Project Report

On

Predictive Health Risk Model Using Lifestyle Data

Submitted in partial fulfilment for the award of

Post Graduate Diploma in Big Data Analytics (PGDBDA)

From Know IT(Pune)

Guided by:

Mrs. Trupti Joshi
&
Mr. Prasad Deshmukh

Submitted By:

Amey Nate	240343025003
Awaiz Bagwan	240343025006
Kshitij Sontakke	240343025023
Muhammad Shaikh	240343025029
Rajat Tikle	240343025057

CERTIFICATE

TO WHOMSOEVER IT MAY CONCERN

This is to certify that

Amey Nate	240343025003
Awaiz Bagwan	240343025006
Kshitij Sontakke	240343025023
Muhammad Shaikh	240343025029
Rajat Tikle	240343025057

Have successfully completed their project on

Predictive Health Risk Model Using Lifestyle Data

Under the guidance of Mrs. Trupti Joshi and Prasad Deshmukh Sir

ACKNOWLEDGEMENT

The project Predictive Health Risk Model Using Lifestyle Data was a great learning experience for us and we are submitting this work to CDAC Know IT (Pune).

We all are very glad to mention the name Mrs. Trupti Joshi and Mr. Prasad Deshmukh for their valuable guidance to work on this project. His guidance and support helped us to overcome various obstacles and intricacies during the course of project work.

We are highly grateful to Mr. Vaibhav Inamdar Manager (Know-IT), C-DAC, for his guidance and support whenever necessary while doing this course Post Graduate Diploma in Big Data Analytics (PG-DBDA) through C-DAC ACTS, Pune.

Our most heartfelt thanks go to Mr. Shrinivas Jadhav (Vice President, Know-IT) who gave all the required support and kind coordination to provide all the necessities like required hardware, internet facility and extra Lab hours to complete the project and throughout the course up to the last day here in C-DAC Know-IT, Pune.

TABLE OF CONTENTS

ABSTRACT

The Predictive Health Risk Model Using Lifestyle Data aims to predict the likelihood of various health conditions based on lifestyle data. By leveraging machine learning techniques and PySpark, this project builds models to classify individuals' risk of developing diseases such as asthma, heart disease, arthritis, and diabetes. The project uses logistic regression, random forest, linear SVM, and gradient-boosted trees to analyze the data and make predictions. The final model's performance is evaluated through accuracy, precision, recall, and F1 score metrics. By accurately predicting health risks, our model aims to provide valuable insights that can inform preventative healthcare measures, enhance patient outcomes, and contribute to public health management. The developed solution will be instrumental in identifying at-risk individuals, tailoring interventions, and ultimately improving health outcomes.

1. INTRODUCTION

In today's data-driven world, the ability to predict health risks based on individual lifestyle factors is becoming increasingly important. With the rise of wearable technology, mobile health apps, and electronic health records, vast amounts of lifestyle data are now available, offering valuable insights into personal health and well-being. Predictive health risk modeling involves analyzing this data to identify patterns and correlations that can forecast potential health issues before they arise. Machine learning techniques have proven to be highly effective in this domain, enabling the development of models that can accurately assess health risks based on lifestyle choices.

The modern healthcare landscape benefits significantly from predictive analytics, especially when it comes to assessing risk based on lifestyle factors. This project focuses on developing a predictive health risk model using lifestyle data, aiming to forecast the probability of various health conditions. By integrating data from diverse health-related questions and features, the model helps identify individuals at risk of diseases such as asthma, arthritis, heart disease, and diabetes.

Objectives:

- To create a predictive model that can classify health risks based on lifestyle data.
- To evaluate and compare the performance of different machine learning algorithms.

Scope:

- The project includes data preprocessing, model training, and evaluation using PySpark.
- The final model's predictions are used to provide insights into individuals' health risks.

2. Dataset Collection and Features

Data Sources

For our project, we utilized the 2022 annual CDC survey data, which includes responses from over 400,000 adults regarding their health status. This dataset is part of the Behavioral Risk Factor Surveillance System (BRFSS), conducted by the CDC. Heart disease is a leading cause of death in the U.S., affecting various demographic groups including African Americans, American Indians, Alaska Natives, and whites. The dataset includes crucial indicators related to heart disease, such as high blood pressure, high cholesterol, smoking, diabetes status, obesity (high BMI), physical inactivity, and excessive alcohol consumption.

The dataset used for this project is obtained from a CSV file stored on Google Drive. It includes various features related to an individual's lifestyle and health conditions. Key features include:

- SleepHours: Average hours of sleep per night.
- WeightInKilograms: Current weight of the individual.
- HeightInCentimeters: Height of the individual in centimeters.
- BMI: Body Mass Index, derived from height and weight.
- SexIndexed: Gender of the individual
- PhysicalActivitiesIndexed: Whether the individual engages in physical exercise weekly
- HadAnginaIndexed: Experience of angina
- HadStrokeIndexed: History of stroke
- HadSkinCancerIndexed: History of skin cancer
- HadCOPDIndexed: Chronic obstructive pulmonary disease status
- HadDepressiveDisorderIndexed: History of depressive disorder
- HadKidneyDiseaseIndexed: History of kidney disease

- DeafOrHardOfHearingIndexed: Hearing difficulties
- BlindOrVisionDifficultyIndexed: Vision difficulties
- DifficultyConcentratingIndexed: Difficulty in concentrating
- DifficultyWalkingIndexed: Difficulty in walking
- DifficultyDressingBathingIndexed: Difficulty in dressing or bathing
- DifficultyErrandsIndexed: Difficulty in managing tasks
- ChestScanIndexed: History of chest scan
- AlcoholDrinkersIndexed: Alcohol consumption status
- HIVTestingIndexed: HIV testing status
- FluVaxLast12Indexed: Flu vaccination status in the last 12 months
- PneumoVaxEverIndexed: Pneumonia vaccination status
- GeneralHealthIndex: Self-rated general health
- LastCheckupTimeIndex: Time since last health check-up
- SmokerStatusIndex: Smoking status
- AgeCategory: Age category

Data Preprocessing:

- Data is loaded from a CSV file and cleaned.
- Features are assembled into a single vector column for model input.
- Stratified sampling is used to split the dataset into training and testing sets.

3. SYSTEM REQUIREMENTS

3.1 Software Requirements

- Python: The programming language used.
- PySpark: For distributed data processing and machine learning.
- Google Colab: For running the code in a cloud environment.
- Google Drive: For storing and accessing datasets and models.

3.2 Hardware Requirements

- CPU: A standard processor is sufficient as PySpark performs distributed computing.
- RAM: At least 8 GB recommended for smooth data processing.
- Storage: Sufficient storage for datasets and models on Google Drive**.**

4. FUNCTIONAL REQUIREMENTS

Python 3:

Versatility: Python is a versatile, high-level programming language ideal for developing diverse applications, including predictive health models. Its flexibility supports various tasks from web and desktop application development to advanced data analysis.

Machine Learning Support: Python is a preferred language for machine learning due to its rich ecosystem of libraries such as scikit-learn, TensorFlow, and PySpark. These libraries facilitate the implementation of complex algorithms and data processing tasks required for building and evaluating health risk prediction models.

Ease of Development: Python's user-friendly syntax and high-level abstractions streamline the development process, allowing for efficient construction and refinement of the predictive health risk model without getting bogged down by lower-level programming details.

Comprehensive Libraries and Frameworks: Python offers a robust set of libraries and frameworks that are essential for statistical analysis, data manipulation, and machine learning. Libraries such as PySpark, pandas, and scikit-learn provide the necessary tools for processing lifestyle data and generating accurate predictions of health risks.

Evolution and Adaptability: Python's evolution incorporates advanced features and best practices from various programming paradigms. This evolution ensures that Python remains a powerful and adaptable tool for modern data science and predictive analytics tasks, including those involved in health risk modeling.

5. ARCHITECTURE:

The architecture of the Predictive Health Risk Model involves the following components:

1. **Data Input**: CSV files stored in Google Drive are loaded into PySpark DataFrames.
2. **Data Preprocessing**: Features are engineered, and data is transformed using PySpark.
3. **Model Training**: Machine learning models are trained on the processed data.
4. **Model Evaluation**: Models are evaluated using various metrics to determine their effectiveness.
5. **Prediction**: Trained models are used to make predictions on user input data.
6. **Model Saving**: Trained models are saved to Google Drive for future use.

6. PySpark for Predictive Health Risk Modeling

PySpark, the Python API for Apache Spark, facilitates large-scale data processing and analytics, making it essential for predictive health risk modeling. Here's how PySpark is utilized in the context of your model:

1. **Scalable Data Processing**
2. **Overview:**
    - **Distributed Computing**: PySpark leverages distributed computing to manage extensive datasets by splitting tasks across multiple machines. This scalability is crucial for handling the large volumes of lifestyle data involved in health risk modeling.
3. **Application:**
    - **Health Data Analysis**: Health risk models often require processing substantial datasets, which include records of physical activity, dietary habits, and medical history. PySpark's distributed processing capabilities enable efficient analysis and model training without being constrained by the limitations of a single machine's resources.
4. **ETL (Extract, Transform, Load) Processes**
5. **Overview:**
    - **ETL Framework**: PySpark supports comprehensive ETL processes to prepare data for analysis. It provides tools to extract data from various sources, transform it into a clean and structured format, and load it into a data warehouse or analytics platform.
6. **Application:**
    - **Data Extraction**: Retrieve data from multiple sources such as health records, wearable devices, and surveys.
    - **Data Transformation**: Clean and preprocess data, including handling missing values, normalizing numerical features, and encoding categorical variables. For example, this includes converting raw lifestyle data into a consistent format and aggregating information from different sources.
    - **Data Loading**: Load the preprocessed data into a format suitable for analysis or machine learning, such as data frames or tables.

7. **Real-Time Data Processing**
8. **Overview:**
    - **Streaming Capabilities**: PySpark supports real-time data processing through Spark Streaming, allowing data to be processed in micro-batches or real-time streams.
9. **Application:**
    - **Dynamic Risk Assessment**: Continuously monitor and analyze incoming lifestyle data to update risk assessments in real-time. For instance, as new health metrics are recorded, the model can provide immediate insights or alerts about potential health risks.
10. **Data Transformation**
11. **Overview:**
    - **Data Manipulation**: PySpark offers powerful tools for data transformation, such as filtering, aggregation, and feature engineering, which are vital for preparing data for modeling.
12. **Application:**
    - **Feature Engineering**: Derive meaningful features from raw data, such as creating new variables that capture interactions between different lifestyle factors (e.g., a wellness score based on physical activity and diet).
    - **Aggregation**: Summarize and aggregate data to identify trends and patterns, such as aggregating monthly physical activity data to analyze long-term lifestyle trends.
13. **Integration with Big Data Tools**
14. **Overview:**
    - **Compatibility**: PySpark integrates seamlessly with big data tools like Hadoop (HDFS) and Hive, enhancing its ability to manage and analyze large datasets.
15. **Application:**
    - **Data Storage**: Store large volumes of lifestyle data in Hadoop's distributed file system (HDFS) and process it with PySpark.
    - **Querying and Analysis**: Utilize Hive for querying and managing large datasets, and perform complex analyses using PySpark's data processing capabilities.

7. MLlib for Machine Learning

Overview:

- **Machine Learning Library**: MLlib is Spark's scalable machine learning library, providing efficient algorithms and tools for building and deploying models on large datasets.

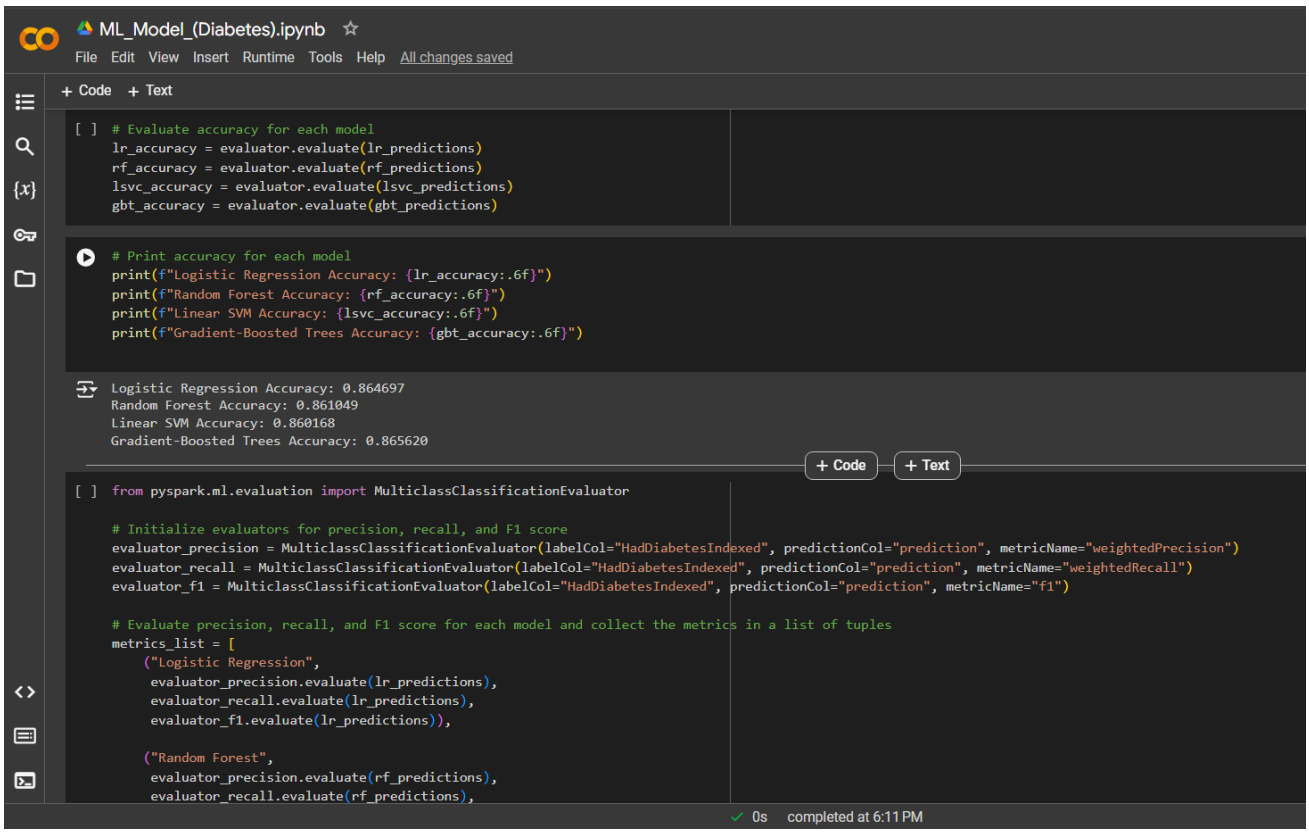
Application:

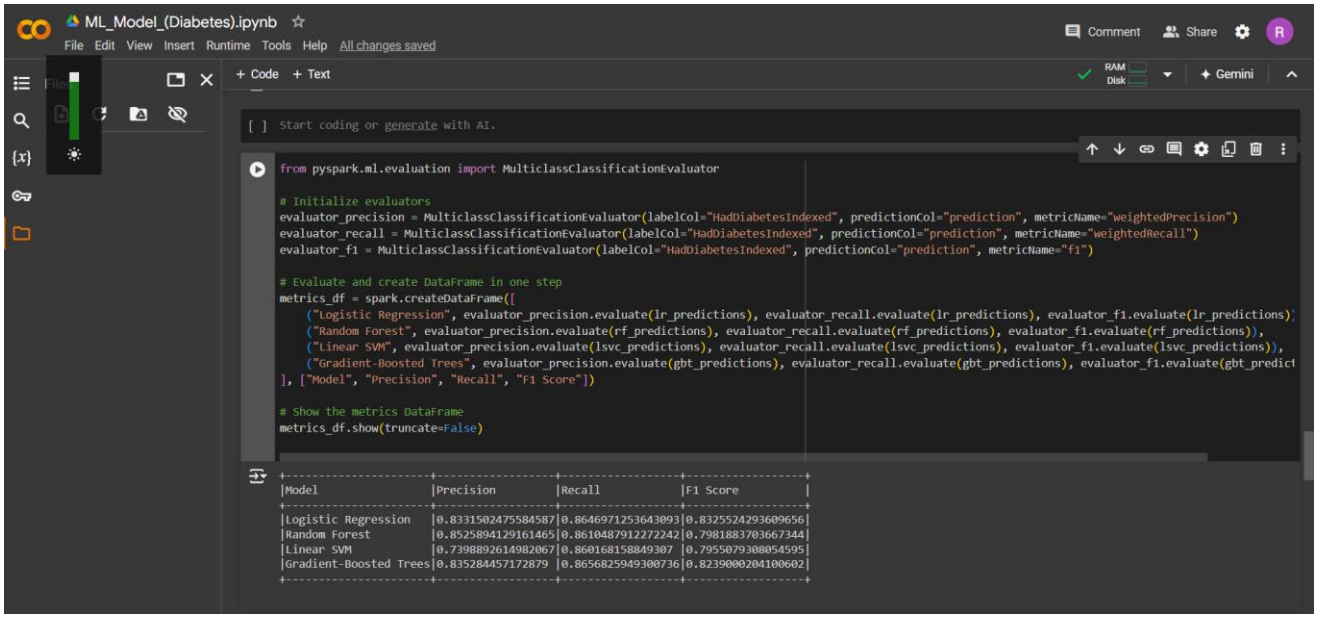
- **Model Building**: Use MLlib's algorithms to train models for assessing health risks based on lifestyle data. Examples include logistic regression for binary classification and decision trees for complex risk assessments.
- **Model Evaluation**: Employ MLlib's tools for evaluating model performance and tuning hyperparameters to enhance predictive accuracy.
- **Scalability**: MLlib leverages Spark's distributed computing to handle large-scale data and complex models efficiently.

Example Workflow:

1. **Data Ingestion**: Load lifestyle data into Spark DataFrames using PySpark.
2. **Data Cleaning and Transformation**: Perform data cleaning and preprocessing, such as managing missing values and scaling features.
3. **Feature Engineering**: Develop features that capture significant health indicators from lifestyle data.
4. **Model Training**: Train predictive models using MLlib on the processed data, such as logistic regression for risk prediction.
5. **Real-Time Processing**: Implement real-time data processing to continuously update risk assessments.
6. **Integration**: Combine with other big data tools for efficient data management and analysis.

FINAL OUTPUT :





7. CONCLUSION AND FUTURE SCOPE:

Conclusion: The Predictive Health Risk Model Using Lifestyle Data successfully demonstrates the ability to predict health risks based on various lifestyle factors. The models trained using PySpark exhibit varying levels of accuracy, with Gradient-Boosted Trees performing the best in this case. The model's predictions can be used to identify individuals at risk of certain health conditions and provide personalized recommendations.

Future Scope:

- **Model Improvement**: Exploring additional features and hyperparameter tuning to improve model performance.
- **Integration**: Integrating the model with a user interface for real-time predictions.
- **Expanded Data**: Using a larger and more diverse dataset to enhance model accuracy and generalizability.
- **Deployment**: Deploying the model in a production environment for broader usage.

8. References

[1] Apache Spark Documentation. Retrieved from https://spark.apache.org/docs/latest/

[2] PySpark Documentation. Retrieved from https://spark.apache.org/docs/latest

[3] Kaggle - https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

[4] Github - https://github.com/mdsk01/Predictive-Health-Risk-Model-Using-Lifestyle-Data