Course: COMP 4334
Assignment: 2
Author: Michael Ghattas

Project Overview

This project applies **Apache Spark's MLlib and Structured Streaming** to solve a binary classification problem using a real-world medical dataset. The task involves training a **logistic regression model** to predict the presence of **heart disease** based on demographic and clinical features. After training, we simulate a streaming environment where test data is incrementally evaluated using the trained model.

Dataset Description

- **Source**: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction
- **Format**: CSV (`heart.csv`)
- **Records**: 918 patients
- **Target Variable**: `HeartDisease` (1 = has heart disease, 0 = no heart disease)
- **Key Features Used**: `Age`, `Sex`, `RestingBP`, `Cholesterol`, `FastingBS`, `RestingECG`, `MaxHR`, `ExerciseAngina`, `Oldpeak`, `ST_Slope`

Machine Learning Pipeline

The pipeline includes the following transformations: - **Bucketizer**: Bins `Age` into 5 categories. - **StringIndexers**: Encode categorical variables (`Sex`, `ChestPainType`, `RestingECG`, `ExerciseAngina`, `ST_Slope`). - **VectorAssembler**: Combines all features into a single feature vector. - **LogisticRegression**: Binary classification model trained to predict `HeartDisease`.

Training:
- 70% of the dataset was used for training.
- The remaining 30% was used for testing and streaming simulation.

Streaming Architecture

- The test set was re-partitioned into multiple CSV files to **simulate streaming ingestion**.
- Spark's **Structured Streaming API** was used to:
  - Read files one at a time from a DBFS directory.
  - Apply the trained pipeline to each incoming batch.
  - Store predictions in a **memory sink** for query and inspection.

Model Evaluation

Static Test Evaluation

- **Accuracy**: `0.8608`
- **AUC (ROC)**: `0.9073`
- **Confusion Matrix**:

Actual	Predicted	Count
0	0	82
0	1	17
1	0	16
1	1	122

Streaming Output

- **Total streamed records processed**: `237`
- **Sample predictions**:

Age	Sex	HeartDisease	Prediction	Probability
60	M	1	1.0	[0.02, 0.97]
42	M	0	0.0	[0.85, 0.15]

Exploratory Data Analysis (EDA)

Summary Statistics

- Mean Age: `53.5`, Cholesterol: `198.8`, MaxHR: `136.8`
- Some data entries (e.g., `Cholesterol = 0`) may represent missing values or poor data quality.

Heart Disease Class Distribution

- `HeartDisease = 1`: 508 patients
- `HeartDisease = 0`: 410 patients

Heart Disease Rate by Age Group

Age Group	Total	Disease Rate
< 40	80	32.5%
40–49	211	40.3%
50–59	374	56.7%
60–69	222	73.4%
70+	31	70.9%

Cholesterol by Heart Disease Status

HeartDisease	Avg Cholesterol
0	227.12
1	175.94

Heart Disease Rate by Sex

Sex	Total	Disease Rate
F	193	25.9%
M	725	63.2%

Conclusion

This project successfully integrates SparkML and Structured Streaming to simulate a real-time machine learning workflow. A logistic regression classifier achieved **86% accuracy and 0.91 AUC**, driven by meaningful age and gender patterns in the dataset. The streaming pipeline is robust and can process incoming batches while maintaining prediction accuracy. The analysis findings highlight the importance of demographic factors—especially age and sex—in heart disease prediction.

The exploratory analysis revealed clear patterns in the dataset:

- Heart disease prevalence increases with age, particularly after 50, with rates rising from ~32% in the under-40 group to over 73% in the 60–70 age group.
- Men were significantly more likely to have heart disease than women, with a prevalence of ~63% versus ~26%.
- Surprisingly, average cholesterol levels were higher in individuals without heart disease, which may point to confounding factors or treatment history not captured in the dataset.

END.