**Project Charter Document**



**Project Name :** MULTIPLE DISEASE PREDICTION

**Department : HEALTHCARE**

**Focus Area :** DIEASESE Analytics

**Product/Process :** DISEASE ANALYSIS



**Prepared By**

|  |  |
| --- | --- |
| **Document Owner(s)** | **Project/Organization Role** |
| AKASH SHERAJI SATHAWANE | DATA SCIENCE INTERN |

**Project Charter Version Control**

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Author** | **Change Description** |
| 1.0 | JUNE 2024 |  | Document created JUNE 2024 |

PROJECT CHARTER PURPOSE

The purpose of this project is to enhance the early detection of multiple diseases by using machine learning models. By evaluating the recall performance of different models, the project aims to identify the best-performing algorithms that can minimize false negatives, ensuring that patients receive timely diagnosis and treatment.



# PROJECT EXECUTIVE SUMMARY

The **Multiple Disease Prediction** project aims to improve the accuracy of machine learning models used for predicting various diseases, with a particular emphasis on recall performance. In healthcare, detecting diseases early is critical to providing effective treatment and improving patient outcomes. This project focuses on evaluating the recall scores of different predictive models, prioritizing those that excel at identifying positive cases and minimizing false negatives.

To achieve this, several machine learning models were assessed, and their recall performance was visualized through a horizontal bar chart. The chart compares the recall scores of models used for predicting multiple diseases, offering clear insights into their ability to correctly classify patients with a specific disease. Models with higher recall values were identified as more reliable for disease detection, ensuring that fewer cases are missed.

The primary deliverable of this project is a recall score visualization that simplifies model evaluation for healthcare professionals and data scientists. By presenting an easy-to-interpret comparison of model performance, this project enables informed decision-making, ensuring the selection of models that are best suited for clinical use. The outcome supports more accurate disease diagnosis, leading to earlier interventions and better healthcare outcomes.



# PROJECT OVERVIEW

The **Multiple Disease Prediction** project focuses on the development and evaluation of machine learning models for the early detection of multiple diseases. In medical diagnostics, accurate prediction is crucial to ensuring patients receive timely treatment. The project emphasizes the use of the **recall score** as the key performance metric, which measures a model’s ability to correctly identify patients with a disease, minimizing the occurrence of false negatives.

By analyzing recall scores, this project aims to determine the most effective models in detecting diseases, particularly in scenarios where false negatives could result in critical health consequences. A high recall score is particularly important for diseases that require urgent intervention, as missed diagnoses can lead to delayed treatments and worsening health outcomes.



# PROJECT SCOPE

The project involves:

1. **Model Evaluation**: Various machine learning models are assessed for their performance in predicting multiple diseases.
2. **Recall Visualization**: Recall scores are visualized using a horizontal bar chart to provide a clear and concise comparison of the models.
3. **Insight Generation**: The results allow for the identification of models that are most suitable for disease prediction tasks, based on their ability to accurately detect positive cases.

Through the visualization of recall scores, healthcare professionals and data scientists can easily interpret which models are best for minimizing false negatives, leading to more accurate diagnoses and better patient outcomes.

## Goals and Objectives

|  |  |
| --- | --- |
| **Goals** | **Objectives** |
| * Clean | * Data |

## Project Deliverables

The **Multiple Disease Prediction** project includes the following key deliverables:

1. **Model Performance Evaluation**:
   * A comprehensive assessment of several machine learning models used for predicting multiple diseases.
   * Evaluation based on the **recall score** for each model, focusing on the model’s ability to correctly identify positive cases and minimize false negatives.
2. **Recall Score Visualization**:
   * A **horizontal bar chart** that visually compares the recall scores of the models. Each bar represents a model's recall score, with annotations displaying the exact recall values.
   * The chart provides a clear comparison of model performance, allowing for quick identification of the best-performing models.
3. **Data Insights and Analysis**:
   * Detailed insights derived from the recall score visualization, including analysis of which models are most effective at disease prediction.
   * Highlights of the top-performing models based on their ability to correctly diagnose patients with a disease, reducing the likelihood of missed diagnoses.
4. **Recommendation Report**:
   * A summary of model performance, with recommendations on the most suitable models for disease prediction based on recall scores.
   * Suggestions for healthcare data scientists and practitioners on model selection, emphasizing models that prioritize high recall to ensure accurate diagnosis.
5. **Python Code for Data Analysis and Visualization**:
   * A well-documented Python script used for sorting the recall scores, generating the bar chart, and annotating the results.
   * Code is provided to ensure that the process is reproducible and can be used for future model evaluations or in other disease prediction tasks.



# PROJECT CONDITIONS

## Project Assumptions

The **Multiple Disease Prediction** project is based on the following key assumptions:

1. **Data Quality and Availability**:
   * It is assumed that the data used for training and evaluating the machine learning models is accurate, reliable, and representative of real-world disease cases. The models are expected to perform well based on the quality of the dataset provided.
   * It is also assumed that the dataset includes a sufficient number of examples for each disease, ensuring balanced and meaningful predictions.
2. **Class Imbalance**:
   * It is assumed that some diseases may have imbalanced data (i.e., fewer positive cases than negative cases), which makes recall a crucial metric. This project assumes that the recall score is an appropriate measure for evaluating model performance in this context.
3. **Model Interpretability**:
   * The project assumes that healthcare practitioners and data scientists are familiar with machine learning concepts, including recall as a performance metric. This ensures that the deliverables, such as recall score visualization, are understandable and actionable by the target audience.
4. **No External Factors Affecting Model Accuracy**:
   * It is assumed that the models are trained in a controlled environment without external biases, such as changes in medical practices, diagnostic technologies, or variations in disease definitions that could affect model predictions.
5. **Predictive Models are Generalizable**:
   * The project assumes that the machine learning models trained on the provided data are generalizable, meaning they can predict multiple diseases with similar performance when deployed in different healthcare settings or applied to new patient data.
6. **Technology Compatibility**:
   * It is assumed that the models and visualization tools developed in this project are compatible with the technology infrastructure of the healthcare institutions where they will be deployed, including necessary software and hardware for implementation.
7. **Timely Implementation of Recommendations**:
   * It is assumed that recommendations based on model evaluation will be implemented in a timely manner by healthcare organizations to improve disease prediction accuracy, ultimately benefiting patient care

### ****Project Overview****:

The project aims to analyze and visualize the recall performance of various models with respect to the positive class. The primary deliverable is a clear and intuitive visualization that showcases the recall values for different models to evaluate their performance in distinguishing the positive class.

### ****Objectives****:

1. **Recall Score Visualization**: Create a horizontal bar chart that illustrates the recall values for different models, with clear annotations for each.
2. **Performance Analysis**: Provide a visual tool to compare and interpret the recall values across multiple models, aiding in model selection and optimization.

### ****Scope****:

1. **Data Preparation**: Data is already available in the notebook with a specific column representing recall scores. These will be sorted and used for plotting.
2. **Visualization**:
   * The visualization will consist of a horizontal bar chart.
   * Each bar will represent the recall score of a particular model.
   * The bars will be color-coded and annotated with recall values.
   * The chart will include a descriptive title, axis labels, and will not display y-ticks to improve readability.
3. **Model Evaluation**: Focus on evaluating models based solely on their recall values for the positive class, which is crucial for imbalanced classification tasks.

### ****Deliverables****:

1. **Recall Score Chart**:
   * A horizontal bar chart visualizing recall scores for the positive class.
   * Annotated recall values for each model, with clear color-coded bars.
   * Descriptive title and axis labels for better interpretation.
2. **Analysis Report**:
   * An analysis of the results explaining the recall scores and their significance.
   * Recommendations on the models that perform best in terms of recall.

### ****Roles and Responsibilities****:

* **Data Scientist** : Responsible for executing the notebook, generating recall score plots, and analyzing the results.
* **Project Manager**: Oversees the overall progress, ensures that deliverables are produced on time, and communicates with stakeholders.

### ****Success Criteria****:

* Successful generation of a recall performance chart that provides meaningful insights into model selection.
* Clear visualization that allows for quick and easy comparison of recall scores across models.

### ****3.3. Tools and Technologies****

* **Python**: The project used Python as the primary programming language.
* **Libraries**: Key libraries used include:
  + **Pandas**: For data manipulation.
  + **Matplotlib**: For plotting the horizontal bar chart.

## ****4. Results****

### ****4.1. Recall Score Chart****

The output of the project is a horizontal bar chart visualizing the recall scores for the positive class. Each bar corresponds to a specific model, with the following features:

* **Annotation**: The recall values are directly labeled on each bar, providing a clear view of model performance.
* **Visualization**: The chart is visually appealing, with clear and bold fonts used for both the recall scores and model names.

The chart provides a quick and intuitive way to compare the recall scores across models. By sorting the models based on their recall scores, the visualization highlights which models perform best in terms of minimizing false negatives.

### ****4.2. Performance Analysis****

The key insight from the visualization is the performance difference among the models based on recall:

* **High Recall Models**: Models with higher recall values are better at correctly identifying positive instances, making them suitable for tasks where false negatives are costly.
* **Low Recall Models**: Models with lower recall values might not be ideal for critical applications where missing positive instances is a major concern.

## ****5. Conclusion and Recommendations****

### ****5.1. Conclusion****

The recall score chart provides a clear and straightforward method to evaluate and compare the performance of different models based on their ability to correctly classify positive instances. By focusing on recall, this visualization is particularly useful in scenarios where the costs of false negatives are high.

### ****5.2. Recommendations****

* **Model Selection**: Based on the recall scores, it is advisable to prioritize models with the highest recall values when the goal is to minimize false negatives.
* **Further Analysis**: It is recommended to complement the recall analysis with other metrics, such as precision, F1 score, and overall accuracy, to gain a comprehensive understanding of model performance.

## ****6. Future Work****

Future work can involve:

* **Incorporating Precision and F1 Scores**: Combining recall with precision to generate F1 scores would give a more balanced view of model performance.
* **Extending to Other Classes**: The current project focuses on the positive class recall. Expanding the analysis to other classes can provide a more holistic evaluation.
* **Hyperparameter Tuning**: Further fine-tuning of the models could improve recall performance.

Project successfully achieved its goal of visually representing the recall scores for various models, providing useful insights for model evaluation and selection.