**CLASSIFICATION OF DIFFERENT MACHINE LEARNING MODELS FOR OBESITY PREDICTION**

**ABSTRACT**

**Obesity presents a complex health problem with profound social implications.** **This study presents the classification of different machine learning based model for obesity prediction. Through data processing, research analysis and ML model application, the project provides an overview of obesity prognosis and prevention measures. The results show the accuracy of the system and highlight the importance of customization. In addition, Tableau visualizations provide additional information about user settings and content. The goal of the project is to contribute to the classification of obesity and promote individual health initiatives. The methodology emphasizes the synergy between data analysis, ML algorithms and visualization tools that promote personalized health recommendations.**

**Keywords— Obesity Classification, Machine Learning, PySpark, Tableau, Data Analysis, Personalized Healthcare, Preventive Measures.**

1. **INTRODUCTION**

Obesity, a complex and multifaceted health problem, has become a global epidemic and poses a major challenge to individuals and society. Obesity, with its negative effects on physical health, mental health and overall quality of life, has become a major public health problem. The World Health Organization (WHO) estimates that more than 650 million adults worldwide suffer from chronic diseases, and this number is growing and is a real concern. Obesity increases the risk of developing a variety of chronic diseases, including heart disease, type 2 diabetes, some cancers, and cardiovascular disease [1]. It also places a huge economic burden on the healthcare system, increasing healthcare costs and reducing productivity. Accurate and timely classification of obesity levels is essential for effective interventions, personal health care strategies, and public health planning. However, current approaches to obesity classification based solely on body mass index (BMI) and simple anthropometric measurements often lack the precision and clarity needed to capture the complexity of this phenomenon. This is where the power of Big data analytics (BDA) and Machine Learning (ML) techniques is introduced. [2]. The use of advanced research methods, especially ML, has emerged as a powerful way to improve predictive capabilities in the field of individual obesity. BDA, with its ability to identify causal factors individually, will become an important tool in this complex environment.

Previous approaches to obesity classification have been largely based on traditional methods that underestimate the complexity of this multifaceted phenomenon. The commonly used body mass index (BMI) is a common index for classifying levels of obesity, but it does not capture changes in body composition, fat distribution, and other important factors [ 3 ]. Although BMI provides a primary screening tool, it lacks the precision and granularity needed to accurately classify obesity and its associated health risks. Other traditional methods based on anthropometric measurements such as waist circumference, weight, family history with overweight provide better insights but still have limitations in discriminatory power [4]. This experiment is aimed at developing a recommendation system using Random forest and logistic regression for classification [5]. This includes building a model for predicting obesity levels based on the attributes supplied in the dataset. The model will strengthen healthcare systems thereby enhancing user satisfaction and healthcare facilities.

The structure of this experiment is as follows: Section 2 reviews the current literature research using sequential and parallel tagging methods. Section 3 describes the dataset. Section 4 describes how to perform the experiments. Section 5 describes the experimental setup of the system. Section 6 presents the results obtained. Section 7 concludes the paper with the conclusions, future directions, and societal implications of this project.

**II. Related Research**

Many researches have been done on the use of machine learning and content-based techniques for obesity level prediction and classification systems. This section summarizes previous research to present the current literature. The summaries are ordered by parallel and sequential implementation.

The authors of [6] presented a new application of ML techniques in obesity assessment using human plasma lipidomics data. They used a large population cohort and used ML algorithms to analyze and extract significant patterns from complex lipidomic profiles. By training the model with lipidomic data and corresponding adiposity measures, they developed a predictive model that can accurately estimate levels of adiposity based on plasma lipidomic profiles. However, further validation and external validation for different populations are needed to ensure the reliability and generalizability of the proposed approach.

Experiment in [7] on Data mining has become more important nowadays. The process of seeking knowledge, by collecting various types of data, allows us to discover previously hidden but important information. This means that data mining is essential to obtain useful secret information. The use of machine learning algorithms in data mining has proven to be successful in quickly generating relevant information. IMC is used to compare a small number of ML algorithms. A person is considered obese if their body mass index (BMI) is greater than 30. Depression, poor functioning, and disability are just some of the ways obesity can reduce life expectancy. In this paper, we test and evaluate different classification machine learning methods, including KNN, XGB, logistic regression, and DT, using obesity data.

In [8], Data mining has become more important nowadays. Through the process of information retrieval, which involves gathering different types of information, we can uncover previously hidden but important information. This makes data mining crucial for extracting useful hidden information. The use of machine learning algorithms in data mining has proven successful in quickly extracting relevant information. BMI is used to compare a small number of ML algorithms. If a person has a body mass index (BMI) of 30 or more, they are considered obese. Depression, poorer work performance and disability are just some of the ways obesity reduces quality of life. In this paper, we used obesity data to test and evaluate many different machine-learning methods for classification, including KNN, XGB, logistic regression, and DT.

Authors of [9] examined body mass index category and mortality risk in US adults: the effect of overweight and obesity on mortality. Using Cox proportional hazards regression, they estimated rates of progression for all causes of death and number of deaths, and relative to their normal weight counterparts, they estimated CVD-related mortality in obese and obese adults, stating that CVD mortality was greater than 20% . obese adults who are of normal weight. - adults with body weight.

**III DATASET**

The dataset used in the experiment focuses on the classification and analysis of obesity levels in selected countries populations [10]. Compiled from various sources, it provides a detailed picture of the factors influencing obesity in different population groups and geographic regions. The dataset includes information on obesity levels among people in Mexico, Peru and Colombia, covering a wide age range from 14 to 61 years. With a total of 17 attributes and 2,111 records, it provides a comprehensive picture of participants' eating habits, physical fitness and demographic characteristics. Below presents the composition of the dataset.

**Dataset Composition**

**Attributes Related to Eating Habits**

1. Frequent consumption of high-caloric food (FAVC)

2. Frequency of consumption of vegetables (FCVC)

3. Number of main meals (NCP)

4. Consumption of food between meals (CAEC)

5. Consumption of water daily (CH2O)

6. Consumption of alcohol (CALC)

**Attributes Related to Physical Condition**

7. Calories consumption monitoring (SCC)

8. Physical activity frequency (FAF)

9. Time using technology devices (TUE)

10. Transportation used (MTRANS)

**Demographic Attributes**

11. Gender

12. Age

13. Height

14. Weight

15. Family history with overweight

**Target Variable**

16. Obesity level (NObeyesdad)

The main goal of using this dataset is to develop robust machine learning models that can accurately predict obesity levels based on given characteristics. By analyzing the relationships between dietary habits, physical condition and demographic factors, the study aims to identify important factors influencing obesity and provide insight into preventive measures and interventions.

**IV. METHODOLOGY**

A systematic method to reach the goals of the project included several steps and techniques. First, a virtual environment was created and Ubuntu was installed on Microsoft Windows 11 using a virtual machine. This environment provided the necessary platform for continuous data analysis. Then, after the default Ubuntu installation of python3, the core components of the Ubuntu installation such as Java-JDK, PySpark and Hadoop were downloaded and installed. These tools were essential for effective data processing and analysis. Installation figures are presented in the appendix section.

The power of PySpark [11], a robust data processing framework, was used to analyze the data. PySpark enables efficient processing of large-scale data operations and ensures optimal performance during the analysis. Using a random forest and Logistic regression algorithm for classification was a key choice. This algorithm is known for its versatility in dealing with complex datasets and producing accurate results. Tableau was also used as a powerful data visualization software for extensive research and visualization of the material. Tableau made it possible to create a comprehensive visual representation that helped identify patterns, trends and correlations in the data.

**V. EXPERIMENTAL SECTION**

The outline presented below describes an experimental procedure to create and evaluate a obesity classification system with PySpark, specifically focusing on the processing of obesity dataset. Here is the code breakdown and its steps:

1. **Import Libraries:**

The code starts by importing the required libraries. These libraries contain PySpark components for data processing and machine learning, and Matplotlib for visualization.

1. **SparkSession Platform:**

SparkSession is the starting point for Spark programming using the DataFrame and Dataset APIs. It is used to create DataFrames, register DataFrames as arrays, run SQL over arrays, save arrays, and read parget files. In this step, a SparkSession named "Obesity\_Level\_Classification" is created.

1. **Load the dataset:**

The dataset "ObesityDataSet.csv" is loaded into a data frame named df. The dataset contains information about obesity levels and related features.

1. **Handling missing values:**

Missing values in the DataFrame df are handled by dropping rows with missing values using the df.na.drop() command.

1. **Select the first 600 rows :**

The command df.limit(600) selects only the first 600 rows of the data set.

1. **Target column distribution:**

The distribution of the target column "NObeyesdad" is followed with matplotlib's histogram function. This helps visualize the distribution of obesity levels in the dataset.

1. **Select relevant columns:**

Only relevant columns for analysis and modeling are selected from the DataFrame df. These columns contain attributes for demographics, dietary patterns, physical condition, and the target variable NObeyesdad.

1. **Encoding of categorical variables:**

Categorical variables for the selected columns are encoded using a StringIndexer. This step converts the categorical values ​​into numeric indices that can be processed by machine learning algorithms.

1. **Feature assembly:**

The selected features are assembled into a single feature using VectorAssembler. This step is necessary to feed the data to the machine learning algorithms.

1. **Scaling functions:**

Aggregate functions are scaled using StandardScaler. Scaling ensures that all features have the same scale, which is important for many machine learning algorithms.

1. **Encode the target variable:**

The target variable NObeyesdad is encoded using a StringIndexer. This step prepares the target variable for classification modeling.

1. **Select relevant features:**

ChiSqSelector selects the top 10 features. The chi-square selector selects features based on the strength of the relationship between the feature and the target variable.

1. **Split the data into training and test sets:**

The preprocessed data is split into training and test sets using a random distribution. This step prepares data for model training and evaluation.

1. **Initialize machine learning models:**

Random forest and logistic regression classifiers are initialized for model training.

1. **Training models:**

Initialized models are trained on training data.

1. **Evaluation models:**

Trained models are evaluated by MulticlassClassificationEvaluator. This step calculates the accuracy of the models based on the experimental data.

1. **Stop SparkSession:**

Finally, SparkSession is stopped to release resources used by Spark.

**VI. RESULT DISCUSSION**

A large workload, including data collection, preprocessing, analysis, and modeling with PySpark, Tableau, and Ubuntu Virtual Machines, was involved in the obesity study project. We examine the performance of random forest and logistic regression classification models in this section, intending to assess the system for classifying obesity levels. To assess the accuracy of classification, we have made use of a Multi-Class Classification Evaluation (MCE) tool model, a central part of PySpark's toolkit. MCE is a specific tool for evaluating the performance of multiclassification models and includes an extensive range of evaluation metrics adapted to this task. It is a useful tool for evaluating the quality of predictions based on machine learning models in multitudinous classification scenarios.

A dataset containing demographics, dietary patterns, physical fitness, and physical performance characteristics has been used in our analysis for levels of obesity. We have found a pattern of distribution in the materials, which shed light on factors that influence obesity in individuals, using an experimental data analysis. Figure 1 shows the distribution of NObeyesdad column using a histogram.

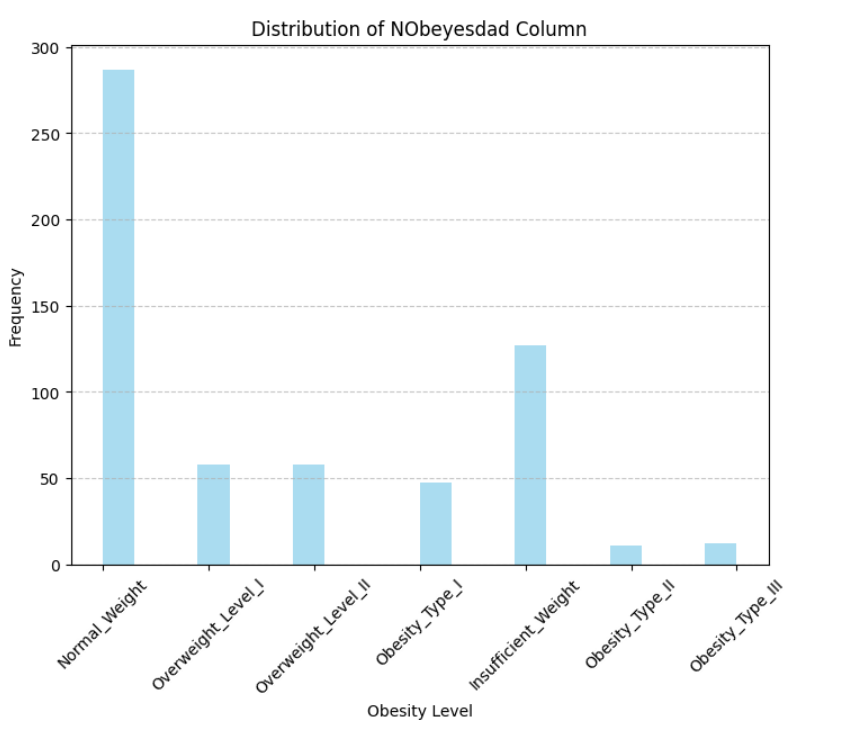


Figure 1 Distribution of Target Column

Figure 1 shows the distribution of the target feature. We see that individuals with normal weight has the highest occurrence but there are still cases of obesity within the distribution. After we evaluated the accuracy of the system using the obesity dataset, we found that the Random Forest model had a commendable overall accuracy of 76% and 68 % for Logistic regression model. Multiple classification estimator measured. These results highlight the effectiveness of classification models in accurately predicting obesity levels based on given characteristics.

In conclusion, the obesity classification system shows strong performance and provides valuable information about health and lifestyle factors influencing obesity. The integration of advanced modeling techniques and comprehensive data analysis will improve our understanding of obesity prognosis and provide potential avenues for preventive measures and interventions.

**VII. CONCLUSION & FUTURE WORKS**

In conclusion, our endeavour started a comprehensive study to exploit the power of PySpark, Tableau and Ubuntu virtual machine to build a robust classification system of obesity level. A detailed breakdown of the code snippet provided a comprehensive overview of how the system works, from initial data processing to model training and evaluation. In particular, the use of the Random Forest algorithm greatly improved the accuracy of the classification system and shed light on effective approaches to predict obesity based on the given characteristics. patterns of different populations and lifestyle factors affecting obesity. Accuracy evaluation using the Multiclass Classification Evaluator showed a commendable accuracy of 76% for the Random Forest model and 68% for the Logistic Regression model. These results highlight the effectiveness of the system to accurately predict obesity levels and highlight important predictors.

The step-by-step analysis highlighted the importance of preprocessing techniques, feature design and model evaluation in optimizing the classification model. This iterative approach introduced a complex interaction of data methods and machine learning algorithms, culminating in a highly accurate obesity prediction system.

In a broader context, this project illustrates the potential of data-driven approaches to complex health problems such as obesity. In the future, one could delve into the actual implementation of the classification system by integrating feedback mechanisms for its continuous improvement and performance improvement. In addition, the exploration of advanced machine learning techniques and the inclusion of additional data sources can further improve the predictive power of the system. Finally, the obesity classification system provides the basis for using data analysis to inform preventive measures and interventions to combat obesity and promote healthier lifestyles.

**IIX. SOCIAL IMPACT OF THIS EXPERIMENT**

The classification system for obesity prediction addresses today's obesity-related health problems, which have a profound social impact. This study provides valuable information on prevention measures and interventions that will improve public health outcomes through accurate prediction of obesity levels according to demographic and lifestyle characteristics. Furthermore, the implementation of such a classification system could have significant beneficial effects on human well-being to promote a healthy lifestyle. People can make informed choices about their diet, exercise, and overall health by knowing the factors that affect obesity.

Moreover, the classification system for obesity can influence public health policy and measures aimed at combating this epidemic. Policymakers may develop targeted strategies to combat obesity at individual and population levels by providing accurate predictions and information on risk factors associated with obesity. Overall, this project will have far-reaching implications for public health and well-being and will provide proactive. Approaches to the fight against obesity and promotion of good dietary habits. An obesity classification system, based on data-driven techniques, can promote positive social change and improve the quality of life of individuals and communities.

**REFERENCES**

[1] Safaei M., Sundararajan, E. A. Driss, M. Boulila, W. & Shapi'i, A., A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity. Computers in biology and medicine, 136, 104754, 2021.

[2] Chatterjee A., Gerdes M. W. & Martinez S. G., Identification of risk factors associated with obesity and overweight—a machine learning overview. Sensors, 20(9), 2734, 2020.

[3] Colmenarejo G., Machine learning models to predict childhood and adolescent obesity: a review. Nutrients, 12(8), 2466, 2020.

[4] Triantafyllidis A., Polychronidou E., Alexiadis A. Rocha C. L., Oliveira D. N., da Silva A. S., & Tzovaras D., Computerized decision support and machine learning applications for the prevention and treatment of childhood obesity: A systematic review of the literature. Artificial Intelligence in Medicine, 104, 101844, 2020.

1. Understand Random Forest Algorithms With Examples (Updated 2023) <https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/>
2. Gerl, M. J., Klose, C., Surma, M. A., Fernandez, C., Melander, O., Männistö, S., ... & Simons, K., Machine learning of human plasma lipidomes for obesity estimation in a large population cohort. PLoS biology, 17(10), 2019.
3. Ramya and K. Rohini, "Comparative evaluation of machine learning classifiers with Obesity dataset," 2021 International Conference on Computing Sciences (ICCS), Phagwara, India, 2021, pp. 38-41, doi: 10.1109/ICCS54944.2021.00016.
4. Ramya and K. Rohini, "Comparative evaluation of machine learning classifiers with Obesity dataset," 2021 International Conference on Computing Sciences (ICCS), Phagwara, India, 2021, pp. 38-41, doi: 10.1109/ICCS54944.2021.00016.
5. Borrell and Samuel, 2014 March L.N. Borrell, L. Samuel Body mass index categories and mortality risk in us adults: the effect of overweight and obesity on advancing death Am. J. Public Health, 104 (3) (2014) available at- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3953803/> Google Scholar.

[10] [Obesity or CVD risk (Classify/Regressor/Cluster) (kaggle.com)](https://www.kaggle.com/datasets/aravindpcoder/obesity-or-cvd-risk-classifyregressorcluster)

[11] Getting started with Spark. <https://www.datacamp.com/tutorial/pyspark-tutorial-getting-started-with-pyspark>

**APPENDIX**

Below is the installation process of the Pyspark/Hadoop Screenshots.

Figure 2. Complete installation of Ubuntu on the virtual machine environment.

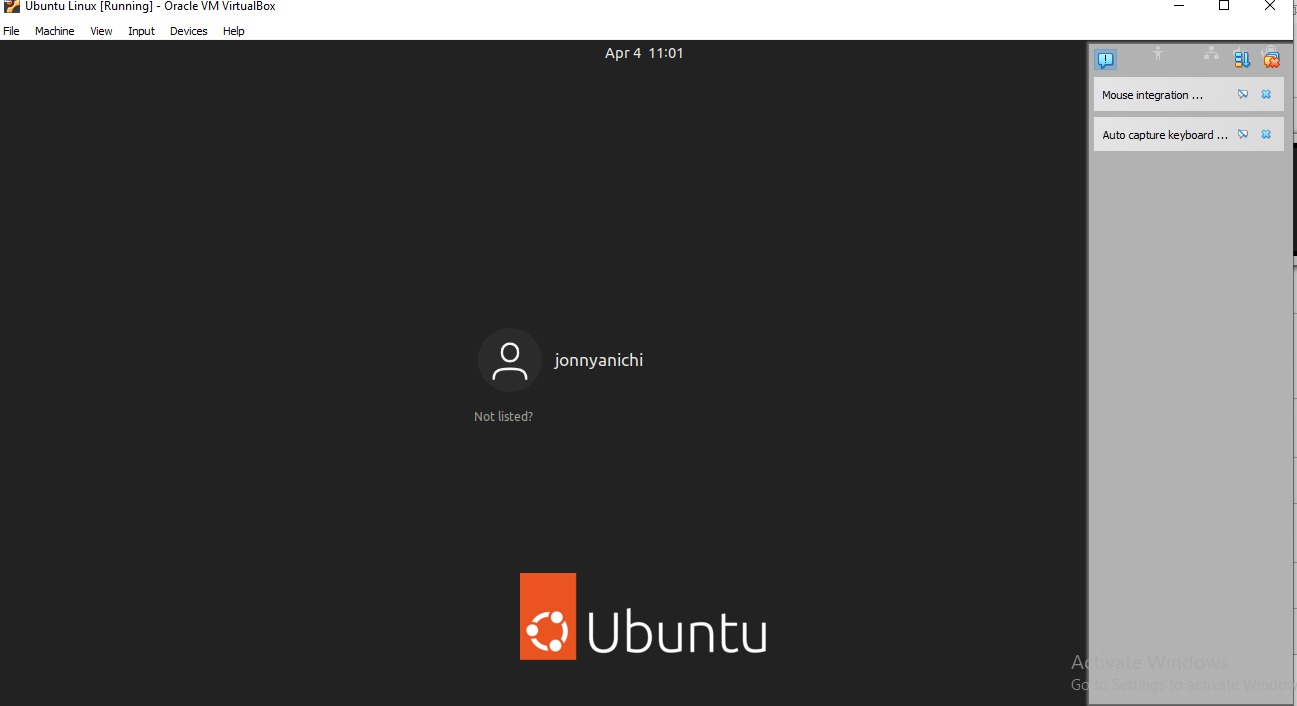


Figure 3. Installing Java jdk for Hadoop and spark support.

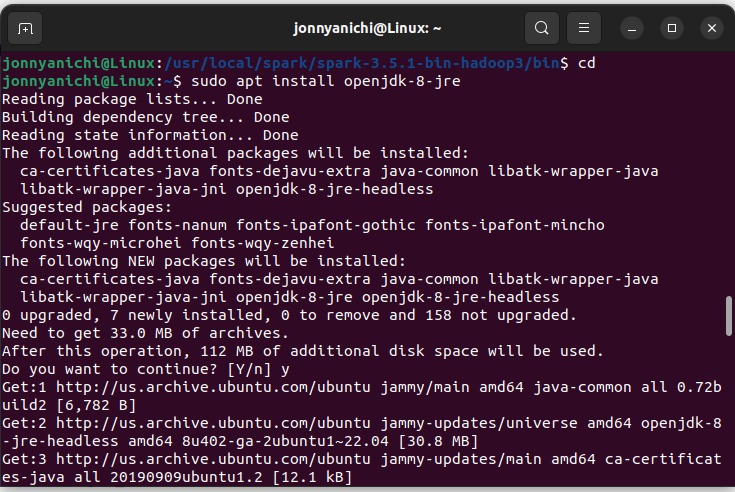
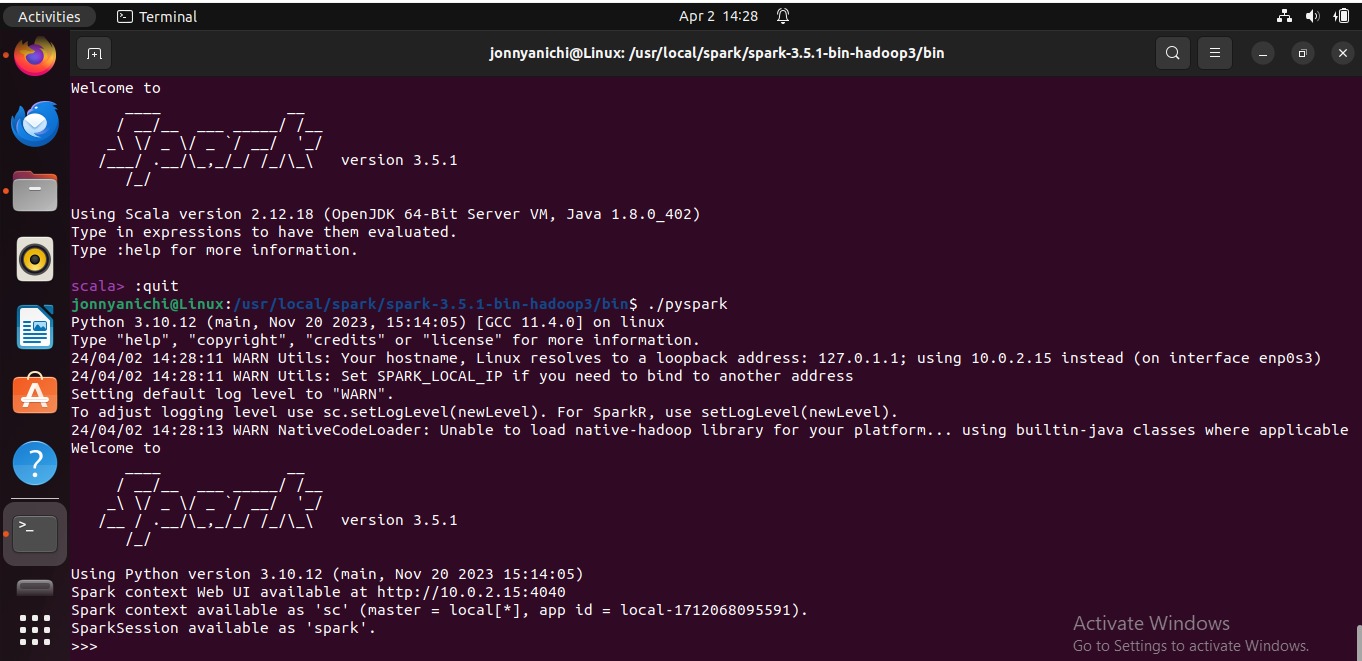


Figure 4. Hadoop and spark configuration.



**Code used for the experiment.**

**from** **pyspark.sql** **import** SparkSession

**from** **pyspark.ml.feature** **import** VectorAssembler, StringIndexer, StandardScaler

**from** **pyspark.ml.classification** **import** RandomForestClassifier, LogisticRegression

**from** **pyspark.ml.evaluation** **import** MulticlassClassificationEvaluator

**from** **pyspark.ml.feature** **import** ChiSqSelector**import** **matplotlib.pyplot** **as** **plt**

**from** **pyspark.ml** **import** Pipeline**from** **pyspark.ml.feature** **import** VectorAssembler, StringIndexer, StandardScaler, ChiSqSelector, OneHotEncoder

# Initialize SparkSession

spark = SparkSession.builder \

.appName("Obesity\_Level\_Classification") \

.getOrCreate()

# Load the Obesity dataset

df = spark.read.csv("ObesityDataSet.csv", header=True, inferSchema=True)# Handle missing values

df = df.na.drop()

# Select the first 600 rows

df = df.limit(**600**)

# Collect the data for the NObeyesdad column to the driver

nobeysdad\_data = df.select("NObeyesdad").toPandas()

# Plot the distribution using matplotlib

plt.figure(figsize=(**8**, **6**))

plt.hist(nobeysdad\_data["NObeyesdad"], bins=**20**, color='skyblue', alpha=**0.7**)

plt.title('Distribution of NObeyesdad Column')

plt.xlabel('Obesity Level')

plt.ylabel('Frequency')

plt.xticks(rotation=**45**) # Rotate x-axis labels for better readability if necessary

plt.grid(axis='y', linestyle='--', alpha=**0.7**)

plt.show()

# Select relevant columns

selected\_columns = ['Gender', 'Age', 'Height', 'Weight', 'family\_history\_with\_overweight',

'FAVC', 'FCVC', 'NCP', 'CAEC', 'CH2O', 'SCC', 'FAF', 'TUE', 'CALC', 'MTRANS', 'NObeyesdad']

df = df.select(selected\_columns)

# Encode categorical variables

categorical\_columns = ['Gender', 'family\_history\_with\_overweight', 'MTRANS']

indexers = [StringIndexer(inputCol=column, outputCol=column+"\_index", handleInvalid="keep") **for** column **in** categorical\_columns]

pipeline = Pipeline(stages=indexers)

df = pipeline.fit(df).transform(df)

# Encode remaining categorical columns

categorical\_columns\_remaining = ['FAVC', 'CAEC', 'SCC', 'CALC']

indexers\_remaining = [StringIndexer(inputCol=column, outputCol=column+"\_index", handleInvalid="keep") **for** column **in** categorical\_columns\_remaining]

pipeline\_remaining = Pipeline(stages=indexers\_remaining)

df = pipeline\_remaining.fit(df).transform(df)

# Assemble features

feature\_columns = ['Gender\_index', 'Age', 'Height', 'Weight', 'family\_history\_with\_overweight\_index',

'FAVC\_index', 'FCVC', 'NCP', 'CAEC\_index', 'CH2O', 'SCC\_index', 'FAF', 'TUE', 'CALC\_index', 'MTRANS\_index']

assembler = VectorAssembler(inputCols=feature\_columns, outputCol="features")

df = assembler.transform(df)

# Scale features

scaler = StandardScaler(inputCol="features", outputCol="scaled\_features")

scaler\_model = scaler.fit(df)

df = scaler\_model.transform(df)

# Update feature\_columns to include the encoded columns

feature\_columns += [column+"\_encoded" **for** column **in** categorical\_columns\_remaining]

# Encode target variable

target\_indexer = StringIndexer(inputCol="NObeyesdad", outputCol="NObeyesdad\_index", handleInvalid="keep")

df = target\_indexer.fit(df).transform(df)

# Select relevant features using ChiSqSelector

selector = ChiSqSelector(numTopFeatures=**10**, featuresCol="scaled\_features", outputCol="selected\_features",

labelCol="NObeyesdad\_index")

selector\_model = selector.fit(df)

df = selector\_model.transform(df)

# Split the data into training and testing sets

(training\_data, testing\_data) = df.randomSplit([**0.8**, **0.2**], seed=**42**)

# Initialize machine learning models

rf = RandomForestClassifier(labelCol="NObeyesdad\_index", featuresCol="selected\_features")

lr = LogisticRegression(labelCol="NObeyesdad\_index", featuresCol="selected\_features")

# Train models

rf\_model = rf.fit(training\_data)

lr\_model = lr.fit(training\_data)

# Evaluate models

evaluator = MulticlassClassificationEvaluator(labelCol="NObeyesdad\_index", metricName="accuracy")

rf\_predictions = rf\_model.transform(testing\_data)

rf\_accuracy = evaluator.evaluate(rf\_predictions)

lr\_predictions = lr\_model.transform(testing\_data)

lr\_accuracy = evaluator.evaluate(lr\_predictions)

**print**("Random Forest Accuracy:", rf\_accuracy)

**print**("Logistic Regression Accuracy:", lr\_accuracy)

# Stop SparkSession

spark.stop()

