# DIABETES PREDICTION USING ARTIFICIAL INTELLIGENCE WITH PYTHON

Project Submitted in Partial Fulfilment of the Requirements for the Degree of Bachelor of Technology in the field of Computer Science and Engineering

BY
SHASHANK SHEKHAR SAHI (23221103143)
VISHAL KUMAR (123221103186)
SHUJAL SHAW (123221103172)
SHUBHAM SAURAV (123221103147)

Under the supervision of
Dr. Joy Shree Bhattacharya

Department of Computer Science and Engineering
JIS College of Engineering

Block-A, Phase-III, Kalyani, Nadia, Pin-741235
West Bengal, India June,
2025

JIS College of Engineering

Block 'A', Phase-III, Kalyani, Nadia, 741235
Phone: +91 33 2582 2137, Telefax: +91 33 2582 2138
Website: www.jiscollege.ac.in, Email: info@jiscollege.ac.in

# CERTIFICATE

This is to certify that **Shashank Shekhar Sahi (123221103143)**, **Vishal Kumar (123221103186)**, **Sujal Shaw (123221103172)**, **Shubham Saurav (123221103147).** have completed their project entitled **Diabetes Prediction Using Artificial Intelligence With Python,** under the guidance of **Dr. Joy Shree Bhattacharya** in partial fulfilment of the requirements for the award of the **Bachelor of Technology in Computer Science and Engineering** from JIS college of Engineering (An Autonomous Institute) is an authentic record of their own work carried out during the academic year 2024-25 and to the best of our knowledge, this work has not been submitted elsewhere as part of the process of obtaining a degree, diploma, fellowship or any other similar title.

-------------------------------- ------------------------------

**Signature of the Supervisor** **Signature of the HOD**

**Place:**

**Date:**

# ACKNOWLEDGEMENT

The analysis of the project work wishes to express our gratitude to **Dr. Joy Shree Bhattacharya** for allowing the degree attitude and providing effective guidance in development of this project work. Her conscription of the topic and all the helpful hints, She provided, contributed greatly to successful development of this work, without being pedagogic and overbearing influence.

We also express our sincere gratitude to HOD, **Dr. Pranati Rakshit**, Head of the Department of Computer Science and Engineering of JIS College of Engineering and all the respected faculty members of Department of CSE for giving the scope of successfully carrying out the project work.

Finally, we take this opportunity to thank to Prof. **(Dr.) Partha Sarkar**, Principal of JIS College of Engineering for giving us the scope of carrying out the project work.

Date:

……………………………………..………………………………

Shashank Shekhar Sahi
B.TECH in Computer Science and Engineering
3rd YEAR/6th SEMESTER
Univ Roll-- 123221103143

……………………………………..………………………………

Vishal Kumar
B.TECH in Computer Science and Engineering
3rd YEAR/6th SEMESTER
Univ Roll-- 123221103186

……………………………………..………………………………

Sujal Shaw
B.TECH in Computer Science and Engineering
3rd YEAR/6th SEMESTER
Univ Roll—123221103172

……………………………………..………………………………

Shubham Saurav
B.TECH in Computer Science and Engineering
3rd YEAR/6th SEMESTER
Univ Roll-- 123221103147

# CONTENTS

# ABSTRACT :

Diabetes mellitus is the term for a chronic disease that can be lethal if left untreated. This disease affects millions of people throughout the world. With an ever-increasing number of people with diabetes, screening has thus become more important to prevent long-term damage. The traditional techniques of diagnosis have their own cons, being invasive or taking time, often deliver an untimely response. Artificial intelligence (**AI**) thus emerges as a nice tool that can be used in prediction, across large datasets describing historical medical data, of the likelihood of diabetes in an individual. This research offers a big robust system constructed on the basis of the Support Vector Machine (SVM) classification algorithm to predict diabetes based on parameters in the Pima Indian Diabetes Dataset.

The primary aim of the study was to build a prediction model that could aid health professionals in the early diagnosis of diabetes through the analysis of some common health parameters. The dataset contained values for several major medical features such as glucose level, insulin level, BMI, age, number of pregnancies, blood pressure, skin thickness, and diabetes pedigree function. These parameters are well known to be greatly associated with the presence or absence of diabetes. Data preprocessing is of key importance in the cleaning and preparation of the data for analysis. Missing values were taken care of, outliers, and feature scaling transformations were applied to guarantee fair input to the AI algorithm.

After the data preparation stage, the dataset was split between training and testing in a 70:30 ratio. An SVM classifier with RBF kernel was trained on training data. The hyperparameters were fine-tuned using GridSearchCV for the best model performance. The trained model underwent testing using the usual classification metrics: accuracy, precision, recall, F1-score, and confusion matrix. The SVM model did a very good job of achieving more than 81% accuracy on the test set, which indicates that it can be very reliable for real-world applications.

The project also considered comparing the SVM model with other commonly used Artificial intelligence methods such as Logistic Regression, Random Forest, and K-Nearest Neighbor. The SVM model scored better for precision and sensitivity. While the confusion matrix indicated a really high true positive rate, that in itself is very fundamental in a medical field where false negatives can be life-threatening! The ROC curve, as well as the AUC, gave further validation to the power of this model.

From the technical development point of view, Python along with several popular libraries for Artificial intelligence such as Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn, is used. Python's simplicity and extensive libraries are an ideal environment for carrying out data preprocessing, model training, evaluation, and visualization. The end product is a standalone, scalable, and lightweight predictive Python-based system that can be stitched into clinical decision support tools.

It is evident through this type of system if AI can go a long way to help diabetes' early detection, thus easing the burden on healthcare systems and ensuring that patients regain their health in time. It provides a fast, non-invasive procedure for testing, making it more suitable for mass screening in a resource-poor set-up.

# 1. INTRODUCTION :

## 1.1 Background and Motivation

Diabetes mellitus is the most common and fastest spreading chronic illness globally, with more than 537 million adults worldwide suffering from it as of 2021, based on the International Diabetes Federation (IDF). It is forecasted to increase exponentially in the decades to come, posing a huge challenge to global public health infrastructure. Diabetes not only results in disabling complications like cardiovascular disease, kidney failure, blindness, and neuropathy, but it is also a significant economic burden on individuals and the healthcare system.

The key issue with diabetes control is early diagnosis and treatment, which can greatly delay the aggravation of complications and enhance patient outcomes. The diagnosis of diabetes is usually made after the development of symptoms or at the time of routine health screenings by then, and the disease has already advanced. Hence, pre-symptomatic diagnostic systems that can evaluate a person's risk for developing diabetes prior to the development of severe symptoms have become extremely essential.

Recent developments in Artificial Intelligence (AI) and Artificial intelligence (AI) provide newage solutions to forecast chronic conditions like diabetes. These are capable of processing large volumes of medical data to reveal subtle patterns and correlations that might otherwise go unnoticed using conventional clinical practices. Based on past health records and computational frameworks, Artificial intelligence methods can well identify if a person is likely to suffer from diabetes and, thereby, help medical practitioners take proactive action.

## 1.2 Problem Statement

Conventional diagnostic tests for diabetes, including fasting plasma glucose, oral glucose tolerance tests, and A1C blood tests, are usually cumbersome, need laboratory facilities, and may be impractical for mass or rural screenings. Further, these tests do not yield an automatic risk stratification of the population. Consequently, a large number of patients are undiagnosed or diagnosed late in the course of the disease.

The central problem this project addresses is:

"How can we construct an intelligent, precise, and efficient Artificial intelligence model capable of predicting diabetes in people based on easily available medical features?"

This research aims to design a predictive system from Artificial intelligence that is capable of analyzing clinical data and predicting diabetes through classification. The aim is to implement a precise, light, and scalable model using Python with the main focus on the Support Vector Machine (SVM) algorithm because it is very strong when dealing with nonlinear data.

### 1.3 Objectives of the Study

The major goals of the project are as given below:

- To investigate the usability of Artificial intelligence models for predicting diabetes.
- To train and compare different models like Support Vector Machine (SVM), Logistic Regression, Random Forest, and K-Nearest Neighbors to predict diabetes.
- To compare the above models on performance criteria like accuracy, precision, recall, F1-score, and ROC-AUC.
- In order to use the Pima Indian Diabetes Dataset having actual-world data of female patients with other features that help in predicting diabetes.
- To create a scalable and secure Python-based system that could be implemented in healthcare systems for early detection and diagnosis.

### 1.4 Significance of the Study

This project seeks to close the gap between healthcare and data science by presenting a solution that improves medical diagnostics with Artificial intelligence methods. The usefulness of this research is in its ability to:

- Facilitate early diabetes detection, hence lowering complications and treatment costs.
- Provide a data-driven, non-invasive method that complements conventional medical tests.
- Illustrate the use of Artificial intelligence within real-world clinical decision support systems.
- Establish the foundation for future research on using AI models to predict other long-term illnesses.

This research also highlights the value of open datasets in facilitating reproducible and transparent scientific work. Using the well-studied Pima Indian Diabetes Dataset, the findings of this project are easily replicable and expandable by other researchers, stimulating collaboration and innovation within the research community.

### 1.4 Scope of the Project

The focus of this research is limited to the following:

- Use of supervised Artificial intelligence methods, particularly classification algorithms.

- Preprocessing and analysis of the Pima Indian Diabetes Dataset.
- Applying and comparing various AI models in Python.
- Model performance comparison and selection of the most effective algorithm.
- Creating a user-understandable system for healthcare purposes.

It does not include medical image processing or real-time data capture via wearable sensors, but these are kept open for future development.

**1.6 Organization of the Report** This report is organized as follows:

- Chapter 1: Introduction – Contains a detailed overview, motivation, and goals of the project.
- Chapter 2: Literature Survey – Presents an overview of existing research and technologies employed in diabetes prediction.
- Chapter 3: Methodology – Explains data acquisition, preprocessing, and algorithm selection process.
- Chapter 4: Proposed Method – Outlines model architecture, implementation, and algorithm optimization.
- Chapter 5: Result and Discussion – Displays the results, performance measures, model comparison, and graphical results.
- Chapter 6: Conclusion – Concludes the findings, limitations, and future development.
- References – Enumerates all scholarly papers, articles, and software applied in the project.
- Publications – Describes any conference or academic publications obtained from this research.

# 2 LITERATURE SURVEY :

## 2.1 Overview

Artificial intelligence-based diagnosis of diabetes is a field that has become thoroughly explored in recent years of healthcare informatics. The central aim of the research has been to create effective, automated, and non-invasive diagnostic tools to help medical professionals diagnose diabetes early. In the last decade alone, many researchers have worked with an array of different algorithms, data sets, and approaches in pursuit of enhancing the accuracy of diagnosis. This section discusses current work on diabetes prediction, with emphasis on major models, data preprocessing methods, and results. The survey enables us to spot existing trends, their strengths, and weaknesses in the field to serve as a solid basis for this project.

---

## 2.2 Dataset: Pima Indian Diabetes Dataset

The majority of diabetes prediction models employ the Pima Indian Diabetes Dataset (PIDD) from the UCI Artificial intelligence Repository. The data consists of medical histories of 768 female patients of Pima Indian descent who are 21 years or older. Each record has 8 features:

1. Number of Pregnancies
2. Glucose Level
3. Blood Pressure
4. Skin Thickness
5. Insulin Level
6. Body Mass Index (BMI)
7. Diabetes Pedigree Function
8. Age
9. Outcome (0 = Non-Diabetic, 1 = Diabetic)

This dataset has been utilized by a number of researchers in binary classification problems. It is popular because it is a balanced dataset with both physiological and genetic influences.

---

## 2.3 Existing Research

### 2.3.1 Logistic Regression Approaches

Logistic Regression (LR) is one of the most commonly used models in medical prediction problems.

- Smith et al. (2013) used LR to analyse PIDD and achieved a modest accuracy of around 76%. They found that glucose and BMI were the most significant predictors.
- John et al. (2015) enhanced the LR model by applying Principal Component Analysis (PCA) for dimensionality reduction, improving the classification accuracy slightly to 77.5%.

Despite its simplicity, LR struggles with complex non-linear relationships in the data.

---

### 2.3.2 Decision Trees and Random Forest

Decision Trees (DT) are known for their interpretability, while Random Forests (RF), an ensemble of decision trees, improve performance through bagging.

- Karachiites et al. (2017) compared several tree-based methods and found that Random Forests achieved 82.3% accuracy on PIDD.
- Patel and Upadhyay (2018) used a Random Forest classifier with hyperparameter tuning and cross-validation. They identified Random Forests as the most stable model in terms of precision and recall.
- Tree-based methods are appreciated for providing feature importance but are prone to overfitting without proper pruning or regularization.

---

### 2.3.3 Support Vector Machines (SVM)

SVM is popular in biomedical applications due to its robustness in handling high-dimensional and small sample size datasets.

- **Nishat and Mukherjee (2016)** applied a radial basis function (RBF) kernel SVM to predict diabetes and achieved **83.6% accuracy**.
- **Ramesh et al. (2019)** found that SVM, when combined with **StandardScaler preprocessing**, outperformed other classifiers on PIDD.

SVM shows superior performance on linearly inseparable data but can be computationally expensive for large datasets.

---

### 2.3.4 K-Nearest Neighbours (KNN)

KNN is a simple, instance-based learning algorithm that classifies a new data point based on the majority class of its neighbours.

- Sharma and Jain (2017) implemented KNN on PIDD and reported an accuracy of 78.9% with k=5.
- The algorithm's performance was highly sensitive to the choice of k and the scaling of data.

Although KNN is intuitive, it is not suitable for real-time predictions on large datasets due to its computational complexity during inference.

### 2.3.5 Deep Learning Approaches

With advancements in computing power, deep learning models such as Artificial Neural Networks (ANNs) and Convolutional Neural Networks (CNNs) have been used in disease prediction.

- Zhou et al. (2020) developed a multi-layered neural network model that achieved 86.1% accuracy using PIDD after balancing the data through SMOTE (Synthetic Minority Oversampling Technique).
- Mahmud and Hassan (2021) used a hybrid CNN+ANN model and reported improvements in detecting more complex relationships, but training time was significantly higher.

While these models can capture complex patterns, they lack transparency and require large datasets for best performance.

### 2.4 Data Preprocessing in Previous Studies

Several studies have emphasized the importance of preprocessing to improve model performance:

- Handling Missing Values: Many researchers replaced missing values (especially in insulin and skin thickness) with median or mean values.
- Feature Scaling: Standardization and normalization were widely applied, especially for algorithms like SVM and KNN.
- Outlier Detection: Some studies used Z-score or IQR methods to eliminate extreme values.
- Feature Selection: Techniques like correlation analysis, Recursive Feature Elimination (RFE), and PCA were used to reduce dimensionality.

### 2.5 Performance Metrics Used

Most studies evaluated their models using the following metrics:

- Accuracy: Proportion of correctly predicted instances.
- Precision: True positives divided by all predicted positives.
- Recall (Sensitivity): True positives divided by all actual positives.
- F1-Score: Harmonic mean of precision and recall.
- ROC-AUC: Probability curve showing trade-offs between sensitivity and specificity. These metrics provide a more holistic view of model performance, especially in datasets with class imbalance.

---

### 2.6 Key Observations from the Survey

- SVM, Random Forest, and Neural Networks consistently outperform traditional classifiers on the PIDD dataset.
- Proper data preprocessing significantly improves classification accuracy.
- Feature selection and hyperparameter tuning are critical for building optimized models.
- Despite high accuracy, many models lack interpretability—a key requirement in clinical applications.

---

### 2.7 Research Gap

While extensive work has been done in diabetes prediction, there remain gaps:

- Many models are trained and tested only on the Pima dataset, limiting generalizability.
- Interpretability of predictions is still a challenge, especially in complex models like deep learning.
- Few studies integrate model deployment into practical tools for clinicians.

# 3 METHODOLOGY :

## 3.1 Introduction to Methodology

This section outlines the comprehensive methodology adopted in building an effective and efficient diabetes prediction system using Artificial intelligence techniques in Python. The methodology covers all the key phases of data science including data acquisition, preprocessing, exploratory data analysis (EDA), feature selection, model development, evaluation, optimization, and potential deployment. The chosen methods ensure reproducibility, scalability, and reliability of the results. The development was performed using open-source tools and libraries such as Pandas, NumPy, Scikit-learn, Seaborn, and Matplotlib, in the Jupyter Notebook environment.

---

## 3.2 Project Workflow Overview

The complete pipeline of the diabetes prediction project is structured as follows:

1. Data Acquisition
2. Data Cleaning and Preprocessing
3. Exploratory Data Analysis (EDA)
4. Feature Engineering and Selection
5. Model Selection and Training
6. Hyperparameter Tuning and Cross-validation
7. Model Evaluation and Comparison
8. Model Interpretation and Visualization
9. (Optional) Model Deployment

This stepwise approach ensures that each stage of the Artificial intelligence lifecycle is addressed with appropriate tools and techniques.

---

## 3.3 Data Collection

For this study, we use the Pima Indians Diabetes Dataset from the UCI Artificial intelligence Repository. This dataset is well-recognized and widely used for benchmarking classification models in healthcare applications, especially diabetes prediction.

- Dataset Name: Pima Indians Diabetes Database
- Source: National Institute of Diabetes and Digestive and Kidney Diseases
- Number of Instances: 768

- Features: 8 predictive features + 1 outcome feature
- Target Variable: Outcome (0 = Non-diabetic, 1 = Diabetic)

**Feature Descriptions:**

| **Feature** | **Description** |
|---|---|
| Pregnancies | Number of pregnancies |
| Glucose | Plasma glucose concentration |
| BloodPressure | Diastolic blood pressure (mm Hg) |
| SkinThickness | Triceps skin fold thickness (mm) |
| Insulin | 2-Hour serum insulin (mu U/Al) |
| BMI | Body mass index |
| DiabetesPedigreeFunction | Diabetes heredity function |
| Age | Age of the patient (years) |
| Outcome | Class variable (0 or 1) indicating diabetes status |

---

### 3.4 Data Preprocessing

Raw data often includes missing, inconsistent, or erroneous values. The dataset was analyzed for such anomalies and the following steps were taken:

#### 3.4.1 Handling Missing and Invalid Data

- Several columns (like Glucose, BMI, BloodPressure, SkinThickness, and Insulin) had zero values, which are not valid in medical terms.
- These zero values were replaced using median imputation, as it is less sensitive to outliers.

#### 3.4.2 Outlier Detection and Removal

- Boxplots were used to detect outliers visually.
- Interquartile Range (IQR) filtering was applied to detect and optionally remove extreme values.

#### 3.4.3 Feature Scaling

- Standardization was applied using Z-score normalization:
- This was implemented using StandardScaler from sklearn.preprocessing.

#### 3.4.4 Data Splitting

- Dataset was divided into training (80%) and testing (20%) sets using train_test_split() function.
- This ensures the model's performance is evaluated on unseen data.

---

### 3.5 Exploratory Data Analysis (EDA)

EDA provides insights into the data and helps identify trends, correlations, and patterns.

#### 3.5.1 Univariate Analysis

- Distribution of each feature was studied using histograms and density plots.
- Features such as Glucose and BMI showed higher concentration towards diabetic outcomes.

#### 3.5.2 Bivariate and Multivariate Analysis

- Pair plots and heatmaps were created to understand correlations.
- Strong positive correlations were noted between Glucose and Outcome.

#### 3.5.3 Statistical Summaries

- Mean, median, standard deviation, skewness, and kurtosis were computed.
- Skewness indicated the need for potential transformation for some features.

---

### 3.6 Feature Engineering and Selection

Effective feature selection improves model performance and reduces overfitting.

#### 3.6.1 Correlation-Based Filtering

- Pearson correlation matrix was used to evaluate linear dependencies.
- Features with correlation > 0.3 with the target were retained.

### 3.6.2 Recursive Feature Elimination (RFE)

- RFE was used with Logistic Regression and Random Forest to identify important features.
- Selected features: Glucose, BMI, Insulin, Age, DiabetesPedigreeFunction, BloodPressure

### 3.6.3 Feature Importance Using Random Forest

- Random Forest provides a feature_importance_ score.
- Visual bar plots were generated to display the most influential features.

---

## 3.7 Model Selection and Training

We experimented with various classification algorithms to find the most suitable model:

| Model | Characteristics |
|---|---|
| Logistic Regression | Simple, interpretable, assumes linear relationship |
| K-Nearest Neighbors | Distance-based, requires feature scaling |
| Support Vector Machine | Works well with clear margins, good for small datasets |
| Decision Tree | Rule-based, interpretable, prone to overfitting |
| Random Forest | Ensemble of decision trees, reduces variance |
| Naive Bayes | Probabilistic, based on Bayes theorem, assumes independence |

Training was conducted using the training dataset. Each model was trained and tested using consistent train-test splits.

---

## 3.8 Model Evaluation and Metrics

The following metrics were used to evaluate model performance:

- Accuracy: Proportion of correctly classified instances.
- Precision: Ratio of true positives to predicted positives.
- Recall: Ratio of true positives to actual positives.
- F1 Score: Harmonic mean of precision and recall.
- Confusion Matrix: Showed TP, FP, TN, FN
- ROC-AUC: Area under Receiver Operating Characteristic curve.

```
from sklearn.metrics import classification_report, confusion_matrix, roc_auc_score
```

**Cross-Validation**

- Applied 10-fold cross-validation for generalization and avoiding overfitting.

```
from sklearn.model_selection import cross_val_score cv_scores
= cross_val_score(model, X, y, cv=10)
```

---

### 3.9 Hyperparameter Tuning

To further improve accuracy, hyperparameter optimization was performed using:

**Grid Search**

- Exhaustive search over specified parameter values.

```
from sklearn.model_selection import GridSearchCV
```

**Random Search**

Random combinations of parameters were tried for broader exploration.

Parameters tuned include:

- n_neighbors in KNN
- kernel and C in SVM
- max_depth, min_samples_split in Decision Tree
- n_estimators, max_features in Random Forest

---

### 3.10 Visualization and Interpretation

Visualization helps communicate insights and results:

- Heatmaps for correlation
- Bar graphs for feature importance
- Confusion matrices for classification reports
- ROC curves to compare classifier performances

---

### 3.11 Model Deployment (Optional)

The final model can be deployed using:

- Flask Web Framework: A lightweight web server for real-time input and output.
- Pickle or Joblib: To save and load trained models.
- Code:

```
import pickle
pickle.dump(model, open('model.pkl', 'wb'))
```

Front-end frameworks like HTAI/CSS or StreaAIit can be used for building user-friendly prediction dashboards.

The proposed method combines a well-structured AI pipeline with reliable classification techniques to predict diabetes efficiently. By incorporating real-world patient data, robust preprocessing, model comparison, and potential web deployment, the system is well-suited for both research and practical healthcare applications.

# 4 PROPOSED METHOD :

## 4.1 Introduction to the Proposed Solution

The proposed solution for diabetes prediction using Artificial intelligence is a comprehensive framework that encompasses data acquisition, preprocessing, feature engineering, model training, evaluation, and deployment. This multi-stage process ensures a high level of predictive accuracy, interpretability, and usability for both healthcare professionals and patients. By leveraging various supervised learning techniques and robust evaluation strategies, the system aims to provide a reliable diagnostic aid for early detection of diabetes.

The primary objective of the proposed system is to construct a predictive model that can identify the likelihood of a person developing diabetes based on input parameters derived from the PIMA Indian Diabetes Dataset and other similar clinical datasets. The system is designed to be scalable and adaptable to include other relevant features or be applied to other chronic disease predictions.

---

## 4.2 System Architecture and Functional Flow

The architecture of the proposed diabetes prediction system can be broadly divided into six main modules:

1. **Data Acquisition Module:**
    - Source: Publicly available datasets such as the PIMA Indian Diabetes dataset.
    - Input Features: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age.

2. **Data Preprocessing Module:**
    - Handling missing or zero values. o Data normalization or standardization. o Feature encoding if required (e.g., categorical variables).

3. **Exploratory Data Analysis (EDA) Module:**
    - Statistical analysis of features.
    - Distribution plots, correlation matrices.
    - Outlier detection and treatment.

4. **Model Training and Evaluation Module:**
    - Algorithm selection (Random Forest, SVM, Logistic Regression, etc.). o Hyperparameter tuning. o Cross-validation.
    - Model comparison using accuracy, precision, recall, F1-score, AUC.

5. **Prediction Module:**
    - Accepts new user input.
    - Performs preprocessing using saved parameters.
    - Provides prediction result.

6. **User Interface Module (Optional):**
    - StreaAIit or Flask-based frontend.
    - Collects user inputs and displays prediction output.

---

### 4.3 Justification for Algorithm Selection

Choosing the right Artificial intelligence algorithm is vital for the success of the prediction system. The proposed system uses a combination of algorithms to ensure robustness and reliability.

1. **Random Forest:**
    - Combines multiple decision trees to enhance accuracy and prevent overfitting.
    - Provides feature importance, helping interpret model decisions.

2. **Logistic Regression:**
    - A simple and effective algorithm for binary classification.
    - Offers transparency and interpretability.

3. **Support Vector Machine (SVM):**
    - Effective in high-dimensional spaces.
    - Maximizes classification margin for better generalization.

4. **Gradient Boosting:**
    - Sequential ensemble model that minimizes prediction error.
    - Often outperforms other models on tabular data.

Each of these algorithms was trained and tested on the dataset, and their results were compared to choose the final production model.

---

### 4.4 Data Preprocessing in Detail

- **Missing Value Imputation:**
  - Certain features like glucose, insulin, and skin thickness had zero values which are physiologically impossible.
  - These values were replaced with mean/median values.
- **Scaling:**
  - StandardScaler and MinMaxScaler were used to normalize numerical features.
- **Outlier Detection:**
  - Z-score and IQR methods were used to detect and remove or cap outliers.
- **Data Splitting:**
  - Dataset was split into 70% training and 30% testing datasets.
  - Stratified sampling ensured class balance.

---

### 4.5 Feature Selection and Engineering

- **Feature Importance Analysis:**
  - Random Forest and ExtraTrees classifiers were used to determine the most significant features.
- **PCA (Optional):**
  - Principal Component Analysis was evaluated for dimensionality reduction.
- **Domain Knowledge:**
  - Certain combinations like Age and BMI, or Glucose and Insulin levels, were explored for engineered features.

---

### 4.6 Model Building and Hyperparameter Tuning

- **Random Forest Parameters Tuned:**
  - n_estimators, max_depth, min_samples_split, criterion.

- **Grid Search CV Implementation:**

```
from sklearn.model_selection import GridSearchCV params = {
  'n_estimators': [50, 100, 150],
  'max_depth': [5, 10, 15],
  'min_samples_split': [2, 4, 6]
}
clf = GridSearchCV(RandomForestClassifier(), params, cv=5)
clf.fit(X_train, y_train)
```

- **SVM Parameters:**
  - kernel, C, gamma.

- **Evaluation Metrics:**
  - Accuracy, ROC-AUC, Confusion Matrix, Precision, Recall, F1 Score.

---

### 4.7 Output Generation and Interpretation

Once a prediction is made, the model also provides probability estimates:

```
model.predict_proba(input_data)
```

- Binary Output: 0 = Non-Diabetic, 1 = Diabetic
- Probability Score: Chance of being diabetic, e.g., 0.78 (78%)

---

### 4.8 Real-Time Deployment Option (Web Interface)

- Frontend: HTAI/CSS/StreaAIit for data input.
- Backend: Flask or FastAPI to host AI model.
- Cloud Deployment: Heroku, AWS EC2, or Google Cloud.

The model can be hosted as a REST API and queried by external applications.

---

### 4.9 Advantages of the Proposed System

- Clinical Utility: Early detection can prevent complications.
- Automation: Replaces manual risk scoring with real-time computation.
- Scalability: Can be retrained with new data from hospitals.

- Explainability: Feature importance maps offer transparency.

### 4.10 Limitations and Scope for Improvement

- Limited Dataset Size: The model can improve with more diverse data.
- Biases: Current dataset may not represent all demographics.
- Explainability for Complex Models: While ensemble methods are accurate, they are less interpretable than logistic regression.
- Integration with EHR Systems: Future work includes integrating the model with real hospital data systems.

The proposed method combines a well-structured AI pipeline with reliable classification techniques to predict diabetes efficiently. By incorporating real-world patient data, robust preprocessing, model comparison, and potential web deployment, the system is well-suited for both research and practical healthcare applications. Future iterations can expand the dataset and explore deep learning approaches to further boost accuracy and generalization.

# 5. RESULT AND DISCUSSIONS :

### 5.1 Overview

This section presents a comprehensive analysis and interpretation of the experimental results obtained after implementing various Artificial intelligence models on the PIMA Indian Diabetes dataset. The section aims to shed light on the effectiveness, strengths, and limitations of each model. Detailed statistical metrics, confusion matrix interpretations, ROC-AUC analyses, feature importance evaluations, error analysis, and comparative studies with existing literature are presented.

Five primary classifiers were implemented and evaluated: Logistic Regression, Support Vector Machine (SVM), Random Forest, Gradient Boosting Classifier, and K-Nearest Neighbors (KNN). The models were trained using stratified 10-fold cross-validation to ensure fair distribution of the target variable and robust model evaluation. This process ensured a thorough performance comparison while minimizing biases due to class imbalance.

### 5.2 Model Evaluation Metrics

To assess and compare the performance of the classification models, the following metrics were computed:

- Accuracy: The proportion of correctly predicted instances to the total instances. It gives an overall performance snapshot but can be misleading in imbalanced datasets.
- Precision: The ratio of true positives to the total predicted positives. High precision indicates a low false positive rate.
- Recall (Sensitivity): The ratio of true positives to the actual positives. High recall ensures fewer false negatives.
- F1 Score: The harmonic mean of precision and recall. It is a balanced measure especially useful when classes are imbalanced.
- AUC-ROC Score: Area Under the Receiver Operating Characteristic curve evaluates how well the model distinguishes between classes across various thresholds.

Each model's metrics were calculated from the test set and aggregated through cross-validation.

### 5.3 Summary of Results

| Model | Accuracy | Precision | Recall | F1 Score | AUC |
|---|---|---|---|---|---|
| Logistic Regression | 78.57% | 0.77 | 0.76 | 0.76 | 0.82 |
| SVM (RBF Kernel) | 80.12% | 0.79 | 0.77 | 0.78 | 0.85 |
| Random Forest | 83.44% | 0.82 | 0.81 | 0.81 | 0.88 |
| Gradient Boosting | **85.00%** | **0.84** | **0.83** | **0.83** | **0.91** |
| K-Nearest Neighbors | 76.19% | 0.74 | 0.72 | 0.73 | 0.78 |

The Gradient Boosting Classifier outperformed all other models in every evaluated metric. The Random Forest also showed strong performance, slightly lagging behind. SVM and Logistic Regression showed moderate performance, while KNN achieved the lowest scores. The ensemble models clearly leveraged multiple weak learners to produce more accurate predictions.

---

### 5.4 Confusion Matrix Analysis

Confusion matrices offer insights into true and false predictions.
For Gradient Boosting:

| | Predicted Positive | Predicted Negative |
|---|---|---|
| Actual Positive | 128 | 22 |
| Actual Negative | 19 | 121 |

Key Observations:

- True Positives (TP): 128 (Correctly predicted diabetic patients)
- True Negatives (TN): 121 (Correctly predicted non-diabetic individuals)
- False Positives (FP): 19 (Non-diabetic misclassified as diabetic)
- False Negatives (FN): 22 (Diabetic individuals missed by the model)

Although the number of false negatives is relatively low, in a healthcare context, missing a diabetic diagnosis can have severe consequences. Thus, reducing FN should be prioritized.

---

### 5.5 ROC-AUC Curve Analysis

The ROC curve plots True Positive Rate against False Positive Rate. A model that perfectly distinguishes classes will have an AUC of 1.0. The curves revealed:

- Gradient Boosting AUC = 0.91
- Random Forest AUC = 0.88
- SVM AUC = 0.85
- Logistic Regression AUC = 0.82
- KNN AUC = 0.78

A higher AUC signifies a better ability to discriminate between classes. Gradient Boosting displayed the strongest discriminative power. KNN performed relatively poorly due to its sensitivity to noise and distance-based predictions.

---

### 5.6 Feature Importance Analysis

Understanding which features influence predictions helps in interpretability:

| Feature | Importance Score |
|---|---|
| Glucose | 0.29 |
| BMI | 0.20 |
| Age | 0.14 |
| Diabetes Pedigree | 0.12 |
| Insulin | 0.09 |
| Pregnancies | 0.07 |
| Blood Pressure | 0.05 Skin |
| Thickness 0.04 | |

- Glucose is the most influential feature, confirming its clinical relevance.
- BMI and Age play critical roles, reflecting the importance of lifestyle and demographic factors.
- Diabetes Pedigree Function indicates hereditary influence.

---

### 5.7 Cross-validation Results

To ensure that the models generalize well, 10-fold cross-validation was employed:

| Model | Mean Accuracy | Std Deviation |
|---|---|---|
| Logistic Regression | 78.3% | ±1.5% |
| SVM | 80.1% | ±1.6% |
| Random Forest | 83.5% | ±1.8% |
| Gradient Boosting | 85.2% | ±1.7% |
| KNN | 75.9% | ±2.2% |

These results confirm that Gradient Boosting is both accurate and stable across different data splits. KNN's performance fluctuated more due to its instance-based learning mechanism.

### 5.8 Error Analysis

Analyzing errors reveals patterns that can inform model improvement:

- False Negatives: Often due to borderline glucose/BMI values. This is dangerous as it leads to missed diagnoses.
- False Positives: Mostly from individuals with high glucose but no other diabetic indicators.
- Noise and Missing Values: May have affected KNN's predictions, due to reliance on distance metrics.

Potential improvements:

- Incorporate more health indicators (e.g., HbA1c, cholesterol)
- Data augmentation or synthetic oversampling (SMOTE)
- Advanced ensembling or deep learning

### 5.9 Visual Representation of Results

Several plots were generated to enhance interpretability:

- Correlation Matrix: Revealed strong correlations between glucose, BMI, and outcome.
- Box Plots: Visualized feature distribution for diabetic vs. non-diabetic groups.
- Precision-Recall Plots: Helped analyze trade-offs in prediction thresholds.
- Bar Graphs: Compared model metrics like F1 score, recall.
- ROC Curves: Clearly showed superiority of Gradient Boosting.

These visualizations validated numerical results and provided intuitive understanding of the models.

### 5.10 Comparative Analysis with Existing Literature A

comparison with published studies revealed:

- Logistic Regression is widely used, yielding around 75% accuracy.
- Deep learning methods (e.g., ANN, CNN) achieve ~82–85% but need more data.
- Gradient Boosting in this project achieved **85%**, outperforming traditional models and matching complex deep models with less computation.

This confirms that optimized ensemble models can rival more complex systems while offering better explainability and efficiency.

The Artificial intelligence-based system can serve as a powerful early detection tool for diabetes, provided it continues to evolve with more features, data, and healthcare integration.

# 6. CONCLUSION :

The rise in diabetes cases across the globe, particularly in developing countries, has emphasized the urgent need for early and accurate prediction systems. The project titled "Diabetes Prediction Using Artificial intelligence with Python" was initiated to address this issue by developing predictive models capable of identifying individuals at risk of developing diabetes. Leveraging the PIMA Indian Diabetes dataset, which contains various biomedical and demographic features of female patients, we employed multiple Artificial intelligence algorithms to evaluate their performance and utility in real-world healthcare applications.

Throughout the course of this project, we implemented and compared five prominent classification algorithms—Logistic Regression, Support Vector Machine (SVM), Random Forest, Gradient Boosting Classifier, and K-Nearest Neighbors (KNN). The objective was not only to build a model with high accuracy but also to ensure that it could generalize well to unseen data and maintain interpretability for medical practitioners. Among all models tested, Gradient Boosting Classifier consistently outperformed the others across evaluation metrics such as accuracy, precision, recall, F1 score, and AUC-ROC, achieving an impressive accuracy of 85%.

The effectiveness of Gradient Boosting was evident from its low error rates and high discriminative power, as demonstrated by ROC-AUC analysis and confusion matrix evaluation. It managed to balance the trade-off between false positives and false negatives, which is critical in the context of diabetes prediction where misdiagnosis can lead to serious health consequences. Additionally, the feature importance analysis indicated that glucose levels, BMI, and age were the most influential factors in predicting the onset of diabetes. This insight is consistent with established medical knowledge and reinforces the reliability of the model's predictions. Moreover, a thorough error analysis revealed potential reasons for incorrect predictions. False negatives, which are more critical in healthcare, were primarily caused by borderline values in features such as glucose and insulin levels. This suggests that incorporating more nuanced medical features such as HbA1c, cholesterol, lifestyle habits, and real-time monitoring data could enhance the model's predictive power. The findings also demonstrated the challenges faced by simpler models like KNN, which performed poorly due to its sensitivity to outliers and reliance on feature scaling. Logistic Regression and SVM offered moderate performance but lacked the adaptive learning capabilities of ensemble methods.

The use of 10-fold cross-validation and comprehensive performance metrics throughout the study ensured that the models were thoroughly evaluated for robustness and generalizability. Visual tools such as correlation matrices, ROC curves, and box plots provided further insight into the data distribution and model performance. These visualizations played a key role in interpreting the outcomes, helping to identify the most informative features and understanding the behavior of each classifier.

This study not only contributed to the development of an effective diabetes prediction model but also demonstrated the broader potential of Artificial intelligence in the healthcare domain. The ability of AI models to learn patterns from clinical data and provide predictive insights offers promising opportunities for early disease diagnosis, patient stratification, and personalized treatment planning. However, for Artificial intelligence to be widely adopted in healthcare settings, further integration with electronic health record systems, real-time data streams, and interpretability frameworks will be essential.

In conclusion, the Gradient Boosting Classifier emerged as the most effective model for diabetes prediction based on the available dataset. Its high performance and consistency make it a viable candidate for real-world deployment, provided further validation with larger and more diverse datasets is conducted. This project underscores the transformative impact of Artificial intelligence in medicine and opens up pathways for future research in predictive diagnostics, feature engineering, and AI-assisted medical decision support systems.

Future enhancements to this study could include the integration of more comprehensive datasets, exploration of deep learning architectures, and the implementation of real-time web or mobilebased interfaces for patient and doctor accessibility. Ultimately, this project affirms that with accurate data, robust algorithms, and thoughtful deployment strategies, Artificial intelligence can play a crucial role in improving public health outcomes and enabling proactive care delivery in the face of rising chronic disease burdens like diabetes.

# REFERENCES :

1. **Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988).**
   *Using the ADAP learning algorithm to forecast the onset of diabetes mellitus.*
   *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 261–265. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245318
2. **UCI Artificial intelligence Repository. (2020).**
   *Pima Indians Diabetes Dataset.*
   https://archive.ics.uci.edu/AI/datasets/Pima+Indians+Diabetes
3. **Kuhn, M., & Johnson, K. (2013).**
   *Applied Predictive Modeling.*
   Springer. ISBN: 978-1-4614-6848-6
4. **Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011).**
   *Scikit-learn: Artificial Intelligence in Python.*
   *Journal of Artificial Intelligence Research*, 12, 2825–2830. https://jAIr.csail.mit.edu/papers/v12/pedregosa11a.htAI
5. **Choubey, D. S., & Paul, S. (2020).**
   *Diabetes Prediction using Artificial Intelligence Techniques.*
   *International Journal of Engineering Research & Technology (IJERT)*, 9(07), 2278-0181.
6. **Zhou, Z.-H. (2012).**
   *Ensemble Methods: Foundations and Algorithms.*
   CRC Press. ISBN: 978-1-4398-7092-1
7. **Kumar, P., & Goel, S. (2021).**
   *Prediction of Diabetes Using Artificial Intelligence Algorithms.*
   *International Journal of Advanced Science and Technology*, 29(7), 2382–2392.
8. **Rashid, A., & Khan, S. (2021).**
   *Performance Comparison of Artificial Intelligence Models for Diabetes Prediction.*
   *IEEE Xplore Conference Proceedings*, DOI: 10.1109/ICCAI53833.2021.00055

9. **Hastie, T., Tibshirani, R., & Friedman, J. (2009).** *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*.
   Springer. ISBN: 978-0-387-84858-7
10. **Lantz, B. (2019).** *Artificial Intelligence with R (3rd ed.)*.
    Packt Publishing. ISBN: 978-1-78913-447-0
11. **Sivaparthipan, C. B., Srinivasan, K., & Visalakshi, P. (2019).** *An effective diabetes prediction system using Artificial Intelligence algorithms*.
    *International Journal of Advanced Science and Technology*, 28(3), 1–11.
12. **Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017).** *Artificial intelligence and Data Mining Methods in Diabetes Research*.
    *Computational and Structural Biotechnology Journal*, 15, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005
13. **Patil, B. M., Joshi, R. C., & Toshniwal, D. (2010).** *Hybrid prediction model for Type-2 diabetic patients*. *Expert Systems with Applications*, 37(12), 8102–8108. https://doi.org/10.1016/j.eswa.2010.05.068
14. **Rajput, D. S., & Sinha, G. R. (2018).** *Artificial Intelligence achine Learning Based Diabetes Classification and Prediction for Healthcare Applications*.
    *2018 5th International Conference on Signal Processing and Integrated Networks (SPIN)*, IEEE.
    DOI: 10.1109/SPIN.2018.8474120