

G H Patel College of Engineering & Technology

**(The Charutar Vidya Mandal (CVM) University)**

**New V. V. Nagar**

# DEPARTMENT OF COMPUTER ENGINEERING

**AI/ML Report**

**on**

***Disease Prediction from Patient Data Using Machine Learning***

# Submitted By

**Name of Student :** Madhav devaiya

**Enrolment Number :** 12302040503004

**Name of Student :** Viraj warhade

**Enrolment Number :** 12302040503012

# Guided By

**Dr. Priyang Bhatt**

# Artificial Intelligence & Machine Learning (202046702)

**A.Y. 2024-25 EVEN TERM**



# CERTIFICATE

This is to certify that the Mini Project Report titled **“****Disease Prediction from Patient Data Using Machine Learning”** has been carried out by **Madhav devaiya (12302040503004)** under guidance in partial fulfilment for the Degree of Bachelor of Technology in **Computer engineering, 6th Semester**, G H Patel College of Engineering & Technology, CVM University, New Vallabh Vidyanagar during the academic year 2024-25.

Dr. Priyang Bhatt Dr. Sudhir Vegad

Internal Guide Head of Department



# CERTIFICATE

This is to certify that the Mini Project Report titled **“Disease Prediction from Patient Data Using Machine Learning”** has been carried out by **Viraj warhade (12302040503012)** under guidance in partial fulfilment for the Degree of Bachelor of Technology in **Computer engineering, 6th Semester**, G H Patel College of Engineering & Technology, CVM University, New Vallabh Vidyanagar during the academic year 2024-25.

Dr. Priyang Bhatt Dr. Sudhir Vegad

Internal Guide Head of Department

**Title of the Project**

**Disease Prediction from Patient Data Using Machine Learning**

**Objective**

The objective of this project is to predict the likelihood of a patient having diabetes based on clinical attributes. The system uses a supervised machine learning model trained on health-related tabular data to classify whether a patient is diabetic or not. The end goal is to assist in early diagnosis and treatment planning using interpretable models and metrics.

**Dataset Used**

* **Source**: Pima Indians Diabetes Database
* **Files**:
  1. diabetes.csv: Contains medical records for female patients of Pima Indian heritage.
* **Key Features:**
  + Pregnancies: Number of times pregnant
  + Glucose: Plasma glucose concentration
  + BloodPressure: Diastolic blood pressure
  + SkinThickness: Triceps skinfold thickness
  + Insulin: 2-Hour serum insulin
  + BMI: Body mass index
  + DiabetesPedigreeFunction: Diabetes likelihood based on family history
  + Age: Age of the patient
  + Outcome: 0 (Non-diabetic) or 1 (Diabetic)

**Model Chosen**

Random Forest Classifier for Binary Classification

1. **Data Preprocessing**:
   * Replaced invalid zero values in medical columns with NaN
   * Filled missing values with column-wise mean
   * Performed train-test split and scaled features using StandardScaler
2. **Model Training:**
   * Used RandomForestClassifier with class\_weight='balanced' to address class imbalance
3. **Model Saving:**
   * Trained model saved as a .pkl file for later use and deployment

**Performance Metrics**

* **Classification Report:**
  + Shows precision, recall, and F1-score for each class.
* **Confusion Matrix:**
  + Provides insights into false positives and false negatives.
* **Example Metrics:**
  + Accuracy: ~77%
  + Precision (Diabetic): ~76%
  + Recall (Diabetic): ~72%

**Challenges & Learnings**

**Challenges:**

1. **Missing Values:**
   * Zero values for glucose, insulin, BMI etc. were unrealistic and needed to be handled carefully.
2. **Class Imbalance:**
   * Number of non-diabetic patients outweighed diabetic ones, requiring model tuning.
3. **Model Interpretability:**
   * Understanding which features contributed most to predictions.

**Learnings:**

1. **Data Cleaning:**
   * Importance of domain knowledge when treating missing or invalid data.
2. **Modeling:**
   * Use of ensemble models like Random Forest for robust classification.
3. **Visualization:**
   * Interpreting feature importance, class distribution, and confusion matrices.

**Tools & Libraries**

* **Python:** Primary programming language
* **Libraries:** Pandas, NumPy, Seaborn, Matplotlib, Scikit-learn, Joblib
* **IDE:** Jupyter Notebook for EDA and prototyping
* **Environment:** Virtualenv or conda environment for isolation