can determine that there is a positive correlation among glucose levels as well as diabetes based on hypothesis test, we are unable to establish causality. “Based on the comparison of the several algorithms used, Random Forest appears to have the best results.

Developing a model that can accurately forecast a patient's likelihood of having diabetes is the goal of this study”. This project has successfully accomplished its primary goal of designing and implementing “diabetes prediction using machine learning techniques and performance analysis.

Several classification and” ensemble learning techniques are employed in the suggested method, including KNN, Gradient Boosting, Logistic Regression, Decision Tree, Random Forest, as well as SVM classifiers. For creating a framework that reduces human labour and yields dependable outcomes, a machine learning algorithm must be employed.



summary

	Model	Training Accuracy	Test Accuracy Score	Test Recall Score	Test Precision Score	AUC
0	LogisticRegression	72.53%	79.22%	77.65%	69.47%	0.8763
1	GaussianNB	71.89%	75.76%	71.76%	65.59%	0.8558
2	k-Nearest Neighbours	75.77%	73.59%	74.12%	61.76%	0.7884
3	Support Vector Machine	73.98%	80.95%	77.65%	72.53%	0.8674
4	DecisionTreeClassifier	75.20%	76.62%	83.53%	63.96%	0.8587
5	RandomForestClassifier	78.83%	80.95%	83.53%	70.30%	0.8790
6	BaggingClassifier	81.51%	80.09%	76.47%	71.43%	0.8663
7	AdaBoostClassifier	78.07%	78.35%	82.35%	66.67%	0.8504
8	GradientBoostingClassifier	82.23%	79.22%	71.76%	71.76%	0.8496

Table: Comparative Analysis Table

Several models' test accuracy falls between 73% as well as 81%, which is same range. Random Forest Classifier yielded best results based on 2metrics: accuracy along with recall.

Conclusion :

- Diabetes prediction could be revolutionized by machine learning with support of advanced computational techniques.
- Early-stage diabetes detection is key for treatment.
- This approach might also assist researchers in creating an accurate as well as effective tool for assisting clinicians to make better decisions regarding disease.
- Future scope of this project would include more factors as well as parameters.
- As parameters are increased, the accuracy will rise even further.
- Accuracy can be improved through data refinement utilizing algorithms along with traditional approaches.

References:

1. https://www.kaggle.com/uciml/pima-indians-diabetes-database
2. Diabetes Prediction using Machine Learning Algorithms - ScienceDirect
3. Predicting Diabetes Mellitus With Machine Learning Techniques (nih.gov)
4. Diabetes Prediction using Machine Learning Techniques – IJERT
5. https://www.researchgate.net/publication/339543101_Diabetes_Prediction_using_Machi ne_Learning_Algorithms