ROADMAP

AI FOR HEALTH: PREDICTING DISEASES WITH ML

**1. Project Planning**

* **Define Objectives**:
  + Accurately predict the likelihood of diseases based on patient data.
  + Identify key health indicators using feature selection.
  + Group patients into meaningful clusters for targeted interventions.
* **Choose a Dataset**: Use publicly available healthcare datasets such as:
  + UCI Machine Learning Repository: Heart Disease Dataset.
  + Kaggle: Diabetes or chronic kidney disease datasets.

**2. Data Collection and Preparation**

* **Collect Data**: Gather patient data containing demographic, lifestyle, medical history, and health indicators.
* **Data Pre-processing**:
  + Handle missing values (mean/mode imputation or deletion).
  + Encode categorical features (one-hot encoding, label encoding).
  + Normalize/standardize numerical features for clustering.
* **Split Dataset**:
  + Training Set: 70%.
  + Testing Set: 30%.

**3. Exploratory Data Analysis (EDA)**

* **Visualizations**:
  + Analyse distributions of features using histograms or boxplots.
  + Use scatterplots and correlation matrices to find relationships between variables.
* **Key Insights**:
  + Identify features most correlated with disease likelihood.
  + Detect and address outliers.

**4. Model Development**

**Naive Bayes:**

* **Goal**: Classify patients into disease-positive or disease-negative categories.
* **Implementation**:
  + Train the model using health indicators as features and disease status as labels.
  + Evaluate using metrics like accuracy, precision, recall, and F1-score.

**Random Forest:**

* **Goal**: Perform feature selection and refine classification accuracy.
* **Implementation**:
  + Train the model on the same dataset.
  + Use feature importance scores to identify the most significant health indicators.
  + Tune hyperparameters (number of trees, max depth) for optimal performance.

**K-Means Clustering:**

* **Goal**: Group patients based on health profiles to identify patterns (e.g., high-risk vs. low-risk groups).
* **Implementation**:
  + Apply clustering on normalized health indicator data.
  + Use the **silhouette score** to evaluate clustering quality.
  + Visualize clusters using PCA or t-SNE for dimensionality reduction.

**5. Model Evaluation**

* **Naive Bayes & Random Forest**:
  + Use a confusion matrix and ROC curve to assess performance.
* **K-Means**:
  + Analyse cluster centroids and their relationships with disease risk factors.
* Compare classification metrics to determine the best algorithm.

**6. Insights and Interpretations**

* **Feature Importance**:
  + Identify top health indicators contributing to disease prediction.
* **Cluster Insights**:
  + Understand patterns in patient profiles and their implications for disease management.
* **Recommendations**:
  + Propose targeted interventions based on identified clusters.

**7. Deployment**

* Develop a simple web or mobile application to accept patient data and return:
  + Predicted likelihood of disease.
  + Significant health indicators.
  + Patient cluster information (e.g., risk group).

**8. Documentation and Reporting**

* Prepare a comprehensive report including:
  + Objectives and problem statement.
  + Dataset description.
  + EDA findings and visualizations.
  + Model performances and insights.
  + Recommendations for healthcare practitioners.

**Tools and Libraries**

* **Languages**: Python
* **Libraries**:
  + pandas, NumPy for data processing.
  + matplotlib, seaborn for visualization.
  + scikit-learn for ML algorithms.
  + flask or streamlit for deployment.

Would you like assistance with implementation steps or help finding a dataset?