Project Proposal

Heart Failure Prediction Using Clinical Records with PySpark and Machine Learning

Bijay Rijal
COLLEGE ID: 230567
COVENTRY ID: 14825321

Table of Contents

Introduction

Cardiovascular diseases, particularly heart failure, are leading causes of mortality worldwide. Early prediction can improve patient outcomes and reduce healthcare costs (Chicco & Jurman, 2020). In this project, I will develop a model to predict heart failure using clinical records. Utilizing PySpark for data processing and machine learning techniques, the goal is to create an accurate prediction model. The dataset contains 13 clinical parameters and 5000 patient records from Kaggle (Heart Failure Prediction - Clinical Records, 2024).

Objective

The goal of this project is to create a machine learning model that accurately predicts heart failure using clinical records. By leveraging PySpark and various machine learning algorithms, I aim to identify key predictors and improve the prediction accuracy.

Dataset Description

The dataset contains the medical records of 5000 patients who had heart failure, where each patient profile has 13 clinical features.

- age: age of the patient (years)
- anaemia: decrease of red blood cells or hemoglobin (boolean)
- creatinine phosphokinase (CPK): level of the CPK enzyme in the blood (mcg/L)
- diabetes: if the patient has diabetes (boolean)
- ejection fraction: percentage of blood leaving the heart at each contraction (percentage)
- high blood pressure: if the patient has hypertension (boolean)
- platelets: platelets in the blood (kiloplatelets/mL)
- serum creatinine: level of serum creatinine in the blood (mg/dL)
- serum sodium: level of serum sodium in the blood (mEq/L)
- sex: woman or man (binary)
- smoking: if the patient smokes or not (boolean)
- time: follow-up period (days)
- [target] death event: if the patient died during the follow-up period (boolean)

Conclusion

By the end of this project, my aim is to create a strong machine learning model that can predict heart failure using medical records. This model could help doctors detect heart failure early and take action sooner, leading to better care and outcomes for patients.

References

Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making, 20(1). https://doi.org/10.1186/s12911-020-1023-5

Heart failure prediction - clinical records. (2024, May 5). Kaggle. Heart Failure Prediction - Clinical Records. (2024, May 5). Kaggle. https://www.kaggle.com/datasets/aadarshvelu/heart-failure-prediction-clinical-records/data

Dataset Link: https://www.kaggle.com/datasets/aadarshvelu/heart-failure-prediction-clinical-records/data