**Capstone Project Proposal: Patient Disease Prediction Using Machine Learning**

**Duration: 15 Days**

**1. Introduction**

With the growing importance of early disease detection in healthcare, machine learning models offer promising results in predicting chronic illnesses. This project aims to develop a machine learning-based system to predict the likelihood of a patient having **breast cancer**, **diabetes**, or **heart disease** based on clinical data. The project will integrate models trained on three real-world datasets and deploy them using **Streamlit** for user-friendly interaction.

**2. Objectives**

* Preprocess and clean three datasets: **Breast Cancer**, **Diabetes**, and **Heart Disease**
* Train and evaluate multiple ML models (e.g., Logistic Regression, Random Forest, SVM)
* Save the best-performing model for each disease as a .pkl file
* Build a unified **Streamlit application** that:
  + Accepts user input
  + Loads appropriate models
  + Returns disease predictions interactively
* Present model performance metrics and analysis
* Create a structured and documented GitHub repository

**3. Project Timeline**

| **Days** | **Tasks** |
| --- | --- |
| **Day 1–3** | Data Preparation and Exploration: Understand datasets, clean/preprocess data (handle nulls, encode categories), and perform EDA |
| **Day 4–7** | Model Training and Evaluation: Train models (Logistic Regression, Decision Tree, Random Forest, SVM). Evaluate with accuracy, precision, recall, F1-score. Save the best .pkl for each disease |
| **Day 8–11** | Streamlit App Development: Design unified interface for user inputs, prediction outputs, and model loading logic |
| **Day 12–13** | Testing and Refinement: Perform usability testing, validate predictions, and refine UI/UX |
| **Day 14–15** | Documentation and Deployment: Create README.md, structure project folders, and prepare for GitHub deployment |

**4. Expected Outcome**

By the end of this project, a fully functional **machine learning-based disease prediction app** will be developed. Users will be able to input medical parameters and receive predictions for **breast cancer**, **heart disease**, or **diabetes**. This tool can serve as a foundational aid in healthcare screening and risk assessment systems.

**5. Tools and Libraries Required**

* Python
* Pandas, NumPy, Scikit-learn
* Streamlit
* Matplotlib, Seaborn
* Pickle
* GitHub for version control

**6. Final Deliverables**

* .pkl files of the best model for each disease
* Streamlit app (Disease\_Prediction\_app.py)
* Cleaned and processed datasets
* Jupyter notebooks for model building
* Full GitHub repository with:

bash

CopyEdit

/models → Saved models (.pkl)

/data → Input datasets

/app → Streamlit app files

/notebooks → ML model development

/docs → Project documentation

README.md → Project overview and setup