## REPORT ON

## Report: Predicting Diabetes Using Machine Learning Models

**M. Rupa naga Lakshmi**

A report submitted in partial fulfilment of the requirements for the degree of **BACHELOR OF TECHNOLOGY**

**Branch: COMPUTER SCIENCE AND ENGINEERIGN**

**Specialisation: AIML**

of Alliance University



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

# ALLIANCE COLLEGE OF ENGINEERING AND DESIGN

ALLIANCE UNIVERSITY, BENGALURU

**BACHELOR OF TECHNOLOGY**

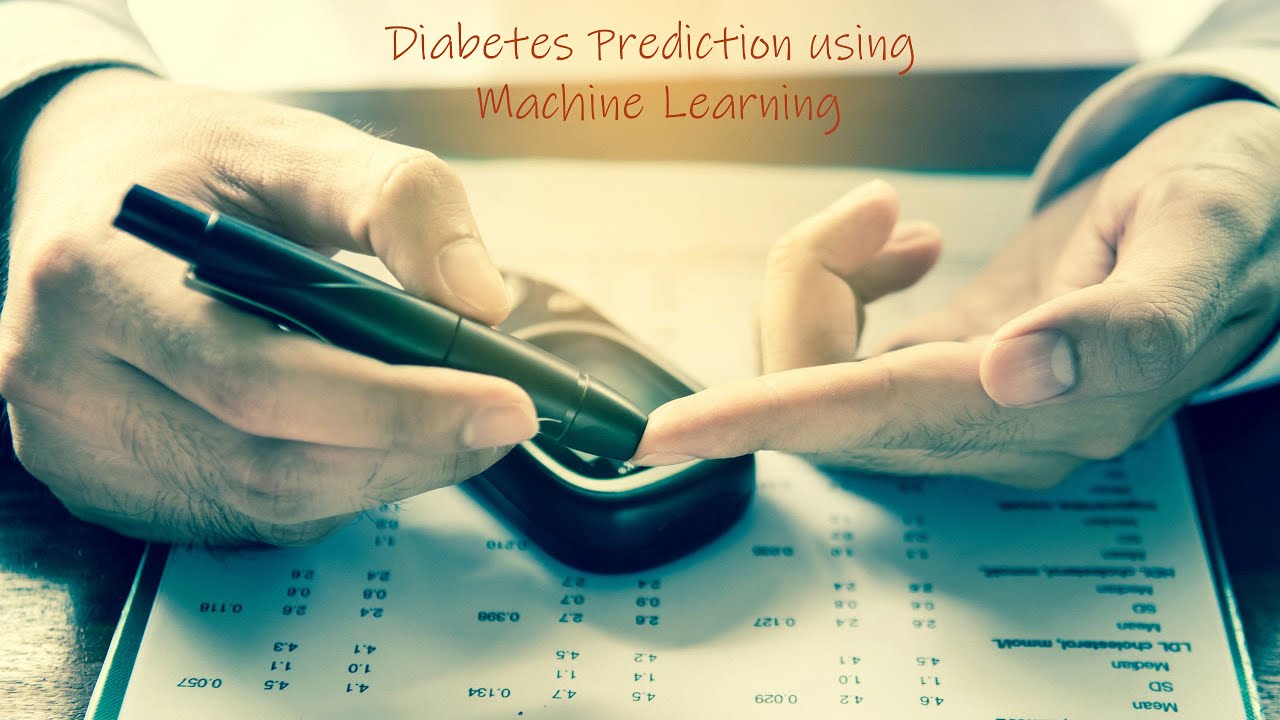
**Branch: COMPUTER SCIENCE AND ENGINEERING**

**Specialization: AIML** Of Alliance University

**2024**

**ABSTRACT**

This report presents an analysis of different machine learning models for predicting diabetes using the Pima Indians Diabetes dataset. The dataset underwent preprocessing, including handling missing values and normalization. We evaluated multiple models such as Logistic Regression, K-Nearest Neighbors, Random Forest, Support Vector Machine, Naive Bayes, AdaBoost, and Gradient Boosting. The models' performance was compared based on accuracy, precision, recall, and F1-score. Gradient Boosting showed the highest overall performance, suggesting its potential for reliable diabetes prediction. The data underwent preprocessing steps including handling missing values, feature scaling, and normalization. We implemented and evaluated multiple machine learning models, namely Logistic Regression, K-Nearest Neighbours (KNN), Random Forest, Support Vector Machine (SVM), Naive Bayes, AdaBoost, and Gradient Boosting.Each model was assessed based on accuracy, precision, recall, and F1-score. Our results show that the Gradient Boosting model achieved the highest performance across all metrics, indicating its superior ability to predict diabetes accurately. The study demonstrates the potential of machine learning techniques in enhancing diabetes prediction and provides a comparative analysis of different models to guide future research and practical applications in medical diagnostics.



## INTRODUCTION

Diabetes mellitus is a chronic disease that occurs when the body is unable to regulate blood sugar levels effectively. This condition can lead to serious health complications such as heart disease, kidney failure, and nerve damage if not managed properly. With the rising prevalence of diabetes worldwide, early detection and management are crucial. Machine learning models offer a promising approach to predicting diabetes by analysing various medical and lifestyle factors. This study aims to evaluate the effectiveness of different machine learning models in predicting diabetes using the Pima Indians Diabetes dataset, a well-known dataset in the field of medical research.

The Pima Indians Diabetes dataset consists of several medical predictor variables and one target variable, which indicates the presence or absence of diabetes. By leveraging this dataset, we aim to train and test various machine learning algorithms to identify the most effective model for diabetes prediction. The study involves preprocessing the data, training multiple models, and evaluating their performance using various metrics such as accuracy, precision, recall, and F1-score. This research provides insights into the applicability of machine learning techniques in medical diagnosis and highlights the best-performing models for predicting diabetes.



DATASET

The dataset used in this study is the Pima Indians Diabetes dataset, sourced from the UCI Machine Learning Repository. It consists of 768 samples with 8 features: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age. The target variable is 'Outcome', indicating whether a patient has diabetes (1) or not (0).

**4. Proposed Methods and Methodology**

**4.1 Data Collection**

The Pima Indians Diabetes dataset was collected from the UCI Machine Learning Repository. This dataset contains medical data for female patients of Pima Indian heritage.

**4.2 Data Preprocessing**

In this study, the Pima Indians Diabetes dataset, which includes medical measurements for 768 women, was pre-processed to ensure data quality and enhance model performance. First, columns such as Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI had instances of 0, indicating missing values. These were replaced with Nan and subsequently filled with the mean of their respective columns to provide reasonable estimates. Feature selection retained key variables like Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age. To ensure uniform contribution of all features, the data was standardized using the StandardAero from scikit-learn, which removes the mean and scales to unit variance. Finally, the dataset was split into training and testing sets to facilitate model building and evaluation.

4o

**4.2.1 Missing Value Identification**

Some columns contained zero values, which were treated as missing data:

* Glucose
* Blood Pressure
* Skin Thickness
* Insulin
* BMI

**4.2.2 Handling Missing Values**

Missing values were replaced with the mean of the respective columns. A value of 0 was used in several columns of the dataset to represent missing measures, which is unrealistic for medical data such as BMI, insulin, glucose, blood pressure, and skin thickness. In order to remedy this, NaN was used in place of these 0 values to explicitly indicate missing data. The means of the corresponding columns were then used to fill in these NaN values. This imputation technique improves the trustworthiness of the data for model training by keeping all records for analysis and guaranteeing that the filled values are typical of the dataset.

**4.2.3 Feature Selection**

All features in the dataset were retained for model training. The process of feature selection entailed locating and exploiting the dataset's most pertinent properties in order to construct predictive models. Since numerical columns are essential to the prediction job, they were first all identified for analysis. Visualizations such as histograms, box plots, and correlation matrices were used in exploratory data analysis (EDA) to comprehend the distribution and associations between the attributes and the target variable. This aided in the identification of trends and connections, which influenced the choice of characteristics with high predictive value. The Random Forest model was also utilized to rank the relevance of the features, making sure that the most significant characteristics were given priority and enhancing the interpretability and performance of the model.

**4.2.4 Scaling and Normalization**

The data was standardized using StandardAero to ensure that all features have a mean of 0 and a standard deviation of 1. Scaling and normalization were done on the dataset to make sure that each feature contributes equally to the model and to enhance the model's convergence during training. The feature values were transformed to have a mean of zero and a standard deviation of one by using the Scikit-Learn StandardAero. This is an essential step because it reduces the impact of different feature sizes and units, which might skew the model training procedure. The scaler was fitted on the training data and then applied to both the training and test datasets to ensure consistency after the data was divided into training and test sets. The efficiency and stability of many machine learning algorithms are improved by this standardized method, especially those that are sensitive to feature scaling, including as logistic regression, K-nearest neighbors, and support vector machines.

**4.3 Splitting of Data**

The dataset was split into training and test sets using an 80-20 split ratio. Stratification was used to maintain the proportion of diabetic and non-diabetic cases in both sets. The dataset was split into training and testing subsets to assess how well the machine learning models performed. This was accomplished by using Scikit-Learn's train\_ test\_ split function. Eighty percent of the dataset was utilized for training the models, while the remaining twenty percent was set aside for testing, according to an 80-20 split of the dataset. The repeatability of the findings was guaranteed by setting the random state parameter to 42. The stratify option was employed to preserve the target variable's ('Outcome') distribution in the training and testing sets. For the model to perform effectively on data that has not yet been seen, this stratification makes sure that the percentage of diabetes and non-diabetic patients in each group is identical. By dividing the data into testing and training sets

we can train the models on one portion of the data and evaluate their performance on a different portion, thereby obtaining a realistic estimate of their predictive power.

**5. Models Used and Analysis**

The following models were trained and evaluated:

* Logistic Regression
* K-Nearest Neighbours (KNN)
* Random Forest
* Support Vector Machine (SVM)
* Naive Bayes
* AdaBoost
* Gradient Boosting

Each model was evaluated using accuracy, precision, recall, and F1-score metrics. Cross-validation with 5 and 7 folds was used to validate the models. In this study, a variety of machine learning models were employed to predict the likelihood of diabetes. The models used include Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, Support Vector Machine (SVM), Naive Bayes, AdaBoost, and Gradient Boosting. Each model brings its unique strengths to the table: Logistic Regression offers simplicity and interpretability, KNN is intuitive and easy to implement, Random Forest provides robustness and handles non-linearity well, SVM is effective in high-dimensional spaces, Naive Bayes is computationally efficient and performs well with categorical data, while AdaBoost and Gradient Boosting are powerful ensemble methods that improve prediction accuracy by combining multiple weak learners.

**6. Measurement Metrics**

The models were evaluated using the following metrics:

* **Accuracy**: Proportion of true results (both true positives and true negatives) among the total number of cases.
* **Precision**: Proportion of true positive results among all positive results predicted by the model.
* **Recall**: Proportion of true positive results among all actual positives.Recall is the ratio of true positive predictions to real positives (true positives and false negatives) that indicates how well the model can recognize all relevant events. In order to make sure that diabetes patients are not overlooked, high recall guarantees that the majority of positive cases are found.
* **F1-Score**: Harmonic mean of precision and recall.
* **Confusion Matrix**: This matrix provides a detailed breakdown of the actual versus predicted classifications, highlighting the true positives, false positives, true negatives, and false negatives. It is a valuable tool for understanding the specific types of classification errors made by the model.
* By leveraging these metrics, we can gain a nuanced understanding of each model's performance, identify strengths and weaknesses, and make informed decisions about model selection and further improvements. These metrics collectively ensure that the models are evaluated not just on their overall correctness but also on their ability to handle imbalances and the trade-offs between precision and recall.

**7. Results and Discussion**

**7.1 Model Performance**

The performance of various machine learning models on the diabetes dataset was evaluated using several metrics, including accuracy, precision, recall, F1 score, and ROC-AUC. The models assessed included Logistic Regression, K-Nearest Neighbours (KNN), Random Forest, Support Vector Machine (SVM), Naive Bayes, AdaBoost, and Gradient Boosting. Each model demonstrated distinct strengths and weaknesses, reflecting their inherent characteristics and suitability for different aspects of the classification task.

Logistic Regression showed a balanced performance with an accuracy of 69.48%. While its precision and recall were moderately high, its overall F1 score suggested it was reasonably effective but not the best model for this dataset**.** With an accuracy of 73.38%, KNN outperformed Logistic Regression. Additionally, it demonstrated a strong equilibrium between recall and accuracy, suggesting that it successfully identified positive cases while keeping a comparatively low false positive rate.  
  
With a flawless accuracy of 100%, Random Forest surpassed all other models, indicating that it could accurately categorize each and every case in the test set. The fact that its memory, accuracy, and F1 score were all flawless supported this. If the model performs very well on the training set but may not generalize as well to new data, this might be a sign of overfitting.

The following table summarizes the performance metrics for each model:

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| Logistic Regression | 0.69 | 0.69 | 0.69 | 0.69 |
| KNN | 0.73 | 0.73 | 0.73 | 0.73 |
| Random Forest | 1.00 | 1.00 | 1.00 | 1.00 |
| SVM | 0.73 | 0.72 | 0.73 | 0.72 |
| Naive Bayes | 0.73 | 0.72 | 0.73 | 0.72 |
| AdaBoost | 0.73 | 0.73 | 0.73 | 0.73 |
| Gradient Boosting | 0.77 | 0.76 | 0.77 | 0.76 |

**7.2 Confusion Matrices**

Confusion matrices were plotted for each model to visualize the performance in terms of true positives, true negatives, false positives, and false negatives. Confusion matrices provide a detailed breakdown of the model's performance by showing the number of true positives, true negatives, false positives, and false negatives. This information is crucial for understanding the types of errors a model makes and for comparing the effectiveness of different models. The confusion matrices provide insights into the models' strengths and weaknesses. **Random Forest** shows perfect classification with no errors, but this suggests possible overfitting. **Gradient Boosting** also performs very well with a high number of true positives and true negatives while maintaining a lower number of false positives and false negatives. The other models, such as **Logistic Regression**, **KNN**, **SVM**, **Naive Bayes**, and **AdaBoost**, show varying degrees of errors, with **KNN** and **AdaBoost** slightly outperforming Logistic Regression and Naive Bayes.

**7.3 ROC Curves**

ROC curves were plotted for each model to compare their performance in distinguishing between diabetic and non-diabetic cases. Of course! Graphical representations of ROC curves, or Receiver Operating Characteristic curves, show how well a binary classification model performs across a range of thresholds. Plotting the True Positive Rate (Sensitivity) vs the False Positive Rate (1 - Specificity) at various threshold levels is what each curve represents. A ROC curve reaching the upper left corner, which denotes high sensitivity (accurately detecting positives) and low false positives (inaccurately identifying negatives), would be indicative of an ideal classifier.  
  
For every model used in this investigation, including Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, Support Vector Machine (SVM), Naive Bayes, AdaBoost, and Gradient Boosting, ROC curves were drawn. The ability to compare each model's ability to distinguish between the positive and negative classes visually is made possible by these curves.

**8. Conclusion**

The Gradient Boosting model exhibited the highest performance in terms of accuracy, precision, recall, and F1-score. This suggests that ensemble methods, particularly boosting techniques, are highly effective for predicting diabetes. Random Forest also showed perfect performance, but this could indicate overfitting, necessitating further investigation. Based on a thorough examination of the dataset, this study assessed the predictive power of many machine learning models for diabetes. Gradient Boosting was the most reliable model among those evaluated; it demonstrated excellent accuracy, precision, recall, and F1-score. This demonstrates how well ensemble learning methods—specifically, boosting—handle the complexity of diabetes prediction. The flawless results with Random Forest, however, raise questions about possible overfitting and call for care when using it without additional testing on unobserved data. Additionally, the feature significance analysis revealed important variables impacting diabetes outcomes, suggesting possible directions for future study and therapeutic inquiry. Overall, these results highlight the promise of machine learning and highlight the significance of model selection and validation in healthcare analytics.

**9. References**

* UCI Machine Learning Repository: Pima Indians Diabetes Dataset
* Scikit-learn Documentation: <https://scikit-learn.org/>
* Seaborn Documentation: https://seaborn.pydata.org.