Early Prediction of Chronic Kidney Disease Using Machine Learning

Milestone 1: Project Initialization and Planning Phase

This project aims to develop a machine learning model for early prediction of chronic kidney disease using adult patient data from electronic health records. By analysing past medical history, lab results, and demographic information, the model will provide real-time risk assessments, enabling early intervention. Key features include high accuracy, EHR integration, and interpretable results for clinicians. The goal is to improve patient outcomes and reduce healthcare costs associated with late-stage CKD diagnosis and treatment.

Activity 1: Define Problem Statement

Problem Statement:A health insurance company's data analyst needs to improve early prediction of chronic kidney disease (CKD) among policyholders. Despite having access to extensive policyholder data, current risk assessment models lack sensitivity for early-stage CKD detection. This limitation hinders the company's ability to promote timely preventive care, potentially resulting in increased healthcare costs and poorer health outcomes for policyholders. The challenge lies in developing a more effective data utilization strategy, leveraging machine learning to identify subtle patterns indicative of early CKD risk.

Early Prediction of Chronic Kidney Disease Using Machine Learning ProblemStatement Report: Click Here

Activity 2: Project Proposal (Proposed Solution)

This project aims to develop a machine learning model for early prediction of chronic kidney disease using adult patient data from electronic health records. By analysing past medical history, lab results, and demographic information, the model will provide real-time risk assessments, enabling early intervention. Key features include high accuracy, EHR integration, and interpretable results for clinicians. The goal is to improve patient outcomes and reduce healthcare costs associated with late-stage CKD diagnosis and treatment.

Early Prediction of Chronic Kidney Disease Using Machine Learning ProblemStatement Report: Click Here

Activity 3: Initial Project Planning

The planning involves collecting anonymized patient data from EHRs, preprocessing it and selecting relevant features for CKD prediction. Multiple machine learning algorithms will be implemented and compared to develop the best performing model, which will then be validated using cross-validation techniques. A user-friendly interface will be created for clinical integration, allowing healthcare providers to input data and receive risk assessments. Finally, the model will undergo pilot testing in a clinical setting to gather professional feedback and refine its performance.

Early Prediction of Chronic Kidney Disease Using Machine Learning ProblemStatement Report: Click Here

Milestone 2: Data Collection and Preprocessing Phase

The Data Collection and Preprocessing Phase involves executing a plan to gather relevant medical data from Kaggle, ensuring data quality through verification and addressing missing values. Preprocessing tasks include cleaning, encoding, and organizing the dataset for subsequent exploratory analysis and machine learning model development.

Activity 1: Data Collection Plan, Raw Data Sources Identified, Data Quality Report

The dataset for "Early Prediction of Chronic Kidney Disease Using Machine Learning" is sourced from Skill wallet(Kaggle). It incorporates applicant details and medical data integrity metrics. Data quality is assured through meticulous verification, addressing missing values, and upholding ethical standards, establishing a dependable foundation for predictive modeling in healthcare.

Early Prediction of Chronic Kidney Disease Using Machine Learning ProblemStatement Report: Click Here

Activity 2: Data Quality Report

The dataset used is sourced from Skill wallet, containing applicant details and metrics to ensure the accuracy of medical data. Rigorous verification processes ensure data quality by addressing missing values and upholding ethical standards. This careful approach establishes a reliable foundation for predictive modelling in healthcare.

Early Prediction of Chronic Kidney Disease Using Machine Learning ProblemStatement Report: Click

Here

Activity 3: Data Exploration and Preprocessing

Data Exploration involves analysing the dataset to understand patterns, distributions, and outliers. Preprocessing includes handling missing values, scaling, and encoding categorical variables. These crucial steps enhance data quality, ensuring the reliability and effectiveness of subsequent analyses in the kidney disease prediction project.

Early Prediction of Chronic Kidney Disease Using Machine Learning ProblemStatement Report: Click Here

Milestone 3: Model Development Phase

The Model Development Phase entails crafting a predictive model for disease prediction. It encompasses strategic feature selection, evaluating and selecting models (Random Forest, Decision Tree, KNN, Logistic), initiating training with code, and rigorously validating and assessing model performance for informed decision-making in the lending process.

Activity 1: Feature Selection Report

The Feature Selection Report highlights five crucial predictors for the chronic kidney disease (CKD) model: Age, Albumin, Sugar, Hemoglobin, and Blood Urea. Age is pivotal due to its strong correlation with CKD risk, Albumin serves as an indicator of kidney function, Sugar is linked to diabetes—a significant contributor to CKD, Hemoglobin is essential for understanding anemia's impact on CKD, and Blood Urea indicates kidney function and overall health. Excluded features such as ID, redundant blood pressure data, and various urine components were omitted due to their lack of specificity or redundancy in relation to CKD. This focused selection strikes a balance between predictive power and model simplicity, ensuring effective and accurate assessment of CKD risk.

Early Prediction of Chronic Kidney Disease Using Machine Learning Problem Statement Report: Click Here

Activity 2: Model Selection Report

The Model Selection Report details the rationale behind choosing Random Forest, Decision Tree, Logistic Regression, and KNN models for chronic kidney disease (CKD) prediction. It considers each model's strengths in handling complex relationships, interpretability, adaptability, and overall predictive performance, ensuring an informed choice aligned with project objectives.

Early Prediction of Chronic Kidney Disease Using Machine Learning Problem Statement Report: Click Here

Activity 3: Initial Model Training Code, Model Validation and Evaluation Report

The Initial Model Training Code employs selected algorithms on the chronic kidney disease (CKD) dataset, setting the foundation for predictive modeling. The subsequent Model Validation and Evaluation Report rigorously assesses model performance, employing metrics like accuracy and precision to ensure reliability and effectiveness in predicting CKD outcomes. This process establishes a robust framework for early detection and risk assessment of chronic kidney disease, potentially improving patient care and management strategies.

Early Prediction of Chronic Kidney Disease Using Machine Learning Problem Statement Report: Click Here

Milestone 4: Model Optimization and Tuning Phase

The Model Optimization and Tuning Phase involves refining machine learning models for peak performance. It includes optimized model code, fine-tuning hyperparameters, comparing performance metrics, and justifying the final model selection for enhanced predictive accuracy and efficiency.

Activity 1: Hyperparameter Tuning Documentation

The Random Forest model was chosen for its exceptional performance, demonstrating remarkable accuracy following rigorous hyperparameter tuning. Its capacity to manage intricate data relationships, mitigate overfitting risks, and enhance predictive precision perfectly meets the project's objectives. By leveraging an ensemble of decision trees, the model achieves robustness and generalizability across various tasks. Furthermore, the Random Forest's inherent feature importance analysis reveals crucial insights into the influential factors, bolstering result interpretability. With these attributes and its strong predictive capabilities, the Random Forest emerges as the optimal choice for our final model.

Early Prediction of Chronic Kidney Disease Using Machine Learning Problem Statement Report: Click Here

Activity 2: Performance Metrics Comparison Report

The Performance Metrics Comparison Report evaluates the baseline and optimized performance metrics across four models: Random Forest, Decision Tree, Logistic Regression, and K-Nearest Neighbors (KNN). This analysis offers insights into each model's predictive accuracy before and after fine-tuning. Notably, the report emphasizes the Random Forest model's superior performance, showcasing the highest accuracy among all models tested. This detailed comparison allows us to quantify the enhancements achieved through optimization and reinforces our choice of the Random Forest as the preferred model for its outstanding predictive capabilities.

Early Prediction of Chronic Kidney Disease Using Machine Learning Problem Statement Report: Click Here

Activity 3: Final Model Selection Justification

The Random Forest model was chosen due to its impressive 95.8% accuracy achieved during hyperparameter tuning. Its ensemble method effectively manages complex data relationships and minimizes overfitting. Additionally, the model's capability to offer feature importance rankings, along with its strong performance, makes it well-suited for the project's goals of high predictive accuracy and interpretability.

Early Prediction of Chronic Kidney Disease Using Machine Learning Problem Statement Report: Click Here

Milestone 5: Project Files Submission and Documentation

For the documentation, kindly refer to the link. Click Here

Milestone 6: Project Demonstration

For the demonstration video, kindly refer to the link. Click Here