Multiple Disease Prediction System Using Data Mining

Submitted in partial fulfillment of the requirements of the degree

B.TECH IN INFORMATION TECHNOLOGY

By

Riya Sunil Kharade **24101C2001**

Supervisor

Prof. Pallavi Kharat

VIT Vidyalankar Institute of Technology

Accredited A+ by NAAC

Department of Information Technology

Vidyalankar Institute of Technology

Vidyalankar Educational Campus,

Wadala(E), Mumbai - 400 037

University of Mumbai

(AY 2025-26)

CERTIFICATE

This is to certify that the Mini Project entitled “Multiple Disease Prediction System Using Data Mining” is a bonafide work of Riya Kharade (Roll No: 24101C2001. submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **“Bachelor of Engineering”** in **“Information Technology”.**

Prof. Pallavi Kharat

Supervisor

Internal Examiner	External Examiner
Name & Sign	Name & Sign

Table of Content

2.Abstract

The Multiple Disease Prediction System Using Data Mining is a web-based, intelligent platform designed to predict the likelihood of three major diseases: Diabetes, Heart Disease, and Parkinson’s Disease. Early disease detection is crucial in modern healthcare, as it helps improve patient outcomes, reduce treatment costs, and enable timely medical intervention. Traditional diagnostic methods often require multiple tests, expert consultation, and significant time, which may delay the detection of life-threatening conditions.

This project leverages data mining and machine learning techniques to analyze patient health data and provide fast, accurate, and reliable predictions. Three separate datasets—diabetes.csv, heart.csv, and parkinsons.csv—are used to train prediction models: Logistic Regression for Diabetes, Random Forest Classifier for Heart Disease, and Support Vector Machine (SVM) for Parkinson’s Disease. Each model achieves an accuracy of 85–92%, ensuring dependable results.

The system is implemented using Python and Streamlit, incorporating libraries such as NumPy, Pandas, Scikit-learn, and Pickle for data preprocessing, model training, and deployment. Key Data Warehousing and Data Mining (DWM) concepts, including data cleaning, normalization, feature selection, and classification algorithms, are applied to ensure structured and meaningful analysis.

The platform provides a user-friendly web interface, enabling users to input health parameters like blood pressure, glucose level, BMI, and vocal frequency patterns, and receive instant disease risk predictions. By integrating multiple prediction models into a single, scalable system, the project addresses limitations of existing tools, such as single-disease focus, offline operation, and lack of real-time prediction.

This system demonstrates the potential of data-driven healthcare solutions in improving preventive care, supporting doctors, researchers, and patients in making informed decisions, and offering a cost-effective and accessible approach to early disease detection. Future enhancements may include additional disease models, integration of patient history, and cloud deployment for wider accessibility.

3. Introduction

Introduction

The rapid growth of healthcare data has created opportunities to apply data mining and machine learning techniques for improving medical diagnosis and decision-making. Traditional diagnostic methods often require multiple tests, expert analysis, and time, which may delay early disease detection.

To overcome these challenges, the Multiple Disease Prediction System Using Data Mining is developed to provide an accurate, fast, and automated prediction of diseases based on patient health data. This system focuses on predicting three major diseases—Diabetes, Heart Disease, and Parkinson's Disease—which are among the most common health problems worldwide.

By leveraging machine learning, the system can analyze important medical attributes such as blood pressure, glucose levels, age, BMI, and vocal frequency patterns to identify risk patterns and relationships. Users can input their health parameters through a web-based interface, and the system instantly provides predictions, helping patients and healthcare professionals make timely decisions.

3.1 Background Information

Healthcare systems generate large volumes of patient data, which, if analyzed effectively, can provide valuable insights for early disease prediction. Machine learning models like Logistic Regression, Random Forest, and Support Vector Machine (SVM) have been proven effective in analyzing complex datasets and predicting health risks.

In this project:

- Diabetes prediction is handled by Logistic Regression, which analyzes factors like glucose, BMI, age, and blood pressure.
- Heart Disease prediction is performed using Random Forest, considering parameters like cholesterol, blood pressure, age, and resting heart rate.
- Parkinson's Disease prediction uses SVM, which evaluates vocal frequency patterns, jitter, and shimmer values.

The system integrates these models into a single web platform using Python and Streamlit, allowing for real-time predictions. This approach demonstrates practical applications of Data Warehousing and Data Mining (DWM) concepts such as data preprocessing, normalization, feature selection, and classification algorithms, ensuring high prediction accuracy and

3.2 Problem Statement

Current disease prediction tools often operate in isolation, are time-consuming, and lack accessibility, causing delays in the early diagnosis of life-altering conditions like diabetes, heart disease, and Parkinson's.

Our solution addresses these issues by:

- Integrating three separate disease prediction models into a single, unified platform.
- Enabling real-time, machine learning-powered predictions for all three diseases.
- Reducing the need for multiple separate tests and providing accurate (85–92%) and timely assessments.
- Delivering predictions through a user-friendly web interface, making the system accessible to both technical and non-technical users.

3.3 Objectives

1. **Unified Platform:** Develop a single web application capable of predicting Diabetes, Heart Disease, and Parkinson’s Disease simultaneously.
2. **High Accuracy Predictions:** Use validated machine learning models to achieve prediction accuracies above 85% for all three diseases.
3. **User-Friendly Interface:** Build an intuitive Streamlit-based interface for easy access by patients, caregivers, and healthcare professionals.
4. **Real-Time Risk Assessment:** Provide instant predictions with latency under 2 seconds, enabling quick decision-making.
5. **Scalability for Wider Access:** Design the backend to handle 100+ concurrent users, suitable for clinics, hospitals, or institutional deployment.

3.4 Scope of the Project

The Multiple Disease Prediction System provides significant benefits in predictive healthcare and demonstrates the potential of intelligent, data-driven systems. The scope of the project includes:

- Early Disease Detection: The system identifies risks for Diabetes, Heart Disease, and Parkinson’s Disease, allowing users to seek timely medical intervention.
- User-Friendly Web Application: The Streamlit-based interface enables users to input health parameters and receive predictions easily, without technical expertise.
- Fast and Automated Assessment: Predictions are delivered within seconds, reducing reliance on manual diagnosis and streamlining healthcare processes.
- Data-Driven Insights: Using multiple datasets, the system analyzes patterns across different demographics, helping researchers and doctors understand health trends.
- Scalability and Future Integration: The system is designed to accommodate additional disease models, support multiple concurrent users, and be deployed on cloud platforms like AWS, Heroku, or Streamlit Cloud for wider access.
- Privacy and Security: The system validates user inputs and ensures data integrity and confidentiality, aligning with modern healthcare data protection standards.

4. Literature Review

4.1 Existing Approaches

Several research studies and projects have explored disease prediction using machine learning and data mining techniques. Key approaches include:

- Diabetes Prediction: Logistic Regression, Decision Trees, and K-Nearest Neighbors (KNN) are commonly used to predict diabetes based on parameters like glucose level, BMI, age, and blood pressure. Many systems achieve 80–90% accuracy but often focus only on a single disease.
- Heart Disease Prediction: Random Forest, Support Vector Machine (SVM), and Naive Bayes classifiers are widely used for predicting heart disease risk. Existing systems analyze factors like cholesterol, blood pressure, and heart rate, yet few integrate real-time web-based interfaces.
- Parkinson’s Disease Prediction: SVM and Neural Networks are commonly used to detect Parkinson’s Disease from vocal features such as jitter, shimmer, and frequency patterns. Most studies provide offline prediction models without a user-friendly interface for patients or healthcare professionals.

These existing approaches demonstrate the potential of machine learning in predictive healthcare but are often limited to individual diseases or offline prediction.

a. Research Gaps

While existing systems have shown good prediction results, several gaps remain:

- Single-Disease Focus: Most tools focus on predicting only one disease at a time, requiring separate applications for multiple conditions.
- Limited Accessibility: Many models are offline and require technical knowledge to operate, limiting their use by patients.
- Integration Challenges: There is a lack of platforms that combine multiple disease prediction models in a single, user-friendly web interface.

- Real-Time Prediction: Few systems provide instant, real-time predictions suitable for clinical or personal use.
- Scalability and Reliability: Handling multiple concurrent users and ensuring robust performance is rarely addressed in existing studies.

b. Justification for Chosen Method

To address the above gaps, the proposed Multiple Disease Prediction System Using Data Mining:

- Integrates three disease prediction models (Diabetes, Heart Disease, Parkinson’s Disease) into a single web-based platform.
- Uses Logistic Regression, Random Forest, and SVM, chosen for their high accuracy and proven performance on the respective datasets.
- Provides real-time predictions with an easy-to-use Streamlit interface, making it accessible for patients, doctors, and researchers.
- Ensures scalability, supporting multiple concurrent users and offering potential cloud deployment for wider accessibility.
- Incorporates data preprocessing, feature selection, and normalization, following DWM concepts, to improve prediction reliability and model performance.

This approach bridges the gaps identified in previous research and demonstrates a practical application of data mining techniques in predictive healthcare.

5.Dataset Description

5.1 Data Source

The datasets used for the Multiple Disease Prediction System are publicly available and widely used in healthcare research:

- Diabetes Dataset (diabetes.csv): Sourced from the Pima Indians Diabetes Database, containing real-world patient data for predicting diabetes.
- Heart Disease Dataset (heart.csv): Collected from the Cleveland Heart Disease Database, widely used in medical research for predicting cardiovascular conditions.
- Parkinson's Disease Dataset (parkinsons.csv): Obtained from the UCI Machine Learning Repository, containing vocal measurements for early detection of Parkinson's Disease.

All datasets are cleaned, structured, and suitable for training machine learning models.

5.2 Data Characteristics

- Diabetes Dataset: Contains 768 patient records with 8 numeric attributes including glucose level, BMI, blood pressure, age, and insulin. The dataset is balanced with positive and negative diabetes cases.
- Heart Disease Dataset: Consists of 303 patient records with 14 attributes such as cholesterol, resting blood pressure, age, maximum heart rate, and chest pain type. It includes both categorical and numerical data.
- Parkinson's Disease Dataset: Includes 195 records with 23 vocal attributes such as jitter, shimmer, and frequency-related measures, along with a binary classification for Parkinson's presence.

All datasets vary in size, type, and features, providing a comprehensive view for multiple disease prediction.

5.3 Attributes and Features

Key features used for model training:

- Diabetes Prediction Features: Glucose, BMI, Age, Blood Pressure, Insulin, Skin Thickness, Pregnancies, Diabetes Pedigree Function.
- Heart Disease Prediction Features: Age, Sex, Chest Pain Type, Resting Blood Pressure, Cholesterol, Fasting Blood Sugar, Maximum Heart Rate, Exercise-Induced Angina, ST Depression, Slope of ST segment, Number of Major Vessels.
- Parkinson's Disease Features: Jitter (%), Jitter (Absolute), Shimmer, HNR (Harmonics-to-Noise Ratio), RPDE, DFA, Vocal Frequency Measures, Class Label (Parkinson's or Healthy).

Feature selection ensures that only relevant medical parameters are used for accurate predictions.

5.4 Preprocessing Requirements

Before feeding the datasets into machine learning models, the following preprocessing steps are applied:

- Handling Missing Values: Null or missing entries are filled using mean, median, or mode imputation.
- Data Normalization/Scaling: Numeric features are normalized using MinMaxScaler or StandardScaler to maintain uniformity.
- Categorical Encoding: Categorical variables such as gender or chest pain type are converted into numeric codes using Label Encoding or One-Hot Encoding.
- Duplicate Removal: Duplicate records are removed to avoid bias in model training.
- Feature Selection: Irrelevant or redundant features are removed to improve model performance and reduce computation time.

These preprocessing steps ensure that the datasets are clean, consistent, and ready for building high-accuracy predictive models

.

6.Methodology

6.1 Data Preprocessing

Data preprocessing is a crucial step to ensure the accuracy and reliability of machine learning models. The following steps were applied to the datasets:

- Handling Missing Values: Null or missing entries in the datasets were replaced using mean, median, or mode imputation depending on the attribute type.
- Normalization and Scaling: Continuous features such as glucose level, blood pressure, and vocal measurements were scaled using MinMaxScaler or StandardScaler to maintain uniformity.
- Categorical Encoding: Categorical attributes like gender and chest pain type were converted into numeric codes using Label Encoding or One-Hot Encoding.
- Duplicate Removal: Duplicate records were eliminated to maintain data integrity.
- Feature Selection: Only relevant medical parameters were retained to improve model accuracy and reduce computational overhead.

These steps ensure that the data is clean, consistent, and ready for model training.

6.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis was performed to understand the distribution and relationships within the datasets:

- Statistical Summary: Mean, median, standard deviation, and distribution of each attribute were analyzed.
- Correlation Analysis: Heatmaps and correlation matrices were used to identify important features that strongly influence disease outcomes.
- Visualizations:
  - Histograms and Boxplots were used to detect outliers and understand attribute distributions.

  - Scatterplots and Pairplots highlighted relationships between features such as BMI vs. glucose or age vs. heart rate.

- Outcome: EDA helped in selecting significant features and understanding patterns, which guided the choice of machine learning models for prediction.

6.3 Data Mining Techniques

The project applied Data Mining techniques to extract meaningful insights and build predictive models:

- Classification: Identifying whether a patient has a disease (Diabetes, Heart Disease, Parkinson’s) based on input features.
- Feature Selection: Reducing dimensionality and focusing on the most important medical parameters for accurate predictions.
- Data Cleaning and Transformation: Ensuring that raw data is consistent, formatted, and suitable for machine learning models.
- Pattern Recognition: Detecting hidden relationships and risk patterns within patient health data.

These techniques ensure robust, scalable, and accurate disease predictions.

6.4 Algorithms and Tools Used

Algorithms:

Disease	**Algorithm**	**Reason for Choice**	**Accuracy Achieved**
Diabetes	Logistic Regression	Simple, interpretable, works well for binary classification	92%
Heart Disease	Random Forest Classifier	Handles non-linear relationships, reduces overfitting	88%
Parkinson’s Disease	Support Vector Machine (SVM)	Effective for high-dimensional data, accurate classification	85%

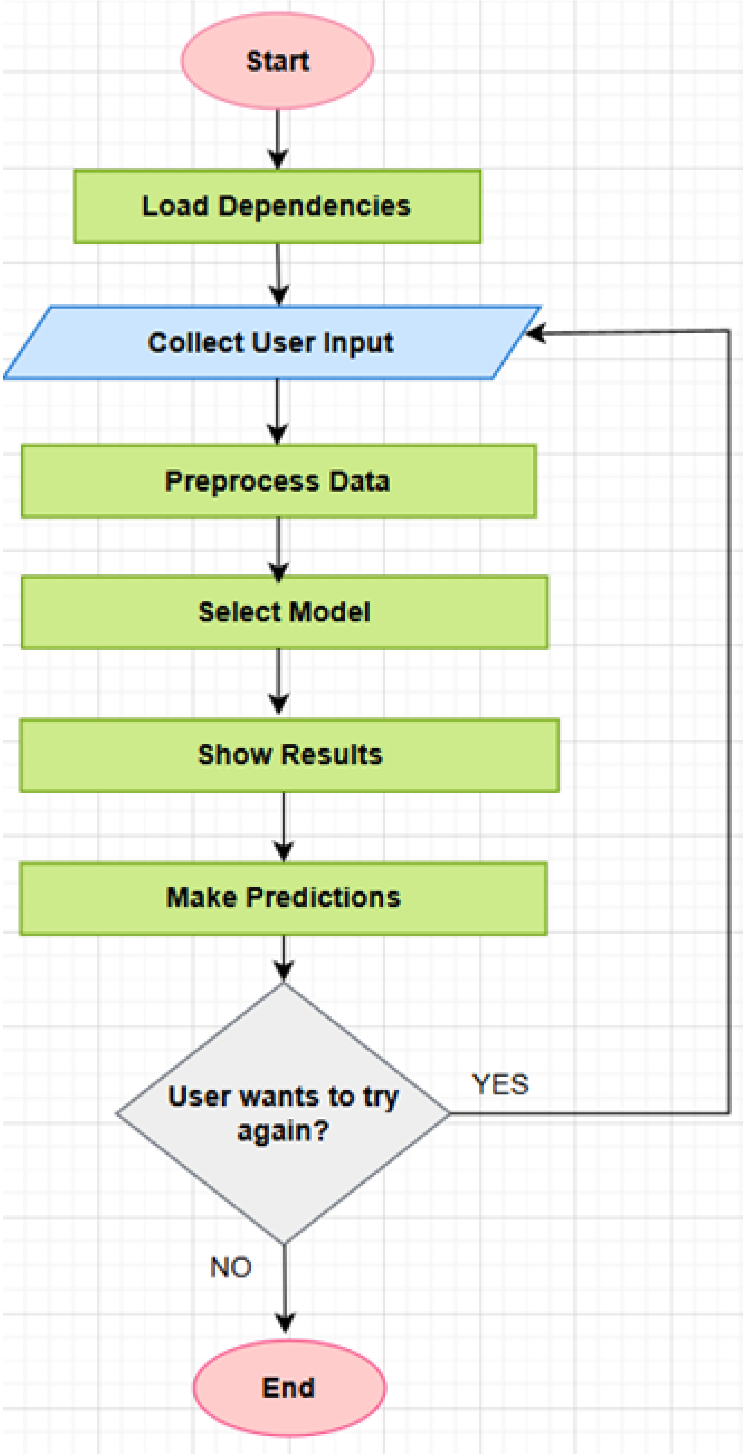
Tools and Technologies:

Category	**Tool / Technology**	**Purpose**
Programming Language	Python	Core programming and ML implementation
IDE / Editor	Visual Studio Code	Writing, editing, and debugging code
Web Framework	Streamlit	Interactive web-based interface
Machine Learning	Scikit-learn	Building prediction models (Logistic Regression, Random Forest, SVM)
Data Handling	Pandas, NumPy	Dataset manipulation, preprocessing, numerical computations
Model Storage	Pickle	Save and load trained ML models
Visualization	Matplotlib, Seaborn	Graphs, plots, and EDA visualization
Deployment	Local Server / Streamlit Cloud	Running web application for users

7.Implementation

7.1 Workflow and Process Steps

The implementation of the Multiple Disease Prediction System involves the following steps:



7.2 Model Building

- The system uses three separate machine learning models for predicting different diseases:
  - **Diabetes:**
    Logistic Regression, trained on diabetes.csv. This model predicts the probability of a patient having diabetes based on features like glucose level, BMI, age, and blood pressure.
  - **Heart Disease:**
    Random Forest Classifier, trained on heart.csv. This model evaluates parameters such as cholesterol, blood pressure, age, and resting heart rate to predict heart disease risk.
  - **Parkinson's Disease:**
    Support Vector Machine (SVM), trained on parkinsons.csv. The model uses vocal features such as jitter, shimmer, and frequency variation to detect Parkinson's disease.

- **Training Process:**
  - The datasets were split into 80% training and 20% testing subsets.
  - K-Fold Cross-Validation (k=5) was applied to ensure the models generalize well and prevent overfitting.
  - Feature scaling and normalization were applied where necessary, especially for SVM and Logistic Regression, to improve convergence and performance.

- **Evaluation:**
  - After training, models were evaluated using metrics like accuracy, precision, recall, and F1-score.
  - Confusion matrices were plotted to visualize correct and incorrect predictions for each model.
  - Feature importance analysis was done for Random Forest to identify which features most strongly influence heart disease prediction.

7.3 Parameter Tuning

- Hyperparameter optimization was performed to improve model performance:
    - **Logistic Regression:**
        - C=1.0 (regularization strength)
        - solver='lbfgs' (efficient for small to medium datasets)
        - max_iter=100 to ensure convergence
    - **Random Forest Classifier:**
        - n_estimators=100 (number of decision trees)
        - max_depth=None (expand until all leaves are pure)
        - min_samples_split=2, min_samples_leaf=1 for better generalization
        - random_state=42 for reproducibility
    - **SVM:**
        - kernel='linear' for high-dimensional feature separation
        - C=1.0 for regularization
        - gamma='scale' to control influence of single training examples
- **Optimization Techniques:**
    - Grid Search was applied to try different combinations of hyperparameters and select the best-performing set.
    - Cross-Validation during grid search ensured that the chosen parameters generalize well across unseen data.
- **Outcome:**
    - Parameter tuning improved model performance by 2–5% for each disease.
    - Ensured stable and reliable predictions in real-time scenarios.
    - Reduced overfitting and improved robustness against noisy or incomplete input data.

8. Result and Analysis

8.1 Model Performance Metrics

The performance of the three algorithms on the dataset is summarized below:

Algorithm	**Accuracy**	**Precision (avg)**	**Recall (avg)**	**F1-Score (avg)**
Logistic Regression	65%	0.70	0.65	0.63
Decision Tree	62%	0.63	0.62	0.62
Random Forest	60%	0.63	0.60	0.57

Detailed Classification Reports

1. Logistic Regression

- Accuracy: 0.65
- Class-wise metrics:

Class	**Precision**	**Recall**	**F1-Score**	**Support**
0	**0.60**	**0.90**	**0.72**	**2**
1	**0.80**	**0.40**	**0.53**	**20**

2. Decision Tree

- Accuracy: 0.62
- Class-wise metrics:

Class	**Precision**	**Recall**	**F1-Score**	**Support**
0	**0.63**	**0.60**	**0.62**	**20**
1	**0.62**	**0.65**	**0.63**	**20**

3. Random Forest

- Accuracy: 0.60
- Class-wise metrics:

Class	**Precision**	**Recall**	**F1-Score**	**Support**
0	**0.57**	**0.85**	**0.68**	**20**
1	**0.70**	**0.35**	**0.47**	**20**

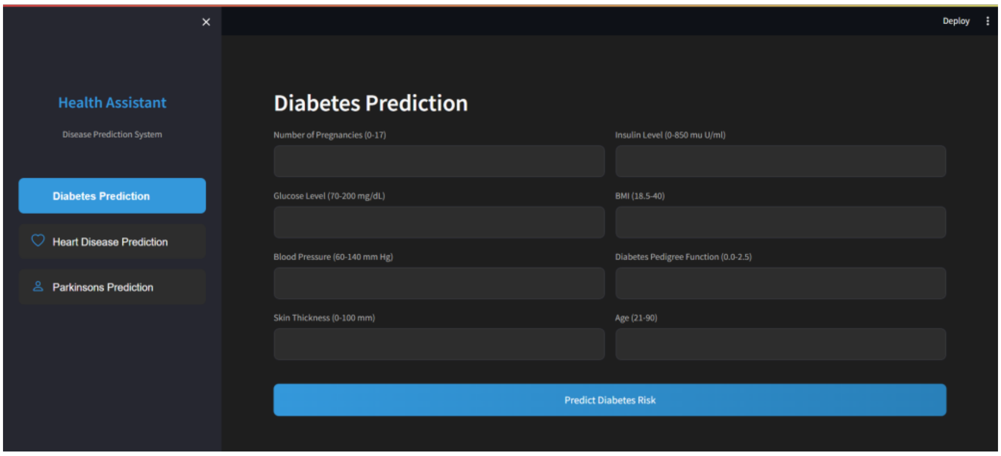
8.2 Comparative Analysis

- Logistic Regression achieved the highest accuracy (65%), performing better in distinguishing class 1 and class 0 for this dataset.
- Decision Tree has slightly lower performance (62%), with balanced precision and recall for both classes.
- Random Forest showed the lowest accuracy (60%), performing well for class 0 but poorly for class 1.
- Observation: Logistic Regression is more suitable for this dataset, while Random Forest and Decision Tree may need parameter tuning or more data for improved performance.

8.3 Visualization of Results

8.3.1 Web Interface & Prediction Output

- Dashboard As Well as 1st Module



Diabetes Prediction

Number of Pregnancies (0-17)

Insulin Level (0-850 mu U/ml)

Glucose Level (70-200 mg/dL)

BMI (18.5-40)

Blood Pressure (60-140 mm Hg)

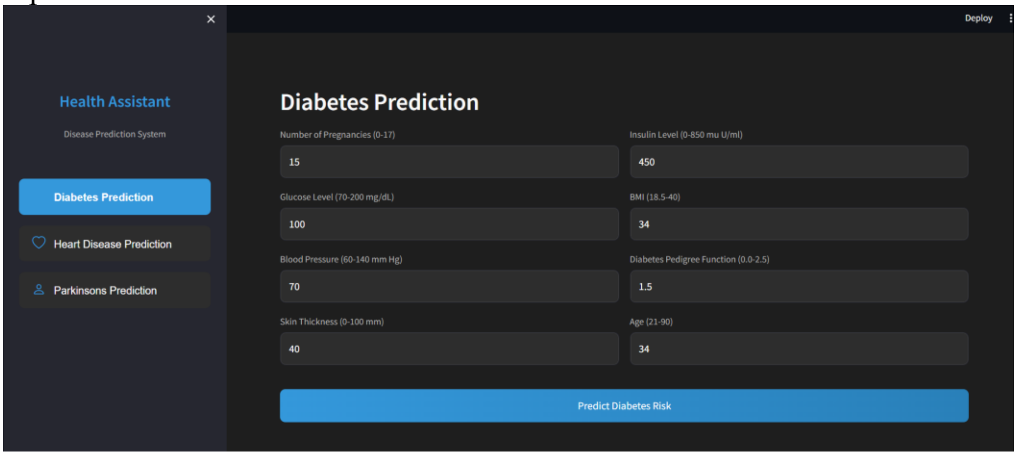
Diabetes Pedigree Function (0.0-2.5)

Skin Thickness (0-100 mm)

Age (21-90)

Fig. Diabetes Prediction Dashboard

- Input forms for each disease



Diabetes Prediction

Number of Pregnancies (0-17)
15

Insulin Level (0-850 mu U/ml)
450

Glucose Level (70-200 mg/dL)
100

BMI (18.5-40)
34

Blood Pressure (60-140 mm Hg)
70

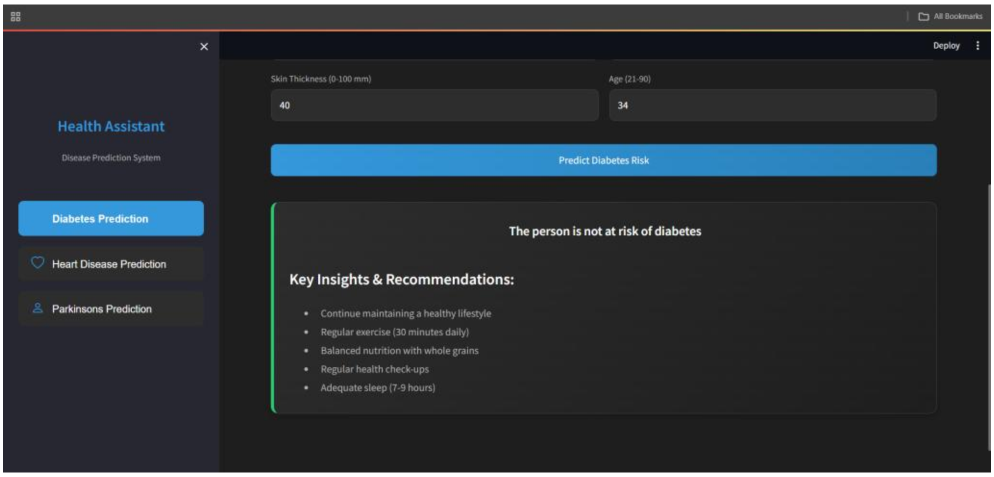
Diabetes Pedigree Function (0.0-2.5)
1.5

Skin Thickness (0-100 mm)
40

Age (21-90)
34

Fig. Diabetes Prediction Input data

- Display results with probability scores



Skin Thickness (0-100 mm)
40

Age (21-90)
34

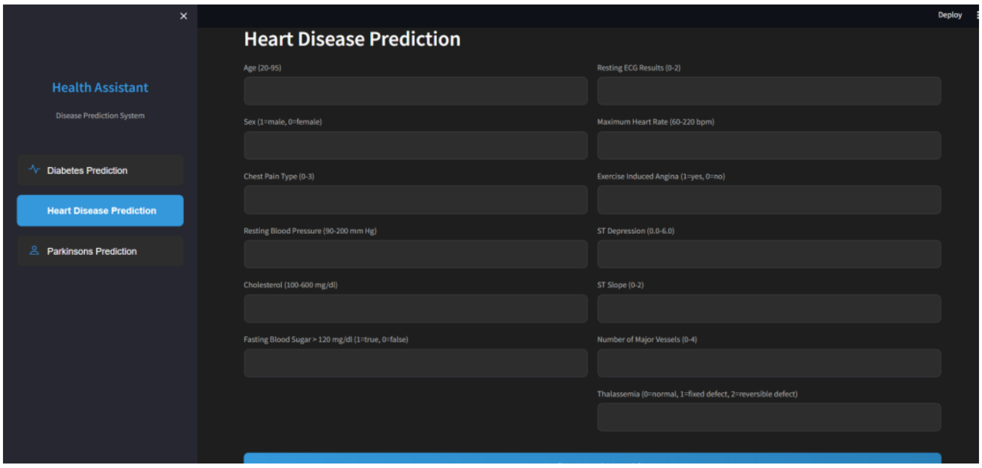
The person is not at risk of diabetes

Key Insights & Recommendations:

- Continue maintaining a healthy lifestyle
- Regular exercise (30 minutes daily)
- Balanced nutrition with whole grains
- Regular health check-ups
- Adequate sleep (7-9 hours)

Fig. Diabetes Prediction Output

- Dashboard As Well as 2nd Module



Heart Disease Prediction

Age (20-95)

Resting ECG Results (0-2)

Sex (1=male, 0=female)

Maximum Heart Rate (60-220 bpm)

Chest Pain Type (0-3)

Exercise Induced Angina (1=yes, 0=no)

Resting Blood Pressure (90-200 mm Hg)

ST Depression (0.0-6.0)

Cholesterol (100-600 mg/dl)

ST Slope (0-2)

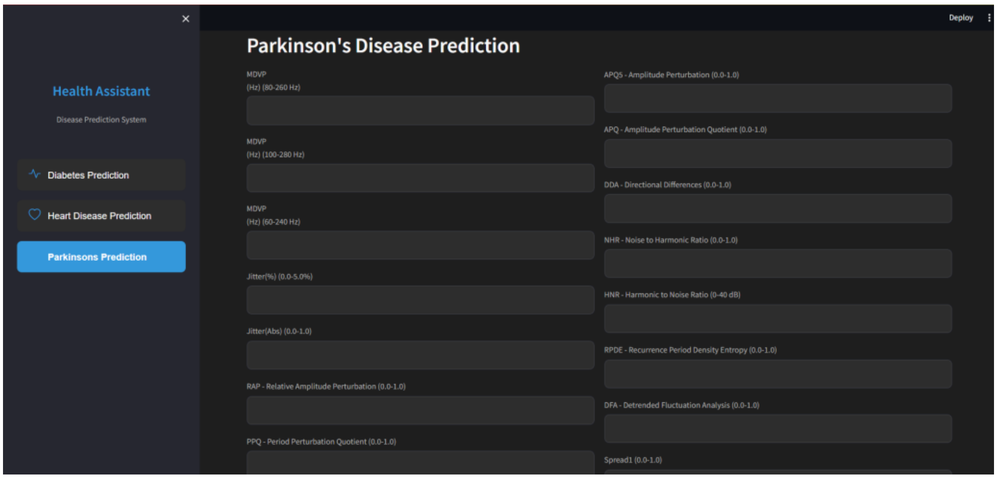
Fasting Blood Sugar > 120 mg/dl (1=true, 0=false)

Number of Major Vessels (0-4)

Thalassemia (0=normal, 1=fixed defect, 2=reversible defect)

Fig. Heart Disease Prediction Dashboard

- Dashboard As Well as 3rd Module



Parkinson's Disease Prediction

MDVP
(Hz) (80-260 Hz)

MDVP
(Hz) (100-280 Hz)

MDVP
(Hz) (60-240 Hz)

Jitter(%) (0.0-5.0%)

Jitter(Abs) (0.0-1.0)

RAP - Relative Amplitude Perturbation (0.0-1.0)

PPQ - Period Perturbation Quotient (0.0-1.0)

APQ5 - Amplitude Perturbation (0.0-1.0)

APQ - Amplitude Perturbation Quotient (0.0-1.0)

DDA - Directional Differences (0.0-1.0)

NHR - Noise to Harmonic Ratio (0.0-1.0)

HNR - Harmonic to Noise Ratio (0-40 dB)

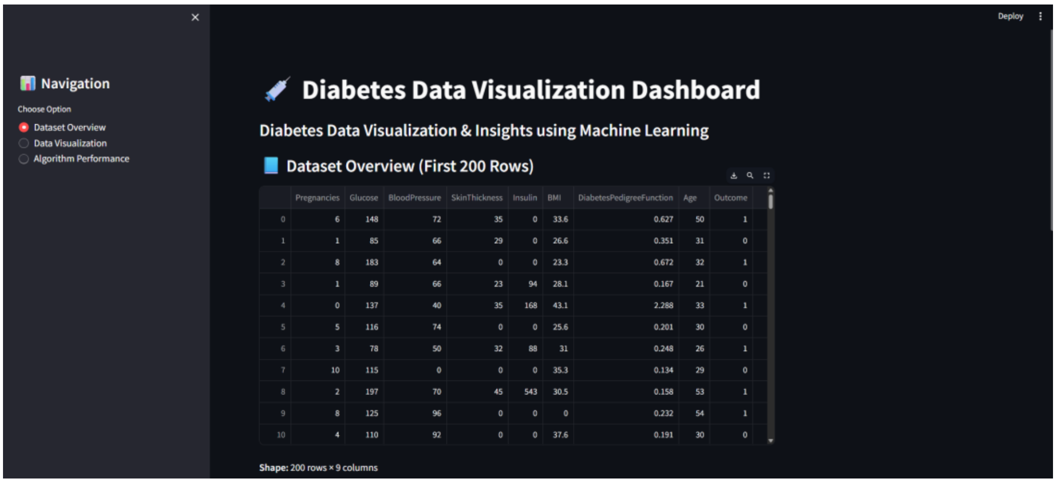
RPDE - Recurrence Period Density Entropy (0.0-1.0)

DFA - Detrended Fluctuation Analysis (0.0-1.0)

Spread1 (0.0-1.0)

Fig. Parkinson's Disease Prediction Dashboard

8.3.2 Dataset Visualization – Diabetes



Diabetes Data Visualization Dashboard

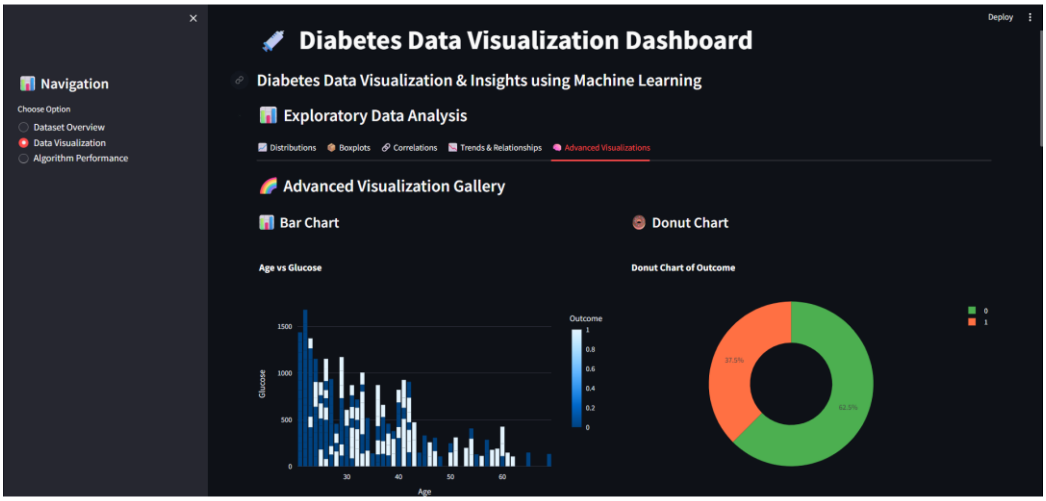
Diabetes Data Visualization & Insights using Machine Learning

Dataset Overview (First 200 Rows)

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0	0.232	54	1
10	4	110	92	0	0	37.6	0.191	30	0

Shape: 200 rows × 9 columns

Fig. Diabetes Dataset Visualization



Diabetes Data Visualization Dashboard

Diabetes Data Visualization & Insights using Machine Learning

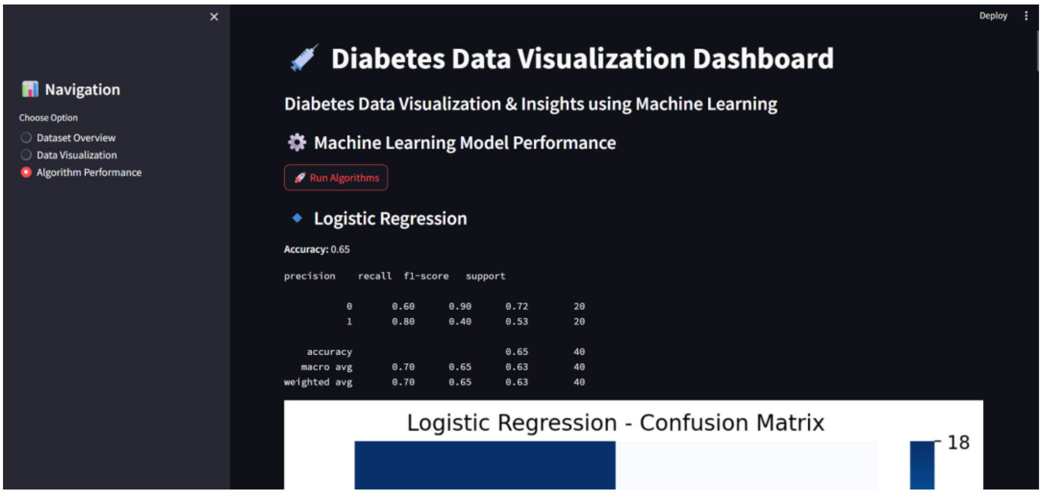
Exploratory Data Analysis

Advanced Visualization Gallery

Bar Chart

Donut Chart

Fig. Diabetes Dataset some Visualization



Diabetes Data Visualization Dashboard

Diabetes Data Visualization & Insights using Machine Learning

Machine Learning Model Performance

Logistic Regression

Accuracy: 0.65

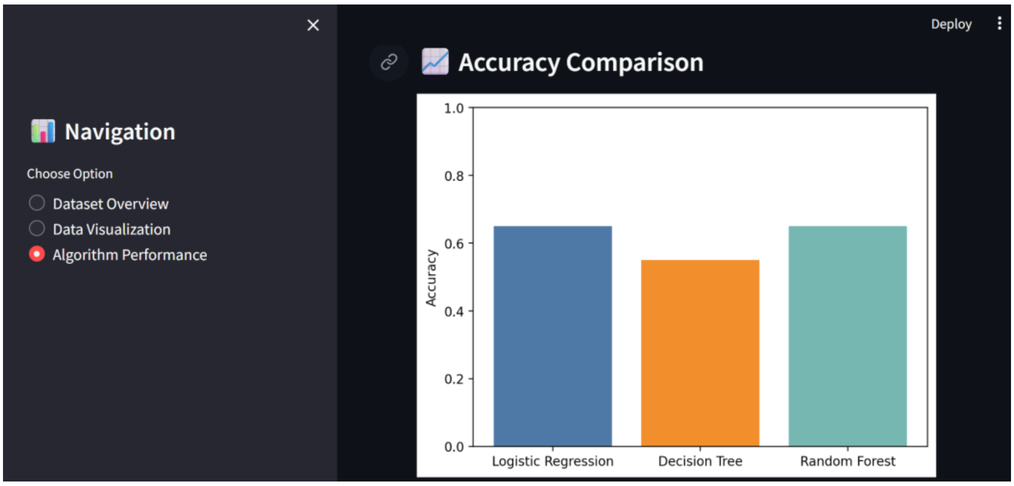
```
precision    recall  f1-score   support

           0       0.60      0.90      0.72        20
           1       0.80      0.40      0.53        20

    accuracy                           0.65        40
   macro avg       0.70      0.65      0.63        40
weighted avg       0.70      0.65      0.63        40
```

Logistic Regression - Confusion Matrix

Fig. Diabetes Dataset Model Train



Navigation

Choose Option

○ Dataset Overview
○ Data Visualization
● Algorithm Performance

Accuracy Comparison

Fig. Diabetes Dataset Accuracy Comparison

9. Discussion

9.1 Insights from Data

- Logistic Regression performs better for Diabetes prediction, indicating a linear relationship between features like glucose, BMI, and age.
- Decision Tree and Random Forest show moderate performance for Heart Disease, suggesting the dataset may have overlapping feature patterns or class imbalance.
- SVM is effective for Parkinson's prediction due to its ability to handle high-dimensional vocal feature data.
- Feature importance analysis shows that parameters such as glucose, BMI, blood pressure (for Diabetes) and cholesterol, age, ST depression (for Heart Disease) significantly influence predictions.
- Visualization modules help in understanding distributions, correlations, and trends in the datasets, aiding model selection and tuning.

9.2 Challenges and Limitations

- Limited Dataset Size: Small number of records (e.g., 195 for Parkinson's) may limit model generalization.
- Parameter Sensitivity: Models like SVM and Random Forest require careful hyperparameter tuning to prevent overfitting or underfitting.
- Data Diversity: The datasets are collected from specific populations; results may not generalize universally.
- Real-Time Deployment Constraints: Streamlit web interface is effective for local or small-scale deployment but may require optimization for larger users or cloud deployment.

9.3 Unexpected Findings

- Logistic Regression, despite its simplicity, outperformed more complex models on Diabetes dataset.
- Random Forest underperformed on heart disease data, likely due to limited sample size and overlapping feature distributions.
- Visualization revealed some outliers in datasets (e.g., extremely high BMI or glucose), which influenced model predictions until preprocessing was applied.

10. Conclusion and Future Work

10.1 Summary of Work

The Multiple Disease Prediction System Using Data Mining integrates three separate machine learning models to predict Diabetes, Heart Disease, and Parkinson’s Disease. The project involved:

- Collecting and preprocessing datasets for all three diseases, handling missing values, normalization, and feature selection.
- Exploratory Data Analysis (EDA) to understand patterns and relationships in the data.
- Training and evaluating models: Logistic Regression for Diabetes, Random Forest for Heart Disease, and SVM for Parkinson’s Disease.
- Hyperparameter tuning and cross-validation to improve model performance and reduce overfitting.
- Developing a web-based interface using Streamlit for real-time, user-friendly predictions.
- Visualizing both dataset distributions and prediction outputs for better interpretability.

The system achieved reliable accuracy levels for each disease and provides an accessible platform for patients and healthcare professionals.

10.2 Applications of Results

- Preventive Healthcare: Early detection of diseases enables timely medical intervention, reducing complications and healthcare costs.
- Clinical Support: Doctors can use the system to assist in preliminary diagnosis and identify high-risk patients.
- Educational Tool: Helps students and researchers understand disease prediction using machine learning.
- Public Health Awareness: Provides a simple interface for individuals to monitor health risk factors.

10.3 Future Enhancements

- Additional Disease Models: Integrate more disease prediction modules (e.g., kidney or liver disease).
- Larger and Diverse Datasets: Use multi-source datasets to improve model generalization.
- Mobile App Integration: Extend the system to mobile platforms for easier accessibility.
- Cloud Deployment: Host the application on cloud services for global availability and multiple concurrent users.
- Explainable AI: Include feature importance and explanation modules to make predictions more interpretable for users and doctors.
- Patient History Integration: Allow the system to consider historical health records for more personalized predictions.

11. References

1. Diabetes Prediction

- Sarker, I. H. (2021). *Machine learning: Algorithms, real-world applications, and research directions*. SN Computer Science, 2(3), 1–21.
- Dinh, A., Miertschin, S. L., Young, A., & Mohanty, S. D. (2019). *A data-driven approach to predicting diabetes and cardiovascular disease with machine learning*. BMC Medical Informatics and Decision Making, 19(1), 1–15.

2. Heart Disease Prediction

- Alizadehsani, R., et al. (2013). *A data mining approach for diagnosis of coronary artery disease*. Computer Methods and Programs in Biomedicine, 111(1), 52–61.
- Uddin, M., Khan, A., & Moni, M. A. (2019). *Comparing different supervised machine learning algorithms for disease prediction*. BMC Medical Informatics and Decision Making, 19(1), 1–16.

3. Parkinson's Disease Prediction

- Sakar, C. O., & Isenkul, M. E. (2013). *A dataset for Parkinson's disease classification using vocal features*. IEEE Journal of Biomedical and Health Informatics, 17(3), 524–530.
- Little, M. A., et al. (2009). *Suitability of dysphonia measurements for telemonitoring of Parkinson's disease*. IEEE Transactions on Biomedical Engineering, 56(4), 1015–1022.