***Github Link:*** [**https://github.com/422323104049/Data-Science-project**](https://github.com/422323104049/Data-Science-project)

**🧠Project Title**

***TRANSFORMING HEALTH CARE WITH AI-POWERED DISEASE PREDICTION BASED ON PATIENT DATA***

**PHASE 2**

**Name**  : SHERIF BASHA . A

**Register Number** : 422323104049

**Institution Name** :THIRUVALLUVAR COLLEGE OF ENGINEERING & TECHNOLOGY

**Department** : COMPUTER SCIENCE AND ENGINEERING

**Date of Submission** : 09:05:2025

**🩺 1. Problem Statement :**

Chronic and life-threatening diseases like diabetes, heart disease, and cancer often go undiagnosed until advanced stages. Early diagnosis can save lives and reduce healthcare costs significantly.

This project aims to build an AI-powered predictive model that analyzes historical patient data — including medical history, lab results, and lifestyle factors — to detect early warning signs of diseases. The goal is to support timely intervention and preventive care using machine learning.

The problem type may be classification (e.g., disease presence or absence) or regression (e.g., risk score).

Predicting diseases early based on patient health records and lifestyle information is a significant challenge in healthcare. Many life-threatening diseases, such as diabetes, heart disease, and cancer, can be managed more effectively with early diagnosis.

This project aims to develop a robust AI-based predictive model that can analyze patient data and accurately forecast disease likelihood.

**🎯 2. PROJECT OBJECTIVES :**

* Develop machine learning models to predict diseases based on historical patient data.
* Identify key health metrics and lifestyle factors contributing to early signs of disease.
* Provide insights into patterns across demographics and risk groups.
* Deliver a user-friendly interface for real-time prediction and doctor support.
* Prioritize model accuracy, interpretability, and ethical usage of patient data.

**🔄 3. Project Workflow (Flowchart):**

**Pic. (**Flowchat for Workflow)

**🧬 4. DATA DESCRIPTION :**

* **Dataset Name:** [e.g., Heart Disease UCI Dataset / Synthetic Health Records].
* **Source:** UCI, Kaggle, or hospital-generated synthetic data.
* **Records and Features:** Varies (typically 1000–50000 records with 10–30 features).
* **Target Variable:** Disease status (Yes/No or Risk Score).
* **Attributes:** Age, gender, blood pressure, glucose, cholesterol, lifestyle habits.
* **Type:** Structured, tabular.
* **Nature:** Static or semi-dynamic.

**🧹 5. DATA PREPROCESSING :**

* Handled missing values using imputation or median fill.
* Removed duplicate and irrelevant records (e.g., null diagnosis entries).
* One-hot encoded categorical variables like gender, smoking status.
* Standardized numeric features using StandardScaler.
* Detected and treated outliers in features like cholesterol and glucose levels.
* Converted categorical values (e.g., gender, smoking) to numeric using encoding.
* Handled outliers using Z-score and domain-specific thresholds.

**📈 6. EXPLORATORY DATA ANALYSIS (EDA):**

**Univariate Analysis:**

* Histograms of age, BMI, blood pressure.
* Count plots for diagnosis, smoking, and obesity status.

**Bivariate & Multivariate:**

* Correlation matrix to find links between vitals and disease
* Scatter plots of glucose vs. risk, age vs. diagnosis
* Bar plots for gender-based or age-group-based disease trends

**Key Insights:**

* Older age, high cholesterol, and smoking are strong risk factors.
* Strong correlation between blood pressure and heart disease.
* Diabetic patients show higher likelihood of multiple conditions.

**Distribution Analysis of Key Variables:**

* Plotted histograms and boxplots for age, glucose, cholesterol, and blood pressure to understand the distribution, detect outliers, and observe skewness.

**Trend and Risk Factor Identification:**

* Identified trends such as increasing disease likelihood with age, higher risk in smokers, and poor outcomes linked to obesity and hypertension.

**Correlation Matrix:**

* Generated a heatmap of feature correlations to identify strong relationships (e.g., high correlation between glucose and diabetes risk).

**🧪 7.FEATURE ENGINEERING :**

1. **Derived Risk Scores:**
   * Combined multiple clinical indicators (e.g., systolic BP, glucose, cholesterol) into a composite "risk score" to simplify prediction and highlight high-risk patients.
2. **BMI Categorization:**
   * Transformed continuous BMI values into categorical labels (e.g., Underweight, Normal, Overweight, Obese) to better capture health risk ranges.
3. **Interaction Features:**
   * Created new features based on interactions, such as age × cholesterol or smoking × glucose, to reveal complex patterns not visible in individual features.
4. **Handling Categorical Variables:**
   * Applied one-hot encoding to multi-class features like occupation or region, and label encoding for binary fields like gender or diabetic status.
5. **Normalization and Scaling:**
   * Scaled numeric features using StandardScaler or MinMaxScaler to bring variables to a uniform range for better model convergence.

**🤖 8.MODEL BUILDING:**

1. **Algorithm Selection:**
   * Used classification models like Logistic Regression, Random Forest, and XGBoost to predict disease presence based on structured patient data.
2. **Baseline Model Development:**
   * Started with Logistic Regression to establish a baseline performance due to its simplicity and interpretability.
3. **Advanced Model Training:**
   * Trained ensemble models (Random Forest, XGBoost) to capture non-linear relationships and improve predictive accuracy.
4. **Handling Imbalanced Data:**
   * Applied techniques like SMOTE (Synthetic Minority Over-sampling Technique) and class weighting to address imbalance in disease vs. no-disease classes.
5. **Cross-Validation Strategy:**
   * Implemented k-fold cross-validation to ensure robustness and avoid overfitting, while comparing model performance across different data splits.

**6.Hyperparameter Tuning:**

* Used GridSearchCV and RandomizedSearchCV to optimize model parameters such as tree depth, learning rate, and number of estimators, leading to improved model performance.

**📈 9.VISUALIZATION OF RESULTS & MODEL INSIGHTS:**

**1.Confusion Matrix Analysis**

* Displayed confusion matrix to understand model performance in terms of True Positives, False Positives, etc., helping assess diagnostic reliability.

**2.Feature Importance Plot**

* Visualized the most influential features (e.g., glucose level, age, blood pressure) using bar charts from models like Random Forest and XGBoost.

**3.ROC Curve and AUC Score**

* Plotted ROC curves for each model to evaluate classification threshold performance and computed AUC to quantify overall model accuracy.

**4.Prediction vs. Actual Chart**

* Used scatter plots or line graphs to compare predicted disease risk scores with actual labels for evaluating model consistency and bias.

🛠️ **10. TOOLS AND TECHNOLOGIES USED :**

**1.Python Programming Language:**

* Used for its versatility and rich ecosystem of libraries for data science, machine learning, and healthcare analytics.

**2.Google Colab / Jupyter Notebook:**

* Provided an interactive environment for data exploration, visualization, and model building with GPU support when needed.

**3.Key Libraries:**

* Utilized pandas and numpy for data manipulation, matplotlib and seaborn for visualizations, and scikit-learn, xgboost, tensorflow for machine learning.

**4.Data Visualization Tools:**

* Plotly and Seaborn were used to create interactive and publication-quality charts for analyzing trends and model performance.

**5.Deployment Tools:**

* Deployed the final predictive model using Streamlit or Gradio to create a user-friendly web interface for real-time disease risk predictions.

**6.Version Control with Git & GitHub**

* Used Git for tracking code changes and GitHub for collaborative development, issue tracking, and maintaining project documentation.

**👩‍🔬 11. TEAM MEMBERS AND CONTRIBUTIONS :**