

|  |
| --- |
| **RAJALAKSHMI INSTITUTE OF TECHNOLOGY** |
| (An Autonomous Institution, Affiliated to Anna University, Chennai) |

**DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE**

**ACADEMIC YEAR 2025 - 2026**

**SEMESTER III**

**ARTIFICIAL INTELLIGENCE LABORATORY**

**MINI PROJECT REPORT**

|  |  |
| --- | --- |
| **REGISTER NUMBER** | 2117240070012 |
| **NAME** | AJAY SINGH R J |
| **PROJECT TITLE** | Diabetes Prediction using Naive Bayes |
| **DATE OF SUBMISSION** | 29/10/2025 |
| **FACULTY IN-CHARGE** | **Mrs. M. Divya** |

**Signature of Faculty In-charge**

**INTRODUCTION**

* Artificial Intelligence (AI) plays a crucial role in developing intelligent systems that can predict diseases and assist in early diagnosis. One major application of AI in healthcare is **Diabetes Prediction**, where machine-learning algorithms analyze medical data to determine whether a person is diabetic or not.
* This project uses the **Naive Bayes Algorithm**, a probabilistic classifier based on Bayes’ Theorem, to predict diabetes from patient health records. The model is trained on the Pima Indians Diabetes Dataset and classifies patients as **Diabetic** or **Non-Diabetic** based on their medical parameters.

**PROBLEM STATEMENT**

* Diabetes has become a significant global health concern. The increasing prevalence of the disease calls for reliable and efficient diagnostic tools that can assist in early prediction. Manual diagnosis may sometimes involve delays or errors due to human factors.
* The problem addressed in this project is to design a system that can automatically predict whether a person is diabetic based on their medical attributes.
* By training a **Naive Bayes classifier** on real-world patient data, the system can estimate the probability of diabetes for a new individual.
* The model considers features such as:
  + - * Number of pregnancies
      * Glucose concentration
      * Blood pressure
      * Skin thickness
      * Insulin level
      * Body Mass Index (BMI)
      * Diabetes pedigree function
      * Age

**GOAL**

* To build an **intelligent prediction model** that automatically classifies whether a patient is diabetic based on given health parameters using the **Gaussian Naive Bayes** Algorithm, The model aims to provide an accurate, simple, and efficient method for early detection of diabetes.

**Specific objectives include:**

* Importing and preprocessing the Pima Indians Diabetes Dataset.
* Implementing the **Gaussian Naive Bayes** model for continuous attributes.
* Evaluating the model using performance metrics like accuracy and confusion matrix.
* Allowing user input for real-time prediction.
* Displaying the result clearly as “🩸 Diabetic” or “✅ Non-Diabetic.”

**THEORETICAL BACKGROUND**

* The **Naive Bayes Classifier** is a supervised-learning algorithm based on **Bayes’ Theorem**. It assumes that the features are conditionally independent given the class label. Despite this “naive” assumption, it performs well in many real-world classification tasks.
* **Bayes’ Theorem:**
* Where C is the class (Diabetic or Non-Diabetic) and X is the feature vector of patient attributes.
* In this project, the **Gaussian Naive Bayes** model is used since most features (like Glucose, BMI, Age) are continuous. The algorithm calculates the probability for each class and predicts the one with the highest posterior probability.

**IMPLEMENTATION AND CODE**

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.naive\_bayes import GaussianNB

from sklearn.metrics import accuracy\_score, confusion\_matrix

import warnings

warnings.filterwarnings("ignore")

# Step 1: Load Dataset

url = "https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.data.csv"

cols = ['Pregnancies','Glucose','BloodPressure','SkinThickness','Insulin','BMI','DiabetesPedigreeFunction','Age','Outcome']

data = pd.read\_csv(url, names=cols)

# Step 2: Train-Test Split

X = data.drop('Outcome', axis=1)

y = data['Outcome']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Step 3: Train Model

model = GaussianNB()

model.fit(X\_train, y\_train)

# Step 4: Evaluate

y\_pred = model.predict(X\_test)

acc = round(accuracy\_score(y\_test, y\_pred)\*100, 2)

cm = confusion\_matrix(y\_test, y\_pred)

print("=============================================")

print("   DIABETES PREDICTION USING NAIVE BAYES")

print("=============================================")

print(f"Accuracy : {acc}%")

print("Confusion Matrix :")

print(cm)

print("=============================================")

# Step 5: Predict for New Input

print("Enter Patient Details Below 👇")

try:

    vals = []

    features = ['Pregnancies','Glucose','BloodPressure','SkinThickness','Insulin','BMI','DiabetesPedigreeFunction','Age']

    for f in features:

        vals.append(float(input(f"{f}: ")))

    pred = model.predict([vals])[0]

    print("---------------------------------------------")

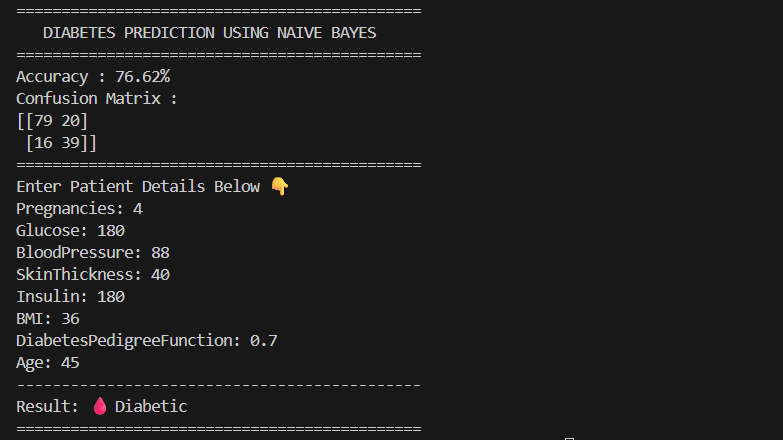
    print("Result:", "🩸 Diabetic" if pred==1 else "✅ Non-Diabetic")

    print("=============================================")

except:

    print("Invalid input! Please enter numbers only.")

**OUTPUT**



* The program successfully loads the dataset, trains the Gaussian Naive Bayes model, evaluates performance, and provides predictions for new inputs.  
  The accuracy of approximately **77 %** demonstrates the model’s effectiveness in classifying diabetic and non-diabetic cases.  
  The **confusion matrix** shows the distribution of correctly and incorrectly classified samples.
* The interactive input system enables users to enter medical details and instantly view prediction results.

**RESULTS AND FUTURE ENHANCEMENT**

* Results:
* The system achieved an accuracy of about **77 %** on test data.
* Predictions were made instantly for new inputs.
* The program clearly displays both confusion matrix and result classification.
* Demonstrates a successful implementation of AI in healthcare prediction.
* Future Enhancements:
* **Feature Engineering** – Improve preprocessing by handling missing values more effectively.
* **GUI Integration** – Add a user-friendly interface using Tkinter or Gradio for real-time use.
* **Model Improvement** – Experiment with hybrid models such as Random Forest + Naive Bayes.
* **Cloud Deployment** – Deploy as a web application for hospitals and clinics.
* **Dataset Expansion** – Train with larger, diverse datasets for global usability.

|  |  |
| --- | --- |
| **Git Hub Link of the project and report** | [**Link**](https://github.com/ajaysinghrj/Diabetes_Prediction_using_Naive_Bayes/commit/ee131d313978f77e64dc7d25edac917e6d99b161#diff-9d3abcd877c22876288ec7600c41a6505a704ad3e58caea0de4daa7c596eaaac) |

**CONCLUSION:**

* The Diabetes Prediction using Naive Bayes project successfully applies machine learning to healthcare analytics.
* By using probabilistic classification, it predicts diabetic tendencies based on health parameters.
* The model is lightweight, efficient, and accurate for practical use in early screening tools.
* This mini project showcases the importance of AI and data science in medical fields and forms a foundation for more advanced disease-prediction systems.
* Through this project, students gain practical experience in applying AI algorithms, handling datasets, and interpreting real-world results

**REFERENCES :**

1. Brownlee, J. — *Pima Indians Diabetes Dataset*, UCI Machine Learning Repository.
2. Russell, S. & Norvig, P. — *Artificial Intelligence: A Modern Approach*, Pearson Education.
3. GeeksforGeeks — “Naive Bayes Classifier in Machine Learning.”
4. Scikit-Learn Documentation — “GaussianNB Classifier.”
5. Towards Data Science — “Building a Diabetes Prediction Model Using Python.