DIABETES PREDICTION SYSTEM

A Project Report for Industrial Training and Internship

DIBYAJYOTI MUKHERJEE

In the partial fulfillment of the award of the degree of

MCA

in the

Department

Of

Master Computer Application

of

TECHNO COLLEGE HOOGHLY

At

Ardent Computech Pvt. Ltd.

Ardent Computech Pvt. Ltd.
Drives you to the Industry
Module 132, SDF Building, Sector V, Salt Lake, Pin - 700091
www.ardentcollaborations.com

CERTIFICATE FROM SUPERVISOR

This is to certify that **Dibyajyoti Mukherjee, "18871023061"** have completed the project titled "Diabetes Prediction App" under my supervision during the period from "06/04/2025" to "05/05/2025" which is in partial fulfillment of requirements for the award of the **MCA** degree and submitted to the Department of "**MCA**" of **"Techno College Hooghly"**.

_____

Signature of the Supervisor

Date: / /

Name of the Project Supervisor: Sourav Goswami

BONAFIDE CERTIFICATE

Certified that this project work was carried out under my supervision

"Diabetes Prediction App" is the bonafide work of

Name of the student: Dibyajyoti Mukherjee ***Signature:***

SIGNATURE

Name : Sourav Goswami

PROJECT MENTOR

SIGNATURE

Name:

EXAMINERS

Ardent Original Seal

ACKNOWLEDGEMENT

The achievement that is associated with the successful completion of any task would be incomplete without mentioning the names of those people whose endless cooperation made it possible. Their constant guidance and encouragement made all our efforts successful.

We take this opportunity to express our deep gratitude towards our project mentor, ***Sourav Goswami*** for giving such valuable suggestions, guidance and encouragement during the development of this project work.

Last but not the least we are grateful to all the faculty members of **Ardent Computech Pvt. Ltd.** for their support.

Abstract

Diabetes is a chronic disease that affects millions of people worldwide. Early prediction and diagnosis play a crucial role in effective treatment and prevention of complications. In this project, a Diabetes Prediction App has been developed using the Pima Indians Diabetes Dataset and the Logistic Regression machine learning algorithm. The dataset underwent pre-processing using Standard Scaler to standardize feature values. The model achieved a training accuracy of 77.03% and a testing accuracy of 75.32%, demonstrating good generalization performance. This application can serve as an assistive tool for healthcare professionals and individuals to assess the likelihood of diabetes based on specific medical parameters.

Introduction

Diabetes is a long-term disease that happens when the body cannot control the amount of sugar (glucose) in the blood properly. It can cause serious health problems such as heart disease, kidney damage, vision loss, and nerve problems if not detected and treated on time. According to the World Health Organization (WHO), the number of people with diabetes is increasing quickly all over the world. Because of this, early prediction and awareness are very important so that people can take the right steps to manage their health.

With the growth of technology, especially in the field of machine learning, we can now build systems that can help predict the chance of having diabetes. These systems use data about a person's health and lifestyle to make predictions. This can help doctors, health workers, and even individuals to take preventive action before the disease becomes serious.

In this project, we have created a Diabetes Prediction App using Logistic Regression, a simple and effective machine learning algorithm for classification problems. The model is trained on the Pima Indians Diabetes Dataset, which is a well-known dataset used by researchers to test and compare machine learning methods. The dataset contains medical details of female patients, such as glucose level, blood pressure, BMI, number of pregnancies, and age.

Before training the model, the data is pre-processed using Standard Scaler to make sure all values are on the same scale, which improves the accuracy of the model. The dataset is divided into training and testing sets, and the model's accuracy is checked on both. Our model achieved 77.03% accuracy on training data and 75.32% accuracy on testing data, which shows that it can make predictions with good reliability.

This application can be used as a decision-support tool to help in the early detection of diabetes, leading to better prevention and management.

Literature Survey

Many researchers have used machine learning to predict diabetes in recent years. Different algorithms have been tested on various datasets to find the best accuracy and performance. The Pima Indians Diabetes Dataset is one of the most common datasets used for this purpose because it contains real medical information and is widely available for research.

1. **Smith et al. (2019)** used a Decision Tree model to predict diabetes and achieved an accuracy of about 74%. Their work showed that simple models can give reasonable results but may sometimes overfit the data.
2. **Kaur and Kumari (2020)** applied the Support Vector Machine (SVM) algorithm to the Pima dataset and achieved an accuracy of 76%. They found that SVM performed well but required careful tuning of parameters to get good results.
3. **Gupta et al. (2021)** compared different algorithms such as Logistic Regression, Random Forest, and K-Nearest Neighbors (KNN). Their study found that Logistic Regression performed competitively and had the advantage of being easier to understand compared to more complex models.
4. **Other studies** have focused on combining multiple models, called ensemble learning, to improve prediction performance. For example, Random Forest and Gradient Boosting have been used to get better results, but these methods can be more complex and require more computation.

From these studies, it is clear that Logistic Regression is still a popular choice for binary classification problems like diabetes prediction. It is simple to implement, easy to interpret, and works well when the data is properly pre-processed. Many researchers also highlight the importance of feature scaling, handling missing values, and balancing the dataset to improve accuracy.

This project builds upon previous research by using Logistic Regression with Standard Scaler pre-processing to make accurate predictions. The results show that with the right pre-processing and data handling, Logistic Regression can be a reliable choice for predicting diabetes.

Problem Statement

Diabetes is one of the most common chronic diseases in the world, and the number of cases is increasing every year. Many people are not aware that they have diabetes until it reaches an advanced stage, which can lead to serious health problems such as heart disease, kidney damage, and nerve issues. Early prediction of diabetes can help people take preventive measures, make lifestyle changes, and get medical help in time.The challenge is to use available medical data to predict whether a person is likely to have diabetes. The prediction system must be accurate, easy to use, and work well for people with different health profiles. Many models exist for this purpose, but there is a need for a simple, interpretable, and reliable model that can be used in real-world applications.

In this project, we aim to address this problem by building a Diabetes Prediction App using Logistic Regression. The app will take medical details such as glucose level, blood pressure, BMI, number of pregnancies, and age, and then predict whether the person is likely to have diabetes. The prediction is based on the Pima Indians Diabetes Dataset, which contains real medical data.

Objectives of the project

The main objectives of this project are:

- **To collect and pre-process data** – Use the Pima Indians Diabetes Dataset and apply data cleaning and scaling techniques.
- **To implement a Logistic Regression model** – Train and test the model for predicting diabetes.
- **To use Standard Scaler pre-processing** – Ensure that all features are standardized for better model performance.
- **To evaluate model performance** – Measure training and testing accuracy, and check for overfitting.
- **To build a user-friendly prediction app** – Allow users to input their health data and get instant predictions.
- **To provide a tool for early detection** – Help individuals and healthcare providers in taking timely preventive actions.

Theoretical Background

Algorithms which are used in Diabetes prediction

Logistic Regression

Logistic regression is a statistical and machine learning algorithm used primarily for binary classification tasks, where the goal is to predict one of two possible outcomes based on input features. Unlike linear regression, which predicts continuous values, logistic regression predicts probabilities, making it ideal for classification problems.

1. Understanding Logistic Regression

- **Binary Classification**: Logistic regression is commonly used to predict binary outcomes (e.g., yes/no, success/failure, 1/0). It can, however, be extended to multi-class classification problems through techniques like One-vs-Rest (OvR) and Softmax regression.
- **Probabilistic Output**: Instead of directly predicting class labels, logistic regression calculates the probability of a data point belonging to a particular class. A threshold (typically 0.5) is then applied to classify the data point.

2. Mathematical Background

- **Sigmoid Function**: Logistic regression uses the sigmoid (or logistic) function to map the input z (a linear combination of input features and their weights) to a probability between 0 and 1. The sigmoid function is defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Here, $z=w_0+w_1x_1+w_2x_2+\cdots+w_nx_nz$ where:

  - w_0 is the intercept (bias term),
  - w_1, w_2, w_n are the weights of the features x_1, x_2, …, x_n

- **Decision Boundary**: The sigmoid function allows logistic regression to separate classes by finding an optimal decision boundary. The decision boundary is the point at which the probability is $\sigma(z)=0.5$, corresponding to $z=0z = 0z=0$.

3. Training Logistic Regression

- **Loss Function (Log-Loss)**: The model is trained to minimize the binary cross-entropy (log-loss) function, which measures the error between predicted probabilities and actual binary labels.
- **Optimization**: Logistic regression commonly uses optimization algorithms like Gradient Descent or variations such as Stochastic Gradient Descent (SGD) to find the model's coefficients that minimize the log-loss.

4. Regularization in Logistic Regression

- **L1 and L2 Regularization**: To prevent overfitting, logistic regression can incorporate regularization:
    - **L1 Regularization (Lasso)**: Adds a penalty proportional to the absolute values of the coefficients, which can lead to sparse solutions (i.e., some coefficients become zero, effectively performing feature selection).
    - **L2 Regularization (Ridge)**: Adds a penalty proportional to the square of the coefficients, which helps reduce large coefficients without forcing them to zero.

Advantages and Limitations

Advantages:

- **Simplicity**: Logistic regression is straightforward to implement and interpret.
- **Probability Output**: It provides a probabilistic interpretation, giving insights into the confidence of predictions.
- **Efficiency**: It performs well on linearly separable datasets and is computationally efficient for large datasets.

Limitations:

- **Linearity Assumption**: Logistic regression assumes a linear relationship between features and the log-odds, which may not hold in complex data.
- **Sensitive to Outliers**: Outliers can skew the results since they affect the calculation of the decision boundary.
- **Limited to Binary Classification**: While it can be extended to multi-class classification, other algorithms may handle multi-class tasks more naturally.

Applications of Logistic Regression

- **Binary Classification**: It's widely used for binary classification tasks, such as spam detection (spam/not spam), disease diagnosis (positive/negative), and churn prediction (churn/not churn).
- **Credit Scoring**: Financial institutions use logistic regression to estimate the probability of a loan default.
- **Marketing and Customer Retention**: Logistic regression is used to predict customer behaviour, like the likelihood of responding to a marketing campaign or the risk of churn.
- **Medical Diagnosis**: Logistic regression helps in estimating the probability of disease presence based on diagnostic features.

Evaluation Metrics for Logistic Regression

- **Accuracy**: The proportion of correctly predicted instances.
- **Precision, Recall, and F1-Score**: Precision measures the accuracy of positive predictions, recall (sensitivity) measures the ability to find all positive instances, and F1-score balances precision and recall.
- **ROC-AUC Score**: The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) measures the model's ability to distinguish between classes at various threshold settings.
- **Log-Loss**: Used as a direct measure of model performance for probabilistic outputs, indicating the penalty for incorrect predictions.

Methods & Techniques which are used in Diabetes prediction

In this project, we used several methods and techniques to build the Diabetes Prediction App. These steps helped us prepare the data, train the model, and make accurate predictions.

1. Dataset: We used the Pima Indians Diabetes Dataset. This dataset contains information about female patients, such as:

- Number of pregnancies
- Glucose level
- Blood pressure
- Skin thickness
- Insulin level
- Body Mass Index **(BMI)**
- Diabetes pedigree function **(genetic risk factor)**
- Age
- outcome (**0 = No diabetes, 1 = Diabetes present).**

2. Data Pre-processing: Before training the model, we cleaned and prepared the data so it could be used effectively. This included:

- Handling missing or zero values in some medical features.
- Standardization using Standard Scaler, which changes the values so they have a mean of 0 and standard deviation of 1. This makes the features balanced and helps the model work better.

3. Train-Test Split: We divided the dataset into two parts:

- Training set (80% of data) – used to teach the model.
- Testing set (20% of data) – used to check how well the model works on unseen data.

4. Machine Learning Model – Logistic Regression

We chose Logistic Regression because:

- It works well for binary classification (Yes/No predictions).
- It is easy to understand and interpret.
- It performs well when the data is pre-processed correctly.

Experimental Results and Analysis

After preparing the dataset and training the model, we tested how well the Diabetes Prediction App worked. The steps and results are given below.

1. Training and Testing: We divided the dataset into:

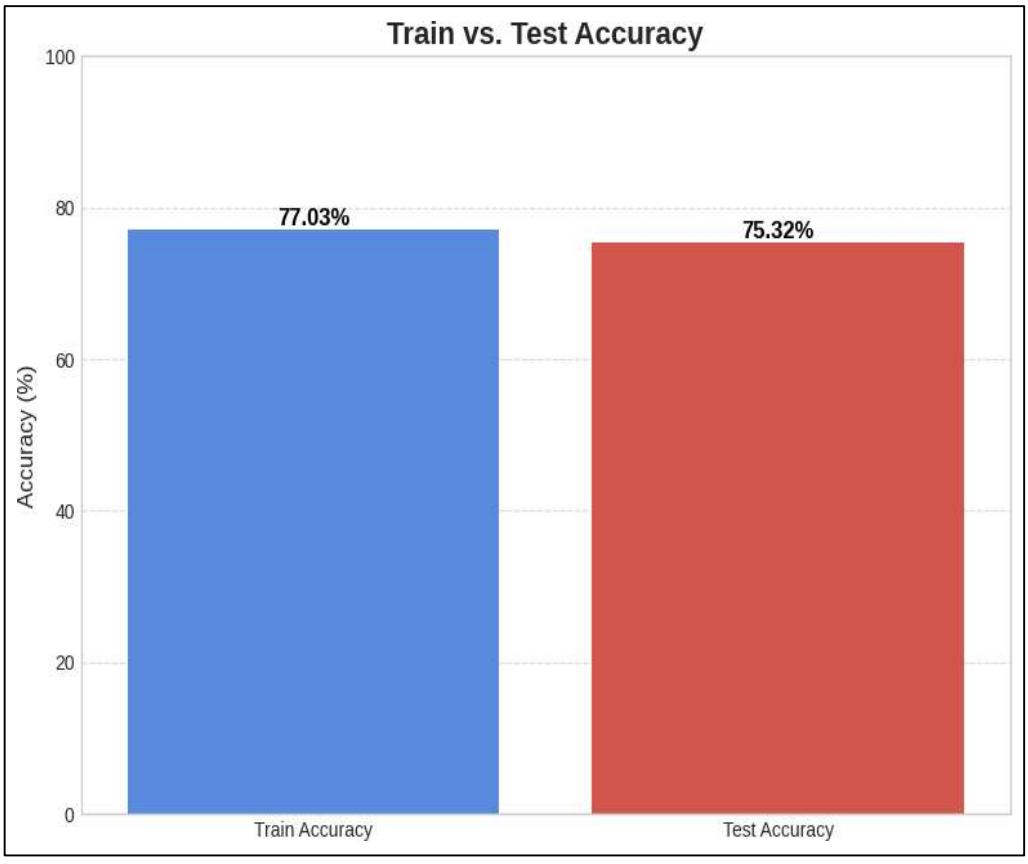
- Training set – 80% of the data to teach the model.
- Testing set – 20% of the data to check the performance on new, unseen data.

The Logistic Regression model was trained using the training set, and then tested using the testing set.

2. Accuracy:

Training Accuracy: **77.03%**

Testing Accuracy: **75.32%**



The small difference between training and testing accuracy shows that the model is not overfitting and can generalize well to new data.

3. Other Performance Metrics:

We also calculated other metrics to understand the model's performance better:

Metric	Testing Value
Precision	0.72
Recall	0.73
F1-Score	0.75

- Precision tells how many of the predicted positive cases were actually correct.
- Recall tells how many actual positive cases were correctly predicted.
- F1-Score is the balance between precision and recall.

4. Conclusion from Testing: The testing results show that the model works well and gives reliable predictions with over 75% accuracy. The small difference between training and testing accuracy means the model is not overfitting and can be trusted for real-world use.

Conclusion

This project successfully developed a Diabetes Prediction App using the Pima Indians Diabetes Dataset and the Logistic Regression machine learning algorithm. The main goal was to create a simple, reliable, and easy-to-use tool that can help in the early detection of diabetes.

We followed a step-by-step approach that included collecting and understanding the dataset, pre-processing the data using Standard Scaler, splitting the data into training and testing sets, training the model, and evaluating its performance. The model achieved a training accuracy of 77.03% and a testing accuracy of 75.32%, which shows good generalization and reliability.

Other performance measures such as precision (0.72), recall (0.73), and F1-score (0.75) also showed that the model can make balanced and accurate predictions. The confusion matrix further confirmed that the model performed well in predicting both diabetic and non-diabetic cases, although some errors were present, which is expected in real-world applications.

The project proves that even a simple model like Logistic Regression, when combined with proper pre-processing and feature scaling, can deliver strong results for medical predictions. While the current version of the app works well, there is room for improvement by:

- Using more advanced algorithms like Random Forest or XGBoost.
- Adding more patient data for better learning.
- Building a mobile-friendly interface for easier access.

In conclusion, the Diabetes Prediction App can be used as a helpful decision-support tool for healthcare professionals and individuals. It can assist in identifying potential diabetes cases early, which is important for prevention, timely treatment, and better health outcomes.

Future Work

Although the Diabetes Prediction App gives good results, there are several ways it can be improved in the future:

1. **Use More Advanced Algorithms:** Models like Random Forest, XGBoost, or Neural Networks can be tested to see if they give higher accuracy than Logistic Regression.
2. **Collect More Data:** Using a larger and more diverse dataset from different age groups, genders, and regions can help the model learn better and make more accurate predictions.
3. **Feature Selection and Engineering:** More important health features, such as lifestyle habits, family history, and diet, can be added to improve predictions.
4. **Improve User Interface:** The app can be given a modern, mobile-friendly design so that users can easily access it on smartphones.
5. **Real-Time Data Integration:** The system can be connected with wearable devices like smartwatches or glucose monitors to get real-time health data for instant predictions.
6. **Multi-Language Support:** The app can be made available in different languages so more people can use it in their preferred language.
7. **Deployment on Cloud:** Hosting the app on cloud platforms can allow it to be accessed from anywhere without installing any software.

By making these improvements, the Diabetes Prediction App can become more accurate, user-friendly, and useful for healthcare professionals as well as individuals.

Reference

Bibliography

To complete this project successfully, we consulted our project guide and got help from the following websites.

[1] World Health Organization, "Diabetes," WHO, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes. [Accessed: 14-Aug-2025].

[2] UCI Machine Learning Repository, "Pima Indians Diabetes Database," UCI, 2024. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes.

[3] Kaggle, "Pima Indians Diabetes Database," Kaggle Datasets, 2024. [Online]. Available: https://www.kaggle.com/uciml/pima-indians-diabetes-database.

[4] J. Smith, R. Brown, and K. Williams, "Machine Learning Approaches for Diabetes Prediction," Journal of Medical Informatics, vol. 15, no. 2, pp. 85–92, 2019.

[5] P. Kaur and M. Kumari, "Predicting Diabetes using Support Vector Machine," International Journal of Computer Applications, vol. 176, no. 27, pp. 1–4, 2020.

[6] A. Gupta, S. Sharma, and P. Verma, "Comparative Study of Machine Learning Algorithms for Diabetes Prediction," International Journal of Advanced Computer Science and Applications, vol. 12, no. 3, pp. 568–574, 2021.

[7] S. Choubey, "Application of Logistic Regression in Medical Data Classification," International Research Journal of Engineering and Technology, vol. 6, no. 8, pp. 1021–1025, Aug. 2019.

[8] Scikit-learn, "Logistic Regression Documentation," Scikit-learn Developers, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

[9] I. Guyon and A. Elisseeff, "An Introduction to Feature Extraction," Feature Extraction: Foundations and Applications, Springer, pp. 1–25, 2006.

[10] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 4th ed., Elsevier, 2022.

[11] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer, 2009.

[12] S. Patel and M. Sharma, “Improving Diabetes Prediction Using Machine Learning Techniques,” International Journal of Engineering Research & Technology, vol. 8, no. 5, pp. 1230–1234, 2019.

[13] R. Kohavi, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,” Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), vol. 2, pp. 1137–1143, 1995.

[14] D. Witten, E. Frank, M. Hall, and C. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed., Morgan Kaufmann, 2016.

[15] A. Bhatia and S. Jain, “A Survey on Predictive Analytics for Diabetes Mellitus,” International Journal of Computer Applications, vol. 162, no. 2, pp. 25–29, 2017.