Evaluating Classifier Performance using PySpark for Cardiovascular Disease Prediction

1st Kavita Dharamshaktu
Information Technology
Indian Institute of Information Technology, Allahabad
iit2020002@iiita.ac.in

2nd Sanskruti Mane
Information Technology and Business Informatics
Indian Institute of Information Technology, Allahabad
iib2020020@iiita.ac.in

3rd Priyamvada Priyadarshani
Information Technology and Business Informatics
BTech in Indian Institute of Information Technology, Allahabad
iib2020037@iiita.ac.in

4th Prachi Gupta
Electronics and Communication Engineering
Indian Institute of Information Technology, Allahabad
iec2020040@iiita.ac.in

***Abstract*—Cardiovascular disease (CVD) poses a significant global health challenge, demanding proactive interventions. Leveraging Internet of Things (IoT) devices and big data analytics, this study aims to transform CVD management. The integration of IoT devices in healthcare, though complex, holds potential, despite challenges of data security, privacy, and data volume management. The methodology relies on PySpark, enabling real-time data processing, storage, and analysis. The implementation involves data preprocessing, machine learning model building, and performance evaluation. Utilizing machine learning models such as Random Forest Regression, Logistic Regression, Decision Tree Classifier, Gradient Boosted Tree (GBT) Classifier, SVM, Multilayer perception, and Random Forest Classifier, this approach aims to facilitate early detection and effective management of CVD. The study emphasizes the potential of this multifaceted approach to revolutionize CVD management, ensuring timely interventions, cost reductions, and ultimately, the preservation of lives.**

Keywords: Cardiovascular disease prediction(CVD), Big data analytics, PySpark, Multilayer Perceptron, Logistic Regression, Decision Tree, Random Forest Classifier, SVM, GBT

I. Introduction

Cardiovascular disease (CVD) remains a global health challenge, contributing significantly to the worldwide disease burden and claiming millions of lives annually. With over 500 million individuals afflicted by CVD and a projected 20.5 million fatalities in 2021 [1], it constitutes a substantial portion of global mortality. This multifaceted condition encompasses various disorders affecting the heart and blood vessels, emphasizing the critical need for preventive measures.

Risk factors for CVD, spanning socioeconomic status, lifestyle choices, and environmental exposures, play a pivotal role in its prevalence. (Fig. 1)The alarming rise in CVD-related deaths from 12.1 million in 1990 to 18.6 million in 2019 underscores the urgency of addressing modifiable risk factors. Elevated LDL cholesterol, air pollution, high fasting plasma glucose levels, obesity-related causes, lack of exercise, and hypertension contribute significantly to the escalating mortality rates [1].

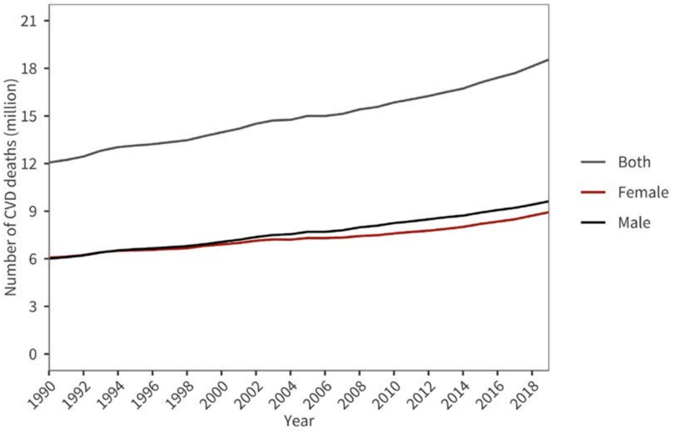

Fig. 1: Global trends in number of deaths due to cardiovascular diseases, 1990-2019 [11]

Recognizing the urgency to mitigate the impact of CVD, our paper aligns its objectives with addressing these risk factors. Leveraging a carefully chosen dataset, we aim to contribute to the understanding and prediction of CVD by utilizing advanced data analytics tools. In recent years, the integration of Internet of Things (IoT) devices has shown promise in transforming CVD management [2]. These interconnected devices, capable of real-time monitoring through wearable sensors and implantable monitors, provide continuous insights into vital signs and physiological parameters. This rich dataset allows healthcare providers to make informed decisions and intervene promptly.

Furthermore, the synergy of big data analytics with IoT technology has facilitated the development of sophisticated algorithms for early detection of cardiovascular changes. Our paper acknowledges the significance of this technological convergence and aims to leverage the Spark architecture, particularly its PySpark API's MLib framework, to fit several models on our heart disease dataset for CVD prediction.

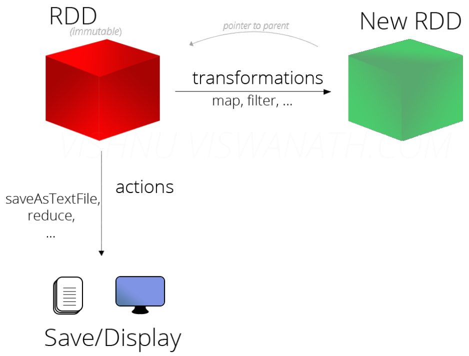

Fig. 2: RDD [12]

PySpark's integration into cardiovascular disease management signifies the fusion of advanced data processing and storage technologies. The paper underscores the efficiency of PySpark in manipulating and analyzing vast datasets from IoT devices, facilitating real-time analysis for accurate diagnoses and personalized treatment.

The Resilient Distributed Dataset (RDD)(Fig. 2), a foundational element of Apache Spark, plays a pivotal role in our analytical approach. RDD's ability to process data in parallel over a cluster of computers, support various data transformations, and be cached in memory enhances the efficiency of our data analysis. The robust ecosystem of Spark libraries further enriches our analytical capabilities.

Despite the potential benefits of IoT and big data analytics in CVD management, challenges such as data security, privacy, and the need for robust infrastructure must be addressed. The paper recognizes these challenges and emphasizes the importance of implementing comprehensive solutions to ensure successful technology adoption in healthcare.

In conclusion, the convergence of IoT devices, big data analytics, and advanced data processing platforms holds immense potential for revolutionizing CVD management. By harnessing the power of real-time data monitoring, predictive analytics, and secure data storage, healthcare organizations can usher in a new era of proactive, patient-centered care. Through continuous innovation and collaboration, the healthcare industry can aspire to significantly reduce the burden of CVD and enhance the quality of life for individuals at risk of or living with cardiovascular conditions. Cardiovascular disease (CVD) remains a leading global cause of death, contributing to a substantial portion of the disease burden. With over 500 million individuals affected worldwide, CVD is projected to cause around 20.5 million fatalities in 2021. This encompasses disorders like coronary artery disease, heart failure, and stroke. Despite advancements in medical technology, the prevalence of CVD poses significant challenges globally. Preventable risk factors include socioeconomic status, lifestyle choices, and environmental exposures such as hypertension, diabetes, obesity, smoking, and stress. From 12.1 million deaths in 1990, CVD-related fatalities rose to 18.6 million in 2019, highlighting the persistent impact of this multifaceted condition.

II. LITERATURE REVIEW

In the dynamic realm of healthcare technology, a series of studies have significantly contributed to the exploration of big data analytics (BDA) and the Internet of Things (IoT) for the monitoring of cardiovascular health. The research conducted in 2019 by [2] was instrumental in developing and testing a system tailored to handle large-scale, high-velocity data originating from diverse IoT sensors. While the study provided valuable insights into the intricacies of managing IoT-generated data, its limitations were apparent in the absence of real-world case studies and a comprehensive analysis, confining its scope to just two specific patient monitoring applications.

Advancing to 2020, another noteworthy endeavor by [4] presented a scalable and real-time system designed for disease prediction through big data processing. The system's architecture processed health-related attributes from user-generated tweets in real-time, employing Spark streaming and Kafka. However, a notable limitation surfaced concerning the oversight of incorporating IoT device data, a potential avenue for 24-hour surveillance that could significantly broaden the system's scope.

Within the same year, [5] implemented a real-time healthcare monitoring system using Apache Spark, Scala, Spark Streaming, and Kafka. This system processed streaming data to predict patient status utilizing the Streaming Linear Regression With SGD algorithm. Despite its promising approach, the study identified challenges encompassing data quality, privacy concerns, model training, and generalization.

Fast-forwarding to 2022, [3] introduced a machine-learning approach for real-time heart disease detection, showcasing the effectiveness of a stacked model in outperforming other models. However, a notable drawback was the absence of real-time analysis and the underutilization of BDA tools. Recommendations for future research include integrating real-time analysis and leveraging BDA frameworks to enhance system efficiency in terms of speed and accuracy, along with delving into the exploration of different machine learning algorithms.

In the same year, [6] provided a comprehensive exploration of the state-of-the-art and future challenges related to big data for real-time processing on streaming data. The study involved data ingestion using the Twitter Streaming API and Apache Flume, with Apache Kafka facilitating efficient data transport. Observations underscored the efficiency of Kafka for handling high-velocity data, albeit acknowledging the potential increase in storage costs due to duplicates. Furthermore, Spark was recognized as an ideal tool for iterative analysis but found less suitable for one-time data processing before storage.

In summary, the combined efforts of these studies have significantly propelled our understanding of harnessing big data, IoT, and real-time processing for cardiovascular health monitoring. While each study contributes valuable insights, the

TABLE I: Table of Literature Review

Title of the Paper	Year	Methodology	Limitations
Visualisation and prediction of Heart disease using Big Data Analytics [3]	2022	Proposes a machine learning approach for real-time heart disease detection, with a stacked model outperforming others.	Not real-time analysis and no BDA tools used. In the future, real-time analysis and BDA framework utilization could enhance system efficiency in terms of speed and accuracy and analyze different ML algorithms further.
Big data and IoT solution for patient behavior monitoring [2]	2019	Develops and tests a system for handling large-scale, high-velocity data from diverse IoT sensors, including those for blood pressure, body temperature, electrocardiogram, and other biometric measurements.	Lacks real-world case studies and comprehensive analysis, focusing on just two patient monitoring applications.
Big Data for Real-Time Processing on Streaming Data: State-of-the-art and Future Challenges [6]	2022	Data ingestion with Twitter Streaming API and Apache Flume has been done. Apache Kafka is employed to transport data efficiently. Data preprocessing entails activities such as filtering out incorrect meanings, removing URL links, and handling user mentions and emoticons in tweet content.	Kafka is great for high-velocity data but can increase storage costs with duplicates. Spark is ideal for iterative analysis but less so for one-time data processing before storage.
Real-Time Healthcare Monitoring System using Online Machine Learning and Spark Streaming [5]	2020	Implemented on Apache Spark using Scala, the system utilizes Spark Streaming and Kafka for data processing and input. The Streaming Linear Regression With SGD algorithm predicts patient status from streaming data.	Data Quality, Privacy Concerns, Model Training, Generalization.
A scalable and real-time system for disease prediction using big data processing [4]	2020	The system's architecture involves users tweeting health-related attributes, processed in real-time via Spark streaming and Kafka. Machine learning predicts health status from the data, returning relevant messages to users via Twitter.	Did not consider the case of IoT device data to be used as input, especially when IoT devices provide the opportunity of 24-hour surveillance.

collective body of work highlights the ongoing need for further exploration, particularly in addressing specific challenges such as real-time analysis, data quality, privacy concerns, and the seamless integration of diverse data sources, including IoT-generated data.

III. Problem Statement

The paper addresses critical challenges in cardiovascular disease (CVD) management. Firstly, current approaches are reactive, identifying CVD only after noticeable symptoms, and missing early intervention opportunities. Secondly, the management of vast volumes of sensitive patient data poses a significant burden. Lastly, the integration of IoT devices with advanced tools like PySpark and MongoDB introduces technical challenges. The paper aims to propose solutions to these issues for proactive detection, efficient data management, and seamless IoT integration in CVD management.

IV. Proposed Methodology

Apache Spark has resilient distributed database rates as its heart does faster processing due to the in-memory usage. It is a batch processing system and converts the real-time data stream into window batches for processing. It is also lambda architecture. Some of its main components are hdfs or Hadoop over which it runs, yarn/mesos/standalone manager for resource management. SparkMl for machine learning libraries, GraphX for graph processing, SparkSQL for SQL querying, and Spark streaming for data stream processing. All these extensions and their drivers and executors are handled by Spark Core responsible for overall management. We are using a standalone single-machine spark. We utilized Google Colab, or Colaboratory, as a free cloud-based platform for our research. It allowed us to work with large datasets and complex tasks without the need for costly hardware. With the ”!pip install pyspark” command, We installed PySpark, which is essential for big data processing and machine learning tasks.

Colab's key advantage is its cloud-based, scalable solution. It provides access to high-performance Spark clusters without the burden of infrastructure management. This was particularly valuable for our work, which involved large datasets and intricate machine-learning models. By establishing a SparkSession, We seamlessly integrated PySpark's capabilities with Colab's collaborative features. This combination enabled us to efficiently preprocess data, train models, and perform in-depth analysis for heart disease prediction, all within a cloud-based environment.

A. Data collection

The dataset, sourced from 'cardio.csv,' includes health-related details on individuals (e.g., age, gender, blood pressure, etc.). Before analysis, preprocessing steps were taken, including removing the 'id' column, handling missing values, and standardizing features. Exploratory data analysis revealed insights into cardiovascular disease factors. The structured dataset is designed for investigating cardiovascular health associations.

B. PySpark Machine Learning

- Spark Version: 3.5.0
- Master: local[]*
- Application Name: Colab

The Architecture of our methodology discussed in our paper is shown in Fig. 3

1) Data Preprocessing : To prepare the data for modeling, we conducted the following data preprocessing steps: After further research, we found that our data does not contain any missing values. We introduced a new feature, `age in years` by dividing the age column by 365, allowing us to work with age in years. We created another feature bmi using weight and height to test it's significance in the dataset also.

In machine learning, it's essential to split the dataset into two subsets: one for training the model and the other for testing the model's performance. The purpose of this split is to evaluate the model's generalization and ensure that it can make accurate predictions on new, unseen data. 'train-data, test-data = assembled-df.randomSplit([0.8, 0.2], seed=123)'. We partitioned the dataset into training and testing subsets, with approximately 80 percent (56,000 data points) allocated to the training set and the remaining 20 percent (14,000 data points) reserved for the testing set. We also scaled the features using StandardScaler.

2) Data Overview: Our dataset comprises several key attributes, including:

1) 'id': A unique identifier for each individual.
2) 'age': Age in days of each participant.
3) 'gender': Gender Categorised as '1' for males and '2' for females.
4) 'height': Height in cm of each participant.
5) 'weight': Weight in kg of each individual.
6) 'ap-hi': Systolic blood pressure.
7) 'ap-lo': Diastolic blood pressure.
8) 'cholesterol': Cholesterol levels are classified into categories '1', '2', or '3'.
9) 'gluc': Glucose levels are classified as '1', '2', or '3'.
10) 'smoke': Smoking status, where '0' represents non-smokers and '1' represents smokers.
11) 'alco': Alcohol intake status, '0' indicating non-drinkers and '1' for individuals who consume alcohol.
12) 'active': Physical activity levels, with '0' indicating inactive and '1' representing active individuals.
13) 'cardio': Binary variable/target to indicate the presence or absence of cardiovascular disease, where '0' signifies no disease and '1' represents the presence of cardiovascular disease.
14) age-in-years: A feature created to represent a person's age in years, which is often more interpretable and directly usable for predictive modeling.
15) bmi: A feature engineered to calculate the Body Mass Index (BMI) of an individual, providing insight into their body weight relative to their height and potential health risks.

For males (gender = 1), the average age is approximately 52.95 years. For females (gender = 2), the average age is approximately 52.63 years. In the dataset, we have a total of 22,616 individuals of male gender and 12,363 individuals of the female gender. Among the male population, 9,316 individuals have been diagnosed with heart disease, while among the female population, 7,263 individuals have received the same diagnosis. It is evident that heart disease affects a substantial portion of both genders, with males showing a higher count of cases compared to females. Understanding the gender distribution and its association with heart disease is a crucial aspect of cardiovascular health research. We employed this dataset to explore and analyze the factors that may influence the likelihood of developing cardiovascular disease. This collection of attributes allowed us to conduct a comprehensive investigation into the relationship between various health-related features and the occurrence of cardiovascular disease.

TABLE II: Dataset Summary

Attribute	Count	Mean	Std Dev	Min	Max
id	70,000	49,972	28,851	0	99,999
age	70,000	19,468.87	2,467.25	10,798	23,713
gender	70,000	1.35	0.48	1	2
height	70,000	164.36	8.21	55	250
weight	70,000	74.21	14.40	10.0	200.0
ap-hi	70,000	128.82	154.01	-150	16,020
ap-lo	70,000	96.63	188.47	-70	11,000
cholesterol	70,000	1.37	0.68	1	3
gluc	70,000	1.23	0.57	1	3
smoke	70,000	0.088	0.28	0	1
alco	70,000	0.054	0.23	0	1
active	70,000	0.804	0.40	0	1
cardio	70,000	0.4997	0.5000	0	1

3) Data Loading and Exploration:: We initiated the workflow by loading a dataset from a CSV file into a PySpark DataFrame. This dataset contains health-related information,

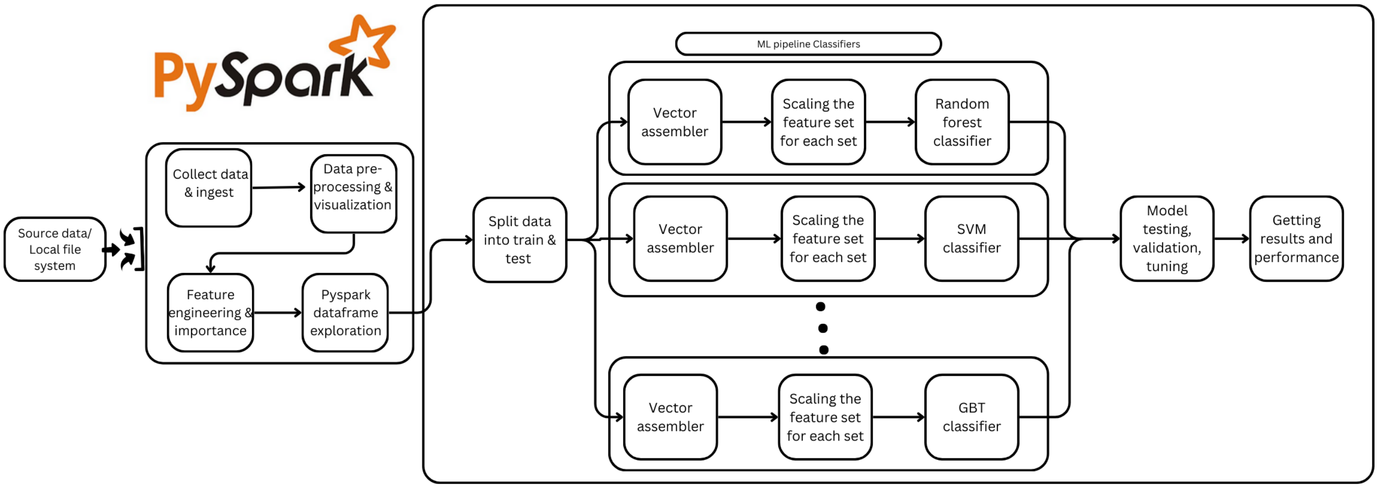
PySpark
Source data/ Local file system
Collect data & ingest
Data pre-processing & visualization
Feature engineering & importance
Pyspark dataframe exploration
Split data into train & test
ML pipeline Classifiers
Vector assembler
Scaling the feature set for each set
Random forest classifier
Vector assembler
Scaling the feature set for each set
SVM classifier
Vector assembler
Scaling the feature set for each set
GBT classifier
Model testing, validation, tuning
Getting results and performance

Fig. 3: Flowchart of Proposed Methodology

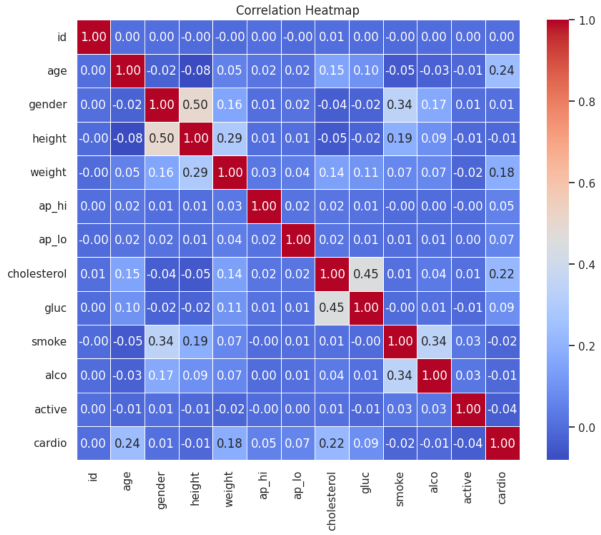

Fig. 4: correlation Heatmap

and we performed preliminary data exploration to understand its characteristics. Specifically, we employed the following operations:

- Summary Statistics: We utilized the `describe()` function to obtain summary statistics, providing insights into the dataset's numerical attributes.
- Unique Value Counts: We counted the unique values in the 'gender' column using the `groupBy()` operation to understand the distribution of gender within the dataset.
- Correlation Analysis: We visualized the correlation between the 'cholesterol' and 'cardio' columns and others as well like Fig. **??**. to identify potential relationships between these variables. We also plotted the correlation heatmap as shown in Fig. 4. A correlation heatmap is a graphical representation of the correlation between variables in a dataset. It is commonly used in data analysis and statistics to visualize the relationships between different variables. Correlation measures how two variables are related to each other, and it can be used to identify patterns, associations, and dependencies within the data. The correlation coefficients for all pairs of variables are organized into a matrix. This matrix is then visualized as a heatmap, where each cell's color represents the strength and direction of the correlation. Typically, a color gradient is used, with warmer colors (e.g., shades of red) indicating positive correlations, and cooler colors (e.g., shades of blue) indicating negative correlations. A cell close to white or a neutral color suggests little to no correlation.
- Cross-Tabulation: We performed a cross-tabulation between the 'cardio' and 'smoke' columns to understand how smoking behavior relates to cardiovascular health.

In terms of feature importance:

- age has a moderate positive correlation (0.238) with the target variable cardio, indicating that age is an important predictor of cardiovascular disease.
- cholesterol also has a moderate positive correlation (0.221) with cardio, suggesting that cholesterol levels play a role in predicting cardiovascular disease.
- weight has a positive correlation of 0.181 with cardio, indicating that it may be relevant in predicting cardiovascular disease.
- gluc (systolic blood pressure) has a positive correlation of 0.089307 with cardio, implying some importance in predicting cardiovascular disease.

On the other hand:

- smoke and alco (smoking and alcohol consumption) have relatively weak correlations with cardio, suggesting that these features may not be as important in predicting cardiovascular disease.
- active (physical activity) has a weak negative correlation (-0.036) with cardio, indicating a small influence in the opposite direction.

We performed a feature importance test using a random forest.

- ap-hi (Systolic Blood Pressure): This feature has the highest importance score (0.4886), making it the most influential feature for predicting cardiovascular disease. It suggests that high systolic blood pressure is a significant indicator of the risk of cardiovascular disease.
- ap-lo (Diastolic Blood Pressure): The diastolic blood pressure (ap-lo) also has a substantial importance score (0.2905), indicating that it is the second most important feature for predicting cardiovascular disease. Both systolic and diastolic blood pressure measures are vital health indicators.
- age: Age is the third most important feature (0.1017). This suggests that age plays a significant role in predicting the risk of cardiovascular disease, which aligns with common medical knowledge that age is a risk factor.
- cholesterol: Cholesterol levels (cholesterol) have a moderate importance score (0.0759). High cholesterol levels can contribute to heart disease, so it is a relevant feature.
- weight: Weight (0.0231) and Body Mass Index (bmi: 0.0148) have some importance, indicating that they are relevant but less influential compared to blood pressure and age.
- gluc (Glucose Levels): Glucose levels (gluc) have a low importance score (0.0029), suggesting that they have a relatively minor impact on predicting cardiovascular disease.
- Other Features: These features have very low importance scores (between 0.0003 and 0.0009), indicating that they have minimal impact on predicting cardiovascular disease in this specific model.

We had some different sets of features that we used while training our models and selected the results with the best accuracy values.

- All features
- Correlation-Based Features: 'age,' 'cholesterol,' 'weight,' and 'gluc'
- Custom-Engineered Features: the 'age-in-years' feature, which might be more informative than the original 'age.'
- Feature Importance-Based Features: Train a model (e.g., Random Forest) and calculate feature importance. Select the top N features based on their importance scores. Applied Random forest and Top 5 Feature came out to be [ap-hi: 0.5448; ap-lo: 0.2651; age: 0.0817; cholesterol: 0.0849; weight: 0.0157;]

4) Data preparation & pipelining : After scaling using StandardScalar, assembling the input features into one column, 'features' using VectorAssembler, and renaming the target column, 'cardio' as 'label' using StringIndexer, they are added as stages in the machine learning pipeline of Pyspark. These are some feature engineering techniques that are offered by Spark Core. StringIndexer is a feature transformer in Spark MLlib that plays a crucial role in managing categorical or string features. It is responsible for mapping unique numerical indices to each distinct category found within a string column. This transformation is essential, primarily because many machine learning algorithms require numerical inputs and struggle to work directly with categorical data. VectorAssembler is a feature transformer available in Spark MLlib, designed to streamline the process of merging multiple feature columns into a single vector column. This capability is especially valuable when you need to aggregate various features into a unified feature vector, a common prerequisite for numerous machine learning algorithms.

5) Model Training: We proceeded to build and evaluate several machine learning models to predict the 'cardio' column, which represents the presence or absence of cardiovascular disease.

- Logistic Regression:
  - Versatile model for binary classification, predicting cardiovascular disease occurrence.
  - Evaluated using metrics like accuracy, precision, recall, and F1-score.
- Decision Tree Classifier:
  - Simple yet powerful model dividing datasets based on feature values for predicting cardiovascular disease presence.
  - Evaluation focused on model accuracy.
- Gradient Boosted Tree (GBT) Classifier:
  - Vital for healthcare, particularly in cardiovascular disease prediction.
  - High predictive accuracy, robustness against noisy data, and feature importance analysis for risk factor identification.
- Random Forest Classifier:
  - Ensemble model combining decision trees for cardiovascular disease prediction.
  - Evaluation emphasized accuracy; effective in capturing complex patterns.
- SVM (Support Vector Machine) Classifier:
  - Aims for an optimal decision boundary in high-dimensional feature spaces.
  - Versatile, applicable to binary and multi-class problems in healthcare, especially cardiovascular disease prediction.
- Multilayer Perceptron (MLP) Classifier:
  - Artificial neural network known for handling complex classification problems.
  - Included for binary classification abilities and capturing intricate data relationships.

To assess and compare the performance of different algorithms for cardiovascular disease identification, the use of k-fold cross-validation ensures the stability and reliability of machine learning models.

While a simple fitting of the model to training data may lead to low bias and variance, we need an ideal model that is able to capture the hidden patterns in data form meaningful insights, and strike a balance between bias and variance. The validation

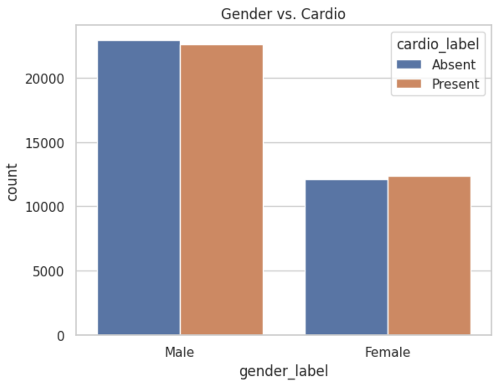

Fig. 5: Gender vs Cardio

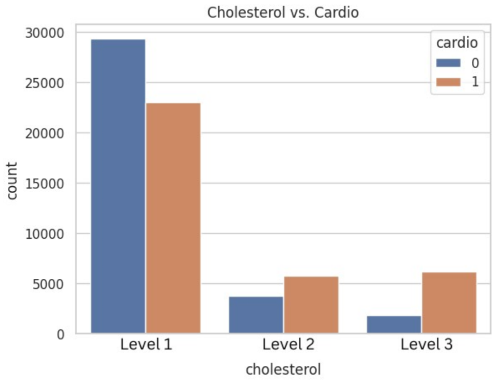

Fig. 6: Cholestrol vs Cardio

set or just plainly dividing the training set and testing set, running one iteration leaves out a significant portion of the dataset and may miss out on some important features. Leave-One-Out Cross Validation leaves one testing data point in each iteration thus being very exhausting and time consuming. It also introduces a lot of variance in the model due to limited testing data. k-fold cross-validation seems to be addressing the need for a balance between training data size, testing data size, and model stability, most accurately. We performed a 5-fold cross-validation and also performed a 10-fold cross-validation for best performing model.

6) Model Evaluation and Visualization: In addition to training the models, we plotted confusion matrices for each classifier. These visualizations provide an in-depth understanding of model performance, particularly in terms of class predictions and the ability to distinguish between classes.

We have successfully implemented a comprehensive machine-learning workflow to predict cardiovascular disease and evaluated the performance of various machine-learning models using the BinaryClassificationEvaluator() of Pyspark. The inclusion of confusion matrices allowed us to visualize and understand the classifiers' strengths and weaknesses.

The work accomplished in this workflow serves as a solid foundation for future model optimization and additional analyses. Further exploration of hyperparameter tuning and ensemble methods may enhance the predictive accuracy of these models for multiple classes or exact heart disease prediction in the future.

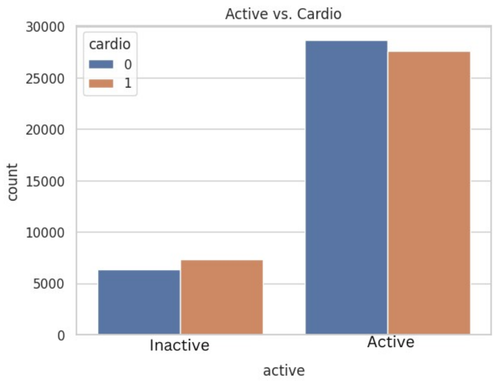

Fig. 7: Active vs Cardio

V. Analysis & Results

TABLE III: Model Performance with Feature Importance-Based Features

Model	Accuracy	F1-score	Precision	Recall
Logistic Regression	0.6109	0.5793	0.6188	0.6109
Multilayer Perceptron	0.6109	0.5793	0.6188	0.6109
Decision Tree	0.6093	0.5783	0.6161	0.6093
SVM (LinearSVC)	0.6060	0.5462	0.6413	0.6060
GBT Classifier	0.6997	0.6958	0.6997	0.6997
Random Forest	0.6091	0.5789	0.6152	0.6091

TABLE IV: Model Performance with all original features

Model	Accuracy	F1-score	Precision	Recall
Logistic Regression	0.7171	0.7166	0.7180	0.7171
Multilayer Perceptron	0.4932	0.3258	0.2432	0.4932
Decision Tree	0.7294	0.7269	0.7365	0.7294
SVM (LinearSVC)	0.7198	0.7177	0.7250	0.7198
GBT Classifier	0.7297	0.7290	0.7312	0.7297
Random Forest	0.7222	0.7214	0.7236	0.7222

TABLE V: Model Performance with Correlation-Based Feature

Model	Accuracy	F1-score	Precision	Recall
Logistic Regression	0.7153	0.7148	0.7162	0.7153
Multilayer Perceptron	0.4932	0.3258	0.2432	0.4932
Decision Tree	0.7226	0.7224	0.7229	0.7226
SVM (LinearSVC)	0.7186	0.7164	0.7240	0.7186
GBT Classifier	0.7247	0.7245	0.7249	0.7247
Random Forest	0.7217	0.7211	0.7229	0.7217

Here are the best-performing models for each feature set:

- Feature Importance-Based Features: GBT Classifier: Accuracy: 0.6997, F1-score: 0.6958, Precision: 0.6997, Recall: 0.6997

TABLE VI: Model Performance with Custom-Engineered Features

Model	Accuracy	F1-score	Precision	Recall
Logistic Regression	0.7163	0.7158	0.7172	0.7163
Multilayer Perceptron	0.7093	0.7090	0.7098	0.7093
Decision Tree	0.7297	0.7290	0.7314	0.7297
SVM (LinearSVC)	0.7199	0.7178	0.7250	0.7199
GBT Classifier	0.7302	0.7296	0.7315	0.7302
Random Forest	0.7242	0.7236	0.7253	0.7242

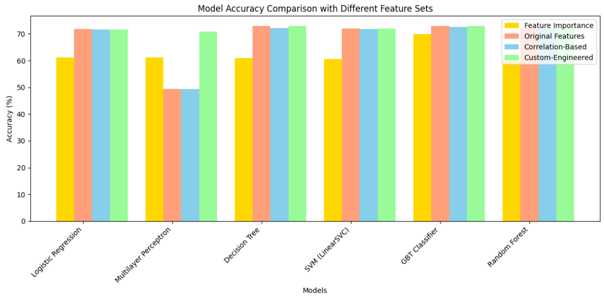

Fig. 8: Accuracies comparison using different features selection.

- All Original Features: GBT Classifier: Accuracy: 0.7297, F1-score: 0.7290, Precision: 0.7312, Recall: 0.7297
- Correlation-Based Features: GBT Classifier: Accuracy: 0.7247, F1-score: 0.7245, Precision: 0.7249, Recall: 0.7247
- Custom-Engineered Features: GBT Classifier: Accuracy: 0.7302, F1-score: 0.7296, Precision: 0.7315, Recall: 0.7302

GBT Classifier consistently performed the best across all feature sets, except for ”Feature Importance-Based Features,” where it performed slightly worse compared to other models. The GBT Classifier likely performed well due to its ability to handle complex relationships in the data and ensemble learning techniques. For feature sets, ”All Original Features” and custom-engineered features yielded the best results, and it’s expected as it contains all the available information. The confusion matrices for all models with all original features selected were drawn and shown in Fig. **??**, while Fig. **??** shows an Accuracy comparison plot. Fig. **??** & Fig. **??** shows Performance metric formulas.

The machine learning model was evaluated on the dataset using an 80-20 train-test split. The GBT model demonstrated an accuracy of 73.20% in predicting cardiovascular disease.

In the testing phase, the model’s performance was further characterized through a confusion matrix, which encapsulates real and predicted classifications, elucidating the model’s effectiveness. Notably, the model correctly identified 4624 true positives (TP) and 5625 true negatives (TN). However, it also produced 1471 false positives (FP) and 2281 false negatives (FN). This translated to a precision of 75.87%, signifying the proportion of positive predictions that were accurate. The recall, also known as sensitivity, stood at 66.97%, depicting the fraction of actual positives correctly identified by the model. The model’s F1 Score, a harmonic mean of precision and recall, reached 71.14%. Furthermore, the model exhibited a specificity of 79.27%, indicating its ability to accurately identify true negatives. We tried KNN which resulted in an accuracy of only 64% which did not seem very promising. 10-fold as well as 5-fold cross-validation when performed on the dataset led to similar results of max 73% accuracy in the case of GBT.

These results reflect the GBT Classifier’s robust performance in distinguishing cardiovascular disease cases, with a balanced trade-off between precision and recall. The tree rules defining the classification were printed and observed as well, for CART, GBT, and random forest. The rules for the Decision tree are shown in Fig. **??**. The figure gives the following information:

- The tree has a depth of 3, meaning that it has three levels or layers of decision nodes.
- The total number of nodes in the tree is 9.
- The model is designed for binary classification, with two classes (numClasses=2).
- It uses 11 features for making classification decisions.

1) The first split is based on ’ap-hi’ (systolic blood pressure), and the decision is made by checking whether ’ap-hi’ is less than or equal to 129.0.
2) If ’ap-hi’ is less than or equal to 129.0, the model proceeds to the left child node, which evaluates ’age’ and checks whether it is less than or equal to 20207.5.
3) If ’age’ is less than or equal to 20207.5, the model goes deeper and examines ’cholesterol’ and checks whether it is less than or equal to 2.5. If this condition is met, the model predicts the class as 0.0.
4) If ’cholesterol’ is greater than 2.5, the model predicts the class as 1.0.
5) If ’age’ is greater than 20207.5, the model makes a similar decision based on ’cholesterol’ as described in steps 3 and 4.
6) If ’ap-hi’ is greater than 129.0, the model directly predicts the class as 1.0.

Similarly, for Random forest, 20 tree classification rules were generated, and 10 tree rules were generated in the case of GBT. In summary, the GBT Classifier is superior in terms of performance across different feature sets. The ”Custom engineered Features” tend to perform better because along with the original, it has some specially engineered features, but feature selection based on importance and correlation can also lead to competitive results while simplifying the model. However, the difference between All original and custom-designed is not very significant so one can say the original features alone can perfectly display the relevance of heart disease predictions. Since we are generalizing the prediction of heart disease here, instead of predicting particular heart diseases, the model may not be able to efficiently capture the heterogeneous patterns of different heart disease data. Moreover, the features we have from weight to blood pressure, glucose, or smoking are very surface-level features and anomalies in them may be indications of other diseases relating to the kidneys, liver, lungs, etc. A better dataset comprising of some polarisation & depolarisation of heart muscles, the ECG wave reports, and its signal processing might result in better accuracy.

VI. Conclusion

In conclusion, this paper presents a compelling approach to addressing the pressing global health challenge of cardiovascular disease (CVD) through the integration of merging big data approaches and machine learning may be precise and useful for doctors without causing them undue worry. Through a variety of machine learning models, including Logistic Regression, Decision Tree Classifier, Gradient Boosted Tree (GBT) Classifier, SVM, Multilayer Perceptron, and Random Forest Classifier, the study aims to revolutionize CVD management. The implementation of IoT devices in healthcare offers the potential for early detection and effective management of CVD, enabling timely interventions and cost reductions, ultimately preserving lives.

However, it is important to note that while the proposed methodology is promising, it comes with its share of challenges, such as data security, privacy, and data volume management. Addressing these issues is essential to ensure the successful implementation of these technologies in healthcare.

Though we have not discussed in detail the integration of its devices, it is essential for data collection as obvious as it is. Due to the limitation of features in the dataset, we might have missed the opportunity to achieve better specificity in our model. It can be solved with the government making better datasets available for research purposes. Due to limited facilities, we couldn't demonstrate parallel processing of spark, but spark is able of parallelising the process and thus makes the application faster. In summary, this multifaceted approach, combining IoT devices, big data analytics, and advanced data processing platforms, has the potential to significantly reduce the burden of CVD and enhance the quality of life for individuals. By harnessing real-time data monitoring and predictive analytics, healthcare organizations can usher in a new era of proactive, patient-centered care that empowers individuals to take control of their cardiovascular health. Continuous innovation and collaboration within the healthcare industry will be instrumental in realizing this vision and improving the well-being of patients worldwide.

Acknowledgment

This paper would not have been possible without the support, guidance, and contributions of many individuals and institutions. We extend our heartfelt thanks to our course professor,*Dr. Nabajyoti Mazumdar*, and mentors, *Mr. Mohammad Shahnawaz* and *Mr. Om Prakash Khushwaha*, for their unwavering support and expert guidance throughout this project.

We're also grateful to the Indian Institute of Information Technology, Allahabad, for providing essential resources and facilities. We appreciate the collective efforts of all who contributed to this research, directly or indirectly, as they played a crucial role in our progress.

References

[1] World Heart Report 2023: Confronting the World's Number One Killer. Geneva, Switzerland. World Heart Federation. 2023.

[2] Big data and IoT solution for patient behavior monitoring Kwok Tai Chui,Ryan Wen Liu, Miltiadis D. LytrasORCID Icon & Mingbo Zhao Pages 940-949 — Received 25 Feb 2018, Accepted 10 Feb 2019, Published online: 27 Feb 2019

[3] A. Shankhdhar, "Visualization and Prediction of Heart Disease using Big Data Analytics," 2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART), Moradabad, India, 2022, pp. 39-43, doi: 10.1109/SMART55829.2022.10046877.

[4] Abderrahmane ED-DAOUDY, Khalil Maalmi, Aziza El ouaazizi et al. A scalable and real-time system for disease prediction using big data processing, 09 March 2023, PREPRINT (Version 3) available at Research Square [https://doi.org/10.21203/rs.3.rs-1567163/v3]

[5] J. Sun and C. K. Reddy, "Big data analytics for healthcare," in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 2013, pp. 1525–1525.

[6] J. Archenaa and E. M. Anita, "Interactive big data management in healthcare using spark," in Proceedings of the 3rd International Symposium on Big Data and Cloud Computing Challenges (ISBCC–16'). Springer, 2016, pp. 265–272.

[7] L. R. Nair, S. D. Shetty, and S. D. Shetty, "Applying spark-based machine learning model on streaming big data for health status prediction," Computers , Electrical Engineering, vol. 65, pp. 393–399, 2018.

[8] S. Alotaibi, R. Mehmood, I. Katib, O. Rana, and A. Albeshri, "Sehaa: A big data analytics tool for healthcare symptoms and diseases detection using twitter, apache spark, and machine learning," Applied Sciences, vol. 10, no. 4, p. 1398, 2020.

[9] A. Ed-Daoudy and K. Maalmi, "Real-time machine learning for early detection of heart disease using big data approach," in 2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS). IEEE, 2019, pp. 1–5.

[10] S. Ashraf, Y. M. Afify and R. Ismail, "Big Data for Real-Time Processing on Streaming Data: State-of-the-art and Future Challenges," 2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), Maldives, Maldives, 2022, pp. 1-8, doi: 10.1109/ICECCME55909.2022.9987770.

[11] Institute for Health Metrics and Evaluation (IHME). GBD Compare Data Visualization. Seattle, WA: IHME, University of Washington, 2020. Available from http://vizhub.healthdata.org/gbd-compare (18 March 2023).

[12] "Spark RDD et Traitement par Lots" IEEE Xplore, 2023. [Online]. Available: https://liliasfaxi.github.io/Atelier-Spark/p4-batch/. Copyright © 2019 - 2020 Lilia Sfaxi Made with Material for MkDocs