Disease Prediction System using Machine Learning for Early Diagnosis

Phase 3: Final Report and Submission

College Name: VIT Bhopal University

Student Name: Taruna Gupta

---

1. Project Title:

Disease Prediction System using Machine Learning for Early Diagnosis

---

2. Summary of Work Done

Phase 1 – Proposal and Idea Submission:

In the initial phase, we proposed building a Disease Prediction System using supervised machine learning techniques. The project was motivated by the rising incidence of chronic diseases such as **diabetes** and **heart disease**, which place a burden on healthcare systems and often go undetected in early stages.

Our objective was to create a digital solution that:

- Accepts patient health data as input.
- Predicts the likelihood of a disease using trained ML models.
- Provides recommendations for further action.

The proposed architecture included:

- Collecting data from reliable medical datasets.
- Preprocessing and cleaning the data.
- Training predictive models using algorithms like Random Forest and Logistic Regression.
- Creating a web application using **Streamlit** for user interaction.

The idea was well-documented with references to datasets, tools, and anticipated benefits such as low-cost preliminary diagnosis, accessibility, and user empowerment.

Phase 2 – Execution and Demonstration:

In the second phase, the concept was translated into a working prototype. Key tasks completed included:

- **Data Acquisition**: Used publicly available datasets like the Pima Indian Diabetes Dataset from the UCI ML repository.
- **Data Preprocessing**: Cleaned data, handled missing values, and selected relevant features like age, BMI, blood pressure, and glucose level.
- **Model Building**: Implemented a **Random Forest Classifier** using scikit-learn. The model was trained, tested, and evaluated using accuracy and classification metrics.
- **Web Interface**: Developed a **Streamlit-based GUI** that allows users to input medical parameters through sliders and buttons. Upon clicking "Predict," the system displays whether the user is at risk of disease or not.
- **Performance Evaluation**: The model achieved a satisfactory level of accuracy with a strong classification report. The interface was intuitive and responsive.

Screenshots, source code, and prediction examples were included in the Phase 2 documentation.

---

3. GitHub Repository Link

🔗 GitHub Repository – Disease Prediction System using Machine Learning for early diagnosis
https://github.com/TarunaGupta22/DiseasePredictionApp.git

---

4. Testing Phase

4.1 Testing Strategy

The Disease Prediction System was evaluated using a comprehensive testing approach that included:

- Verifying **input handling** for different user-provided medical data combinations.
- Checking **model reliability** and **classification accuracy** on test datasets.
- Assessing the **user interface experience** for responsiveness and usability.
- Evaluating system behaviour with **extreme or nonsensical inputs**.

4.2 Types of Testing Conducted

- **Unit Testing**: Validated core components like data preprocessing, model training, and prediction generation.
- **Integration Testing**: Ensured proper coordination between the ML model and Streamlit interface.

- **User Testing**: Test users simulated real interactions, providing feedback on usability and prediction clarity.
- **Edge Case Testing**: Tested extreme inputs (e.g., very high glucose or blood pressure) to ensure the model did not crash and provided plausible outputs.
- **Performance Testing**: Measured prediction speed and resource consumption under different loads.

4.3 Results

- **Model Accuracy**: Achieved 80–85% accuracy depending on input features. The model was consistent in predicting risk for typical patient profiles.
- **Classification Metrics**: Precision and recall were balanced, making the model suitable for early screening.
- **UI Responsiveness**: Streamlit interface processed inputs and delivered results instantly, enhancing user experience.
- **Robustness**: Model handled all inputs without breaking, though predictions for extremely noisy data may be less reliable (a noted limitation).

---

5. Future Work

While the current version is a strong proof-of-concept, several enhancements are planned to increase real-world applicability:

- **Advanced Algorithms**: Incorporating ensemble learning or deep neural networks to improve predictive accuracy.
- **Multi-Disease Capability**: Expanding beyond diabetes and heart disease to include predictions for hypertension, stroke risk, etc.
- **More Features**: Adding lifestyle data (e.g., smoking status, physical activity) and clinical indicators (e.g., cholesterol, HbA1c levels).
- **Real-Time Data Integration**: Syncing with smart health devices or mobile apps to analyse live health metrics.
- **Feedback Loop Integration**: Allowing users or physicians to verify predictions and feed corrections back to the model for retraining.
- **Cloud Hosting & API Access**: Deploying the app on platforms like Heroku or AWS and exposing it as a REST API.
- **Multi-Language Support**: Making the interface accessible in regional languages to improve reach in rural and non-English-speaking areas.

---

6. Conclusion

The **Disease Prediction System using Machine Learning** effectively demonstrates how AI and ML can augment healthcare delivery. By combining predictive analytics with a user-friendly interface, this project enables early risk detection, promotes awareness, and assists healthcare professionals in prioritizing care.

Through structured project phases, the system evolved from an idea to a deployable prototype, proving the potential of machine learning in preventive health applications. With future enhancements, this system could contribute meaningfully to telemedicine, remote health monitoring, and community health outreach programs.