ID2221 Project report

EUGEN LUCCHIARI HARTZ MIHAILO CVETKOVIC ARTEM SLIUSARENKO
eugenlh | mihailoc | artems @kth.se
Group 19

October 27, 2024

1 Introduction

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.

1.1 Tools

- Docker to reliably replicate the application
- Apache Spark MLlib for classification algorithms like Random Forests
- HDFS to manage the large healthcare dataset
- PySpark to implement data processing, feature engineering, and machine learning models
- Results are stored in the Dynamo NoSQL database

2 Dataset

The dataset used for the project was the processed cleveland dataset available for downloading at https://archive.ics.uci.edu/dataset/45/heart+disease. It contains patient health records with various medical indicators and diagnostic results. Each record includes the following columns:

- **age**: Age of the patient in years
- **sex**: Sex of the patient.
  - 1 = Male
  - 0 = Female
- **cp (chest pain type)**: Categorized into four types.
  - 1 = Typical angina: chest pain related to decreased blood supply to the heart.
  - 2 = Atypical angina: chest pain not related to the heart.
  - 3 = Non-anginal pain: typically esophageal or another form of chest pain.
  - 4 = Asymptomatic: no chest pain.

- **trestbps (resting blood pressure)**: Blood pressure measured at rest (in mm Hg) upon admission to the hospital.
- **chol (cholesterol)**: Serum cholesterol level in mg/dl.
- **fbs (fasting blood sugar)**: Indicates if fasting blood sugar ¿ 120 mg/dl.
  - 1 = True (higher than 120 mg/dl)
  - 0 = False (120 mg/dl or lower)
- **restecg (resting electrocardiographic results)**:
  - 0 = Normal.
  - 1 = ST-T wave abnormality (T wave inversions and/or ST elevation or depression of ¿ 0.05 mV).
  - 2 = Left ventricular hypertrophy by Estes' criteria.
- **thalach (maximum heart rate achieved)**: The highest heart rate reached during a stress test.
- **exang (exercise-induced angina)**:
  - 1 = Yes (angina induced by exercise)
  - 0 = No (no angina induced by exercise)
- **oldpeak**: ST depression induced by exercise relative to rest, indicating the difference between the heart's state during rest and exercise.
- **slope**: Slope of the peak exercise ST segment.
  - 1 = Upsloping: better heart rate recovery.
  - 2 = Flat: minimal or no change in the ST segment.
  - 3 = Downsloping: indicative of heart problems.
- **ca (number of major vessels colored by fluoroscopy)**: Number of major blood vessels (0-3) visible after injecting a contrast dye.
- **thal (thalassemia type)**: A blood disorder affecting hemoglobin levels.
  - 3 = Normal.
  - 6 = Fixed defect: no proper blood movement in part of the heart.
  - 7 = Reversible defect: impaired blood flow, but not permanent.
- **target**: Diagnosis of heart disease.
  - 0 = No heart disease.
  - 1 = Presence of heart disease.

3 Methodology

The project's methodology included the following main stages: data preparation, model training, and performance evaluation.

- **Data Loading and Preprocessing:**
  - The dataset was loaded from HDFS, which efficiently manages large datasets.

  - Preprocessing steps included converting data types and encoding categorical variables (like sex and chest pain type).
  - Records with missing values were removed to simplify analysis.

- **Data Balancing:**
  - To correct class imbalances, minority classes were oversampled which helps the model to predict all categories more accurately.

- **Feature Scaling:**
  - Standardization was used to scale features which ensures that attributes with large ranges did not overly influence the model.

- **Model Selection and Training:**
  - Both Logistic Regression and Random Forest were tested.
  - Logistic Regression served as a baseline model and provided an initial accuracy and F1 score.
  - Random Forest, an ensemble of decision trees, improved predictive power and allowed for feature analysis. Hyperparameters (e.g., number of trees, tree depth) were optimized using cross-validation.

- **Evaluation and Tuning:**
  - Accuracy and F1 score assessed the performance of the model which provided insights into both overall prediction accuracy and the precision of class predictions.
  - Cross-validation further optimized the Random Forest model which refines its settings and confirms the best feature set.

- **Storing Results:**
  - Predictions and feature importances were saved in a DynamoDB NoSQL database for easy retrieval and analysis.

4 Results

Model performance was evaluated primarily through accuracy and F1 score, with feature importance analysis to identify key predictors.

- **Logistic Regression Performance:**
  - **Accuracy:** 0.59
  - **F1 Score:** 0.59
  - Logistic Regression, used as a baseline, had moderate scores which indicates limited effectiveness in handling complex patterns in the dataset.

- **Random Forest Classifier Performance:**
  - **Tuned Model:**
    - **Accuracy:** 0.81
    - **F1 Score:** 0.81
    - Cross-validation improved performance which shows Random Forest's strength in capturing complex relationships in the data.

- **Feature Importance:** Key predictors identified in the analysis:
  - **restecg (Resting ECG results):** Measures abnormalities in the heart's electrical activity while at rest, such as irregular rhythms or signs of heart strain. This was the most influential feature, which shows that abnormalities in heart electrical patterns are strong indicators of potential heart disease.
  - **age, exang (Exercise-induced angina), and cp (Chest Pain type):** These features also had significant predictive power:
    - **age:** Higher age often correlates with increased heart disease risk.
    - **exang:** Indicates whether the patient experienced angina (chest pain) during exercise. Angina induced by physical activity can signal reduced blood flow to the heart.
    - **cp (Chest Pain type):** Different types of chest pain are associated with various heart conditions, making this a valuable predictor.
  - **Less impactful features: chol (Cholesterol)** and **fbs (Fasting blood sugar):** These had lower importance in this dataset, suggesting they contributed less to the model's prediction of heart disease.

Glossary of Technical Terms

- **Cross-Validation:** A technique for evaluating model performance by dividing data into multiple subsets, training on some and testing on others, to ensure robust and reliable results.

- **F1 Score:** A metric that combines precision (accuracy of positive predictions) and recall (ability to find all positive instances) into a single score, useful especially in imbalanced datasets.

- **Feature Importance:** A ranking of features based on their contribution to the model's predictions, helping identify which factors most influence outcomes.

- **Hyperparameter Tuning:** The process of adjusting model settings (e.g., number of trees in Random Forest) to improve performance. Cross-validation is often used in conjunction with this process.

- **Oversampling:** A technique to address class imbalance by duplicating examples from underrepresented classes, creating a more balanced dataset for training.

- **Random Forest:** An ensemble learning algorithm that builds multiple decision trees and combines their outputs for more accurate and stable predictions.

- **Standardization (Feature Scaling):** The process of rescaling feature values so they have a mean of zero and a standard deviation of one, allowing for fair comparison between features with different scales.

5 Instructions

The most important prerequisite to be able to run the project is Docker. We use Docker to be able to reliably replicate the project for collaboration.

Below are the steps to run the project.

1. Build docker image that has Apache Spark and Jupyter Notebook.

   From the main project directory.

   ```
   $ cd apache-spark_jupyter
   $ docker build -t spark-jupyter .
   ```

2. Start the HDFS, Dynamo Local and Dynamo Admin, Spark, and Jupyter.

   From the main project directory.

   ```
   $ cd docker-hadoop-m1
   $ docker-compose up -d
   ```

3. Before we can run any machine learning algorithms on our data we need to prepare the folder structure and copy the data inside the input folder.

   From the **‘docker-hadoop-m1‘** directory.

   ```
   $ ./hdfs dfs -mkdir -p /input
   $ ./hdfs dfs -copyFromLocal -f /app/data/heart_decease.csv /input/
   ```

4. Navigate to the Jupyter Notebook UI

   Open Jupyter Notebook and naviagate to the file called **`process_data.ipynb`**.

5. Execute code in the Jupyter Notebook.

6. View the results.

   Navigate to the DynamoDB Admin UI and open the table called **processed** to view the results.

7. Teardown

   From the main project directory.

   ```
   $ cd docker-hadoop-m1
   $ docker-compose down -v
   ```