ID2221 Data Intensive Computing - Project Proposal - Group 19

by Artem Sliusarenko, Mihailo Cvetkovic and Eugen Lucchiari Hartz

Healthcare Data Analysis for Predictive Modeling

1.) Problem Description

The project aims to predict the likelihood of diseases, such as diabetes or heart disease, using patient health records. By analyzing historical healthcare data, the goal is to build predictive models that can identify individuals at high risk for these conditions. This can assist healthcare providers in early detection and personalized treatment plans, potentially improving patient outcomes and reducing healthcare costs.

2.) Tools

- Apache Spark MLlib for classification algorithms like Decision Trees and Random Forests
- HDFS to manage the large healthcare dataset
- Code will be written using PySpark for processing and model building
- Power BI or Tableau for visualizing the results

3.) Data

We plan to use the UCI Heart Disease Dataset. It is publicly available from the UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/heart+Disease
The data will be stored in HDFS for distributed processing with Apache Spark.

4.) Methodology and Algorithm

- Data Preprocessing
- Classification Algorithm:
    - Decision Tree: splits the datasets into subsets to predict outcomes
    - Random Forest: combines the result of multiple decision trees to improve accuracy, provide feature importance which help to identify which patient characteristics are most predictive of disease