

## 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), CDAC, for his guidance and support whenever necessary while doing this course Post Graduate Diploma in Big Data Analytics (PGDBDA) through CDAC 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 CDAC Know IT, Pune.



# TABLE OF CONTENTS

ABSTRACT

1. INTRODUCTION
2. DATA COLLECTION AND FEATURES
3. SYSTEM REQUIREMENTS
   1. Software Requirements
   2. Hardware Requirements
4. FUNCTIONAL REQUIREMENTS
5. ARCHITECTURE
6. PYSPARK
7. CONCLUSION AND FUTURE SCOPE
8. REFERENCES



# ABSTRACT

Predictive health risk modeling is a critical task in the field of healthcare, with applications ranging from early disease detection to personalized health recommendations. The goal of this project is to predict individual health risks using lifestyle data through machine learning techniques. The dataset includes a variety of lifestyle factors such as physical activity, dietary habits, sleep patterns, and demographic information. The methodology involves data preprocessing, model selection, training, and evaluation, including the implementation of advanced techniques such as feature engineering and ensemble learning. 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.



# 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.

This report explores the application of machine learning algorithms, including advanced techniques such as feature engineering and ensemble learning, to predict health risks using lifestyle data. The dataset used in this study includes various lifestyle factors such as physical activity, diet, sleep patterns, and demographic information. By leveraging this data, our goal is to create a model that can provide personalized health risk assessments, thereby aiding in early intervention, preventative care, and the overall improvement of public health outcomes.

Objectives:

The primary objectives of this report are:

* To investigate the effectiveness of machine learning algorithms in predicting health risks based on lifestyle data.
* To explore the application of advanced techniques such as feature engineering and ensemble learning in improving the accuracy of health risk predictions.
* To evaluate the performance of different machine learning models and strategies in predicting individual health risks using lifestyle data.



## 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.

**Dataset Overview**

The dataset originally contains nearly 300 variables. For our analysis, we reduced this to 40 relevant variables that are most predictive of heart disease. The data is provided in two versions: one with missing values (NaNs) and one with those values imputed.

**Key Indicators of Heart Disease**

1. High Blood Pressure: Elevated blood pressure readings.

2. High Cholesterol: Elevated cholesterol levels.

3. Smoking: Smoking status.

4. Diabetes Status: Presence of diabetes.

5. Obesity: High Body Mass Index (BMI).

6. Physical Inactivity: Levels of physical activity.

7. Alcohol Consumption: Amount of alcohol intake.

**Data Treatment**

The dataset is part of the BRFSS, a comprehensive health survey system. Established in 1984, BRFSS now covers all 50 states, the District of Columbia, and three U.S. territories, making it the largest continuously conducted health survey system globally. The most recent dataset from 2023 was selected for this analysis, focusing on variables directly influencing heart disease. We prepared the data by selecting key variables and handling missing values, resulting in two versions of the dataset: one with NaNs and one without.

**Applications**

The dataset is suitable for various analytical tasks, including exploratory data analysis (ETL) and machine learning applications. Key machine learning models such as logistic regression, support vector machines (SVM), and random forests, Gradient Boost can be applied to predict heart disease. The variable "HadHeartAttack" is treated as binary (Yes/No), indicating whether a respondent had heart disease.



**Key Features/Attributes**

1. High Blood Pressure:

- Attributes: Elevated systolic or diastolic pressure readings.

- Significance: Strong indicator of cardiovascular risk.

2. High Cholesterol:

- Attributes: Elevated levels of total cholesterol, LDL, or HDL.

- Significance: Contributes to atherosclerosis and heart disease.

3. Smoking:

- Attributes: Current smoker status, smoking frequency.

- Significance: Major risk factor for heart disease and stroke.

4. Diabetes Status:

- Attributes: Presence of diabetes or prediabetes.

- Significance: Increases risk of cardiovascular complications.

5. Obesity:

- Attributes: High BMI values.

- Significance: Linked to increased risk of heart disease and related conditions.

6. Physical Inactivity:

- Attributes: Levels of exercise or physical activity.

- Significance: Low physical activity contributes to cardiovascular risk.

7. Alcohol Consumption:

- Attributes: Frequency and quantity of alcohol intake.

- Significance: Excessive alcohol use can exacerbate cardiovascular issues.



# SYSTEM REQUIREMENTS

### Hardware Requirements

Computer: A computer with sufficient processing power and memory to run data processing and analysis tasks. A modern multicore processor and at least 8 GB of RAM are recommended.

Storage: Adequate storage space to store the generated dataset and any additional datasets if required. An SSD (Solid State Drive) is recommended for faster data access.

Internet Connection: A stable internet connection for downloading and installing software packages and libraries, as well as for any online resources needed during the project.

### Software Requirements

1. Operating System: Windows 10 or higher
2. Python: The project heavily relies on Python for data generation, analysis, and machine learning. Ensure Python is installed on your system.



1. Python Libraries: Install the following Python libraries and dependencies using package managers like pip or conda.

[EDIT / ADD LIBs]



# FUNCTIONAL REQUIREMENTS

**Python 3:**

- Versatile Programming Language: Python is a general-purpose, high-level programming language known for its versatility in developing various types of applications, including web and desktop applications.

- Machine Learning Integration: Python is widely used for implementing machine learning algorithms and data analysis tasks due to its robust libraries such as scikit-learn, TensorFlow, and pandas.

- Ease of Use: Python simplifies the development process by handling many common programming tasks, allowing developers to focus on building and refining the predictive health risk model.

- Extensive Libraries and Frameworks: Python’s extensive ecosystem includes libraries specifically designed for statistical analysis, data visualization, and machine learning, which are essential for analyzing lifestyle data and predicting health risks.

- Language Evolution: Python incorporates features and best practices from a variety of programming paradigms and languages, making it a powerful tool for modern data science and predictive modeling tasks.



# ARCHITECTURE:

Fig: Add fig here



# PySpark for Predictive Health Risk Modeling

PySpark is the Python API for Apache Spark, a powerful open-source distributed computing framework designed to handle large-scale data processing and analytics. Here’s how PySpark can be leveraged in predictive health risk modeling using lifestyle data:

**1. Scalable Data Processing**

Overview:

- Distributed Computing: PySpark enables the processing of large datasets by distributing tasks across a cluster of machines. This distributed approach allows for the handling of massive amounts of lifestyle data, which is crucial for developing robust predictive models.

Application:

- Health Data Analysis: In health risk modeling, datasets can be extensive, including numerous records of individual lifestyle factors like physical activity, diet, and medical history. PySpark efficiently processes these large datasets, enabling quick analysis and model training without being limited by the constraints of a single machine’s memory or processing power.

**2. ETL (Extract, Transform, Load) Processes**

Overview:

- ETL Framework: PySpark facilitates comprehensive ETL processes, essential for preparing 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.

Application:

- Data Extraction: Extract data from diverse sources such as health records, wearable device logs, and surveys.

- Data Transformation: Clean and preprocess the data, including handling missing values, normalizing numerical features, and encoding categorical variables. For instance, converting lifestyle data into consistent formats or aggregating information from multiple sources.

- Data Loading: Load the transformed data into a format suitable for analysis or machine learning modeling, such as data frames or tables.

**3. Real-Time Data Processing**

Overview:

- Streaming Capabilities: PySpark supports real-time data processing through Spark Streaming, which processes data in micro-batches or real-time streams. This feature allows for continuous data analysis and updating of models as new data arrives.

Application:

- Dynamic Risk Assessment: Continuously monitor and analyze incoming lifestyle data, enabling real-time updates to risk assessments. For example, as new health metrics are recorded, the model can adapt and provide immediate insights or alerts about potential health risks.



**4. Data Transformation**

Overview:

- Data Manipulation: PySpark provides powerful tools for data transformation, including filtering, aggregation, and feature engineering. These tools are critical for preparing data for modeling and analysis.

Application:

- Feature Engineering: Create meaningful features from raw data, such as deriving new variables that capture interactions between lifestyle factors (e.g., combining physical activity and dietary data to create a wellness score).

- Aggregation: Summarize and aggregate data to identify trends and patterns. For example, aggregating monthly physical activity data to understand long-term lifestyle patterns.

5. Integration with Big Data Tools

Overview:

- Compatibility: PySpark integrates seamlessly with other big data tools like Hadoop (HDFS), Hive, and various data storage and processing systems. This integration enhances its ability to handle and analyze large datasets.

Application:

- Data Storage: Store large volumes of lifestyle data in Hadoop's distributed file system (HDFS) and access it using PySpark for processing.

- Querying and Analysis: Use Hive for querying and managing large datasets, and perform complex analysis with PySpark’s data processing capabilities.

6. MLlib for Machine Learning

Overview:

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

Application:

- Model Building: Utilize MLlib’s algorithms to train predictive models for assessing health risks based on lifestyle data. Common algorithms include logistic regression, decision trees, and random forests.



- Model Evaluation: Leverage PySpark’s tools for model evaluation and hyperparameter tuning to improve predictive accuracy.

- Scalability: MLlib handles large-scale data and complex models by leveraging Spark’s distributed computing power, ensuring that even large and intricate models can be trained efficiently.

**Example Workflow:**

1. Data Ingestion: Use PySpark to ingest lifestyle data from multiple sources.

2. Data Cleaning and Transformation: Perform data cleaning and transformation tasks, such as dealing with missing values and feature scaling.

3. Feature Engineering: Create features that capture significant health indicators.

4. Model Training: Train predictive models using MLlib on the processed data.

5. Real-Time Processing: Implement real-time data processing to continuously update predictions and risk assessments.

6. Integration: Integrate with other big data tools for efficient data management and analysis.

## MLlib

MLlib is Apache Spark's machine learning library designed to handle large-scale data processing and provide scalable algorithms for building and deploying machine learning models. Here’s how MLlib can be applied to predictive health risk modeling using lifestyle data:

**1. Scalable Algorithms for Large Datasets**

Overview:

- Scalability: MLlib is engineered to work efficiently with large datasets by leveraging Spark's distributed computing framework. This is crucial for health risk modeling, where datasets often encompass extensive records of individual lifestyle factors and health metrics.

Application:

- Handling Big Data: In predictive health risk modeling, MLlib can process vast amounts of lifestyle data, such as physical activity logs, dietary information, and medical history. This scalability ensures that models can be trained and applied effectively across large populations, providing comprehensive risk assessments.



**2. Algorithms for Classification, Regression, and Clustering**

Overview:

- Diverse Algorithms: MLlib includes a variety of algorithms for different machine learning tasks, including classification, regression, and clustering. Each of these types of algorithms can be used to address specific aspects of health risk modeling.

Application:

- Classification: Use classification algorithms to predict binary outcomes, such as whether an individual has a high risk of heart disease based on their lifestyle data. Examples include logistic regression and decision trees.

- Regression: Apply regression techniques to model continuous health outcomes, such as predicting the likelihood of developing a health condition over time based on lifestyle factors. Examples include linear regression and ridge regression.

- Clustering: Implement clustering algorithms to identify patterns or groups within the data, such as grouping individuals with similar lifestyle profiles. Examples include K-means clustering.

**3. Distributed Computing for Big Data**

Overview:

- Distributed Processing: MLlib is designed to leverage Spark’s distributed computing capabilities, allowing it to handle large-scale data processing across multiple nodes in a cluster. This feature is essential for processing and analyzing extensive health datasets efficiently.

Application:

- Efficient Processing: When dealing with large volumes of lifestyle data, MLlib ensures that computational tasks are distributed across a cluster, reducing processing time and enabling the handling of complex models and extensive datasets.

**4. Predictive Modeling**

Overview:

- Model Building: MLlib provides a range of algorithms for building predictive models that can forecast health risks based on historical and current lifestyle data. These models can be used to predict future health outcomes and identify individuals at high risk.



Application:

- Risk Prediction: Develop models to predict the likelihood of health issues such as cardiovascular disease, diabetes, or obesity. For instance, logistic regression can be used to classify individuals into risk categories, while more complex algorithms like random forests can enhance prediction accuracy.

**5. Recommendation Systems**

Overview:

- Personalized Recommendations: MLlib can also be used to develop recommendation systems, which can provide personalized health recommendations based on an individual's lifestyle data and risk profile.

Application:

- Health Interventions: Create systems that recommend lifestyle changes or interventions to individuals based on their predicted risk levels. For example, if a model predicts a high risk of heart disease, the system could recommend changes in diet, exercise, or medical check-ups.

**6. Data Preprocessing**

Overview:

- Preprocessing Tools: MLlib includes tools for data preprocessing, which are essential for preparing raw lifestyle data for modeling. This includes techniques for handling missing values, normalizing data, and feature extraction.

Application:

- Data Preparation: Utilize MLlib’s preprocessing functions to clean and prepare lifestyle data. This involves filling in missing values, scaling numerical features, and encoding categorical variables to ensure that the data is suitable for machine learning models.

Example Workflow in MLlib

1. Data Ingestion: Load lifestyle data into a Spark DataFrame using PySpark.

2. Data Preprocessing: Clean and preprocess the data with MLlib’s tools, such as handling missing values and normalizing features.

3. Feature Engineering: Extract and transform features relevant to health risk prediction, such as creating composite indices from various lifestyle factors.



4. Model Training: Apply MLlib’s algorithms to train predictive models, such as logistic regression for binary classification of risk levels.

5. Model Evaluation: Assess model performance using metrics provided by MLlib, such as accuracy, precision, recall, and AUC-ROC.

6. Prediction and Recommendations: Use the trained models to predict health risks and provide personalized recommendations based on the results.

1. **Gradio:**



# CONCLUSION AND FUTURE SCOPE:



### Conclusion:



**Future Scope:**



## References

1. Understanding of a convolutional neural network, 10.1109/ICEngTechnol.2017.8308186, Saad Albawi, Tareq Abed Mohammed, Saad Al-Zawi,

[https://ieeexplore.ieee.org/document/8308186 ]

1. Image Captioning Using Inception V3 Transfer Learning Model, 10.1109/ICCES51350.2021.9489111, Sheshang Degadwala, Dhairya Vyas, Haimanti Biswas, Utsho Chakraborty, Sowrav Saha,

[https://ieeexplore.ieee.org/document/9489111]

1. TTransfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images, 10.29322/IJSRP.9.10.2019.p9420, Srikanth Tammina, [https:/[/www.rese](http://www.researchgate.net/publication/337105858_Transfer_learning_using_VGG-)a[rchgate.net/publication/337105858\_Transfer\_learning\_using\_VGG-](http://www.researchgate.net/publication/337105858_Transfer_learning_using_VGG-) 16\_with\_Deep\_Convolutional\_Neural\_Network\_for\_Classifying\_Images]
2. Kaggle

[<https://www.kaggle.com/datasets/puneet6060/intel-image-classification/code>]