# Diabetes Classification Project

**Author**: Parinha  
**Course**: Machine Learning with Python

## 1. Brief Summary of the Project

This project addresses the challenge of early diabetes detection through machine learning techniques. Using the Pima Indians Diabetes dataset, we develop predictive models to identify patients at risk of diabetes based on various medical indicators. The primary goal is to create an accurate classification model that can assist medical professionals in diagnosing diabetes at an early stage. This has significant potential for improving healthcare outcomes by enabling timely intervention and treatment strategies for at-risk patients.

## 2. Healthcare Impact of the Project

This project has significant implications for healthcare:

* **Early Detection**: By identifying high-risk individuals before symptoms become severe, interventions can be initiated earlier.
* **Resource Optimization**: Healthcare systems can prioritize resources for patients with higher predicted risk.
* **Personalized Medicine**: The risk factors identified can help tailor treatment approaches to individual patient profiles.
* **Public Health Planning**: Insights from the model can inform broader public health strategies for diabetes prevention.
* **Educational Tool**: The visualizations and analysis serve as educational resources for understanding diabetes risk factors.

## 3. Research Questions

The project seeks to answer the following research questions:

1. Which demographic and health indicators are most strongly associated with diabetes risk?
2. How accurately can machine learning models predict diabetes based on these indicators?
3. Which machine learning approach provides the best performance for diabetes classification?
4. What threshold of these indicators significantly increases the risk of diabetes?
5. How can these insights be translated into practical screening recommendations?

## 4. Approach to Answering the Research Questions

My approach involved a comprehensive machine learning pipeline:

1. **Data Understanding**: Exploring the Pima Indians Diabetes dataset to understand its structure and characteristics.
2. **Data Preprocessing**: Handling missing values, standardizing features, and preparing the data for modeling.
3. **Exploratory Data Analysis**: Visualizing relationships between features and the target variable.
4. **Model Training**: Implementing multiple machine learning models to compare their performance.
5. **Model Evaluation**: Assessing models using various metrics relevant to healthcare applications.
6. **Feature Importance Analysis**: Identifying which health indicators contribute most to diabetes risk.
7. **Model Selection**: Choosing the best performing model based on appropriate evaluation metrics.
8. **Deployment**: Creating a reusable pipeline for making predictions on new patient data.

## 5. Individual Contribution to the Project

My individual contributions to this project include:

* Implementing a comprehensive data preprocessing pipeline to handle missing and zero values in medical data.
* Creating informative data visualizations to explore relationships between health indicators and diabetes.
* Training and evaluating multiple machine learning models for diabetes classification.
* Developing a thorough model evaluation framework focused on healthcare-relevant metrics.
* Building a prediction system that can be used on new patient data.
* Documenting insights and findings that can inform medical decision-making.

## 6. Details about the Dataset Used, Including Visualizations

The Pima Indians Diabetes dataset is a collection of medical data from 768 female patients of Pima Indian heritage, aged 21 years and older. It contains various health metrics and a binary outcome variable indicating whether the patient developed diabetes within five years.

### Dataset Columns:

* **Pregnancies**: Number of pregnancies the patient has had
* **Glucose**: Plasma glucose concentration (mg/dL)
* **BloodPressure**: Diastolic blood pressure (mm Hg)
* **SkinThickness**: Triceps skinfold thickness (mm)
* **Insulin**: 2-Hour serum insulin (mu U/ml)
* **BMI**: Body Mass Index (weight in kg/(height in m)²)
* **DiabetesPedigreeFunction**: Diabetes pedigree function (a function of diabetes history in relatives)
* **Age**: Age in years
* **Outcome**: Binary variable (1: has diabetes, 0: no diabetes)

### Data Preparation:

The dataset required careful preprocessing due to the presence of zero values in features that biologically cannot be zero (like Glucose or BMI). These were treated as missing data and replaced with median values.

### Visualizations:

Key visualizations in the project include: - Distribution of the target variable (diabetic vs. non-diabetic) - Feature distributions separated by outcome class - Correlation heatmap showing relationships between features - Pairplots of the most important features correlated with diabetes

These visualizations revealed that features like Glucose, BMI, and Age had strong associations with diabetes risk, while others showed more subtle relationships.

## 7. Machine Learning Model Chosen and Justification

This project is a binary classification task as we aim to predict a binary outcome: whether a patient has diabetes (1) or not (0).

After evaluating multiple models, **Random Forest** was selected as the final model for implementation based on its superior overall performance.

**Justification for Choice:** 1. **Balanced Performance**: Random Forest provides excellent balance between precision and recall, which is crucial in a healthcare context where both false positives and false negatives have significant implications. 2. **Feature Importance**: It provides valuable insights into feature importance, helping identify the most significant indicators for diabetes risk. 3. **Robustness to Overfitting**: As an ensemble method, Random Forest is less prone to overfitting compared to individual decision trees. 4. **Non-linearity**: It captures complex non-linear relationships between features that might be missed by simpler models like Logistic Regression. 5. **Minimal Hyperparameter Tuning**: Even with default parameters, Random Forest performs well, making it a practical choice.

## 8. Alternative Models Explored and Comparison

In addition to Random Forest, the following alternative models were implemented and evaluated:

1. **Logistic Regression**: A linear model that is simple, interpretable, and serves as a good baseline.
2. **Decision Tree**: A non-linear model that creates clear decision rules.
3. **Support Vector Machine (SVM)**: A powerful model capable of finding complex decision boundaries.

The models were compared based on multiple metrics including accuracy, precision, recall, F1 score, and ROC-AUC. While Logistic Regression performed reasonably well and offered interpretability advantages, Random Forest consistently showed superior performance across most metrics, particularly in terms of F1 score and ROC-AUC, which are important for balanced classification in healthcare contexts.

## 9. Evaluation Techniques Used

To ensure the model’s reliability and accuracy, I employed the following evaluation techniques:

### Train-Test Split

The dataset was divided into: - Training set (80%): Used to train the models - Testing set (20%): Used to evaluate model performance on unseen data

### Evaluation Metrics

Multiple metrics were used to provide a comprehensive assessment:

1. **Accuracy**: Overall correctness of predictions (correct predictions / total predictions)
2. **Precision**: Ability to avoid false positives (true positives / (true positives + false positives))
3. **Recall**: Ability to find all positive cases (true positives / (true positives + false negatives))
4. **F1 Score**: Harmonic mean of precision and recall, providing balance between the two
5. **ROC-AUC**: Area under the Receiver Operating Characteristic curve, measuring the model’s ability to discriminate between classes

### Confusion Matrix

This visualization provides a detailed breakdown of correct and incorrect classifications, allowing for analysis of: - True Positives: Correctly identified diabetic patients - False Positives: Non-diabetic patients incorrectly classified as diabetic - True Negatives: Correctly identified non-diabetic patients - False Negatives: Diabetic patients incorrectly classified as non-diabetic

## 10. Hyperparameter Tuning

To optimize the performance of our selected Random Forest model, I performed hyperparameter tuning using GridSearchCV, which systematically searches through a predefined set of hyperparameter values to find the optimal combination.

The hyperparameters tuned include: - Number of estimators (trees): [50, 100, 200] - Maximum depth of each tree: [None, 10, 20, 30] - Minimum samples required to split a node: [2, 5, 10] - Minimum samples required at a leaf node: [1, 2, 4]

The tuning process yielded an optimized model with improved F1 score compared to the default Random Forest implementation. The best parameters were determined based on 5-fold cross-validation to ensure robustness.

## 11. Accuracy/Performance of the Model

The tuned Random Forest model achieved excellent performance metrics on the test dataset:

* **Accuracy**: Between 75-80%, indicating good overall predictive performance
* **Precision**: Around 75-80%, showing strong ability to correctly identify diabetic patients
* **Recall**: Approximately 70-75%, demonstrating good capability to find positive cases
* **F1 Score**: Between 72-77%, representing a balanced performance between precision and recall
* **ROC-AUC**: Around 80-85%, indicating excellent discriminative ability

Feature importance analysis revealed that Glucose was the most important feature, followed by BMI, Age, and DiabetesPedigreeFunction. This aligns with medical knowledge about diabetes risk factors and provides actionable insights for healthcare professionals.

## 12. Assessment of Underfitting and Overfitting

To ensure our model is neither underfitting nor overfitting, I evaluated its performance on both the training and test datasets:

* **Training vs. Test Performance**: The difference between training and test set metrics was monitored to detect overfitting. A small gap (typically less than 5-10 percentage points) indicates good generalization.
* **Learning Curves**: Analysis of model performance across different training set sizes helped identify if the model would benefit from more data or was already saturated.
* **Model Complexity**: The optimal hyperparameters from tuning generally favored moderate complexity (e.g., reasonable tree depths), indicating a good balance between underfitting and overfitting.

The final model showed a small difference between training and test performance, suggesting good generalization capabilities without excessive overfitting to the training data.

## 13. Key Learnings from the Project

This project yielded several important insights:

1. **Feature Importance**: Glucose levels, BMI, and Age emerged as the most significant predictors of diabetes risk, which aligns with medical knowledge about risk factors.
2. **Data Quality Challenges**: Missing data indicated as zeros in medical datasets required careful preprocessing to avoid biasing the model.
3. **Model Selection Considerations**: While simpler models like Logistic Regression performed reasonably well, the Random Forest’s ability to capture non-linear relationships provided superior results.
4. **Balance of Metrics**: In healthcare applications, considering multiple performance metrics (not just accuracy) is crucial. The F1 score proved valuable as it balances precision and recall.
5. **Hyperparameter Sensitivity**: Random Forest performance improved with tuning, demonstrating the importance of optimization even for relatively robust algorithms.
6. **Interpretability Trade-offs**: While Random Forest performed well, its “black box” nature presents challenges for explaining predictions to medical professionals compared to simpler models.

## 14. Potential Usefulness to Society or Healthcare Industry

This project has several potential applications in healthcare:

1. **Screening Tool**: The model can be implemented as a preliminary screening tool to identify high-risk individuals who may benefit from more comprehensive testing.
2. **Resource Allocation**: Healthcare providers can use risk predictions to prioritize limited resources for diabetes prevention and management.
3. **Personalized Risk Assessment**: The feature importance analysis can help develop personalized risk profiles based on individual health metrics.
4. **Public Health Campaigns**: Insights about key risk factors can inform targeted public health education and prevention campaigns.
5. **Clinical Decision Support**: The model can be integrated into electronic health record systems to provide real-time risk assessments during patient consultations.
6. **Research Guidance**: The identified relationships between variables can guide further medical research into diabetes risk factors.
7. **Remote Healthcare**: The model can be deployed in telehealth applications where in-person diagnostic testing may be limited.

## 15. Future Extensions of the Project

This project could be extended in several valuable directions:

1. **Additional Features**: Incorporating more health indicators like family history details, lifestyle factors (diet, exercise), and socioeconomic factors.
2. **Temporal Analysis**: If longitudinal data becomes available, developing models that track how risk changes over time based on changing health metrics.
3. **Risk Stratification**: Evolving from binary classification to multi-class classification that identifies different levels of risk.
4. **Deep Learning Approaches**: Experimenting with neural networks to capture more complex patterns, especially if the dataset expands.
5. **Explainable AI Methods**: Implementing techniques like SHAP (SHapley Additive exPlanations) values to provide more interpretable predictions.
6. **Mobile Application**: Developing a user-friendly mobile app that allows individuals to assess their diabetes risk and receive personalized prevention recommendations.
7. **Integration with IoT Devices**: Connecting the model with wearable health monitors to provide continuous risk assessment based on real-time health data.

## 16. Conclusion Based on Findings

The diabetes classification project successfully developed a machine learning model capable of predicting diabetes risk with high accuracy using the Pima Indians dataset. The tuned Random Forest classifier achieved excellent performance metrics, demonstrating its potential as a valuable tool for healthcare professionals.

Key findings include:

1. **Prediction Performance**: The model achieved strong performance metrics, with particular strength in correctly identifying high-risk patients.
2. **Critical Indicators**: Plasma glucose concentration emerged as the single most important predictor, followed by BMI and age, confirming their clinical significance in diabetes risk assessment.
3. **Data Preprocessing Impact**: Proper handling of missing values significantly improved model performance, highlighting the importance of domain knowledge in data preparation.
4. **Model Generalization**: The tuned Random Forest showed good generalization ability with only a small gap between training and test performance, suggesting it would be reliable on new patient data.

This project demonstrates how machine learning can effectively support healthcare decision-making by providing accurate risk assessments based on readily available medical measurements. While not a replacement for clinical judgment, such tools can enhance early detection efforts and help prioritize preventive interventions for at-risk individuals.

## Code Implementation

The complete code implementation for this project can be found at: https://github.com/parinha/diabetes-classification-project

Key components of the implementation include: - Data preprocessing and feature engineering - Exploratory data analysis with visualizations - Model training and comparison - Hyperparameter tuning - Model evaluation - Prediction system for new patient data