

**MACHINE LEARNING AND DATA MINING**

**Machine Learning & Data Mining with Python and Azure Machine Learning Studio**



***STUDENT NAME: TOLUWALOPE OLADIPUPO OJUROYE***

***STUDENT ID: @00690747***

***COURSE: DATA SCIENCE***

***MODULE TITLE: MACHINE LEARNING AND DATA MINING***

***LEVEL: 7***

Table of Contents

[Task 1: Machine Learning Prediction for Employee Attrition 2](#_Toc160203646)

[1.1: Introduction 2](#_Toc160203647)

[1.2: Research Question 2](#_Toc160203648)

[1.3: Literature Review 2](#_Toc160203649)

[1.4: Dataset 2](#_Toc160203650)

[1.5: Explanation and Preparation of Dataset (Exploratory Data Analysis) 2](#_Toc160203651)

[1.6: Implementation in Python (K-Nearest Neighbours – KNN and Decision Tree) 8](#_Toc160203652)

[1.6.1: KNN Classification 10](#_Toc160203653)

[1.6.2: Decision Tree Classification 13](#_Toc160203654)

[1.7: Implementation in Azure Machine Learning Designer (Two-Class Neural Network and Two-Class Decision Forest) 17](#_Toc160203655)

[1.7.1: Two-Class Neural Network 21](#_Toc160203656)

[1.7.2: Two-Class Decision Forest 25](#_Toc160203657)

[1.8: Result Analysis and Discussion 28](#_Toc160203658)

[1.8.1: Results: Performance metrics of the classification models 28](#_Toc160203659)

[1.8.2: Discussion 28](#_Toc160203660)

[1.9: Conclusion 29](#_Toc160203661)

[1.10: REFERENCES 29](#_Toc160203662)

[Task 2: Exploring Obesity Trends – Comparative Analysis of K-Means and Hierarchical Clustering Algorithms 29](#_Toc160203663)

[2.1: Introduction 29](#_Toc160203664)

[2.2: Research Question 29](#_Toc160203665)

[2.3: Literature Review 29](#_Toc160203666)

[2.4: Dataset 29](#_Toc160203667)

[2.5: Explanation and Preparation of Dataset (Exploratory Data Analysis) 30](#_Toc160203668)

[2.6: Implementation of Clustering Algorithms 35](#_Toc160203669)

[2.6.1: K-Means Clustering Algorithm 35](#_Toc160203670)

[2.6.2: Hierarchical Clustering Algorithm 37](#_Toc160203671)

[2.7: Result Analysis and Discussion 39](#_Toc160203672)

[2.8: Conclusion 39](#_Toc160203673)

[2.9: REFERENCES 40](#_Toc160203674)

[Task 3: Analysing Twitter Sentiments towards Technological Products 40](#_Toc160203675)

[3.1: Introduction 40](#_Toc160203676)

[3.2: Research Question 40](#_Toc160203677)

[3.3: Literature Review 40](#_Toc160203678)

[3.4: Dataset 40](#_Toc160203679)

[3.5: Explanation and Preparation of dataset (Exploratory Data Analysis) 40](#_Toc160203680)

[3.6: Implementation of Text Analysis and Sentiment Analysis 41](#_Toc160203681)

[3.6.1: Text Analysis 41](#_Toc160203682)

[3.6.2: Sentiment Analysis 45](#_Toc160203683)

[3.7: Result Analysis and Discussion 46](#_Toc160203684)

[3.8: Conclusion 46](#_Toc160203685)

[3.9: REFERENCES 47](#_Toc160203686)

# Task 1: Machine Learning Prediction for Employee Attrition

## 

## 1.1: Introduction

Understanding employee attrition is pivotal for organizations, as it necessitate the voluntary leaving of employees and crash various aspects of the workplace. The many-sided nature of employee attrition, control by elements like job satisfaction and organizational lifestyle, entails an in-depth examination for successful retention policies. The importance of employee attrition recline in its possibility to disrupt productivity within a workplace, leading to the dropping of institutional understanding and proficiency, also financial burdens linked with hiring and training new staff. The leaving of skilled co-worker can adversely affect the confidence and commitment of remaining employees, creating an ambience of uncertainty and potentially declining job satisfaction. Constant turnover also constitute challenges to team variability, hindering cooperation and cohesion, which are essential for organizational efficacy. Moreover, the negative effects of attrition extend beyond the internal workings of an organization, impacting its reputation and attractiveness to potential employees and clients. Addressing the root causes of attrition becomes a strategic imperative for organizations seeking stability, positive work culture, and sustained success. dynamic measures in employee retention not only conserve institutional knowledge but also promote a custom of trust, engagement, and continuous development. As the workplace evolves, organizations must navigate the complexities of the modern workforce by prioritizing strategies that mitigate attrition and enhance overall organizational effectiveness. (Sinha et al., 2014). Machine learning presents a promising approach for anticipating and managing employee turnover, setting the stage for the study's focus on predictive models. By leveraging historical data, machine learning algorithms can recognize designs and factors contributing to employee attrition, enabling organizations to anticipate and address probable turnover issues. The study emphasizes the development and implementation of predictive models to proactively manage and mitigate the impact of employee turnover on organizational stability and effectiveness.

## 1.2: Research Question

To what extent can machine learning models predict employee attrition by considering factors such as education, tenure, location, payment tier, age, gender, benching status and domain experience?

## 1.3: Literature Review

The literature consistently underscores the pivotal role of job satisfaction in determining employee attrition rates, emphasizing its multifaceted nature encompassing various work-related factors. Studies indicate that a positive organizational culture, characterized by strong leadership, clear communication, and a supportive work environment, significantly reduces attrition by enhancing overall job satisfaction. Career development opportunities emerge as a critical factor, with research highlighting the correlation between robust professional growth programs, employee engagement, and decreased turnover.

Mishra, S., & Mishra, D. (2013 ) emphasize the interconnectedness of these factors, illustrating that a holistic approach addressing job satisfaction, organizational culture, and career development is essential for effective employee retention strategies. This literature review underscores the importance of evidence-based interventions and organizational strategies to create a workplace conducive to employee satisfaction and, consequently, reduced attrition rates. Traditional methods for predicting employee turnover, characterized by simplicity and dependence on human intuition, encounter challenges in scalability and analyzing large datasets. These limitations underscore the necessity for more advanced approaches capable of addressing the intricate factors influencing attrition. Consequently, there is a rising trend towards investigating machine learning techniques in turnover prediction, harnessing their ability to handle extensive data and unveil subtle patterns, with the aim of improving accuracy and effectiveness in anticipating employee turnover. (Zhao et al., 2019 )

## 1.4: Dataset

The dataset was gotten from Kaggle (Elmetwally, 2023). It contains 4653 samples and 9 features of a company’s employee’s information. All features except the LeaveOrNot feature are independent variables. The dataset is made up of four categorical variables and five numerical variables.

The features are as follows:

* Education: Bachelors, Masters and PHD.
* JoiningYear: Year in which employee joined the company.
* City: Location where employee works.
* Payment Tier: Salary categories
* Age
* Gender
* EverBenched: indicating if an employee has been temporarily out of work.
* ExperienceInCurrentDomain: Number of years of experience an employee has in their current field.
* LeaveOrNot: indicating if the employee has left the organization or not.

## 1.5: Explanation and Preparation of Dataset (Exploratory Data Analysis)

The libraries needed for data manipulation, visualization and analysis were loaded. Then the Employee dataset was loaded into a Pandas dataframe named dataset. The structure and characteristics (column names, data types, number of missing values) of the dataset were examined using dataset.info().

A screenshot of a computer

Description automatically generated

Using the Seaborn’s countplot() fuction, the distribution of educational qualification among employees was plotted where each bar represented the frequency of employees with that qualification. Bachelors was the most common educational qualification indicating that the most of the employees’ highest educational attainment was a bachelor’s degree.

A screen shot of a graph

Description automatically generated

A box plot was used to depict how the year of joining the company varies across the three different cities where the employees are located giving an overview of hiring trends across the different locations.

* *Bangalore*: Employees joined between 2012 to 2018. The interquartile range of joining year was in the middle of 2013-2017 point out when majority of the employees in Bangalore joined. The median joining year was 2015 indicating that partly of the employees joined before 2015, and half joined after 2015.
* *Pune*: Pune employees also joined the company between 2012 to 2018. Nevertheless, majority joined in the middle of 2014 to 2017. This was depict by the interquartile range. Also, the median appears that partly of their employees joined before 2015 and the other partly, after 2015.
* *New Delhi:* In New Delhi, the joining year was also between 2012 to 2018. Like Pune, nearly all employees joined in the middle of 2014 to 2017. However, the median joining year was 2016 spelling that half of the employees joined before 2016 and half after 2016.

A screenshot of a graph

Description automatically generated

Another boxplot was used to compare the distribution of experience across different payment tiers providing insights into how experience relates to compensation within the company.

* Payment Tier 1 employees had 0 to 7 years of experience in their current domain. Most of these employees had 1.5 to 4 years of experience (interquartile range) with half having less than 3 and the other half having more than 3 years of experience (median) in their current domain.
* Employees in payment tiers 1 and 2 have similar distributions. The employees in those payment tiers had 0 to 7 years of experience in the current domain. The interquartile range of experience in current domain is between 2 to 4 years, of which the median experience is 3 years.

A screenshot of a graph

Description automatically generated

A bar chart was used to depict the proportion of males and females in the company. The bar chart revealed that the company employed more males than females.

A screen shot of a graph

Description automatically generated

The distribution of age groups across genders was visualized using a grouped bar chart. This showed that majority of the employees were between the ages of 24 and 28. It also showed that across all the ages, there were more males than females.

A screenshot of a graph

Description automatically generated

A bar chart was used to visualize the prevalence of employees leaving or staying in the company. The bar chart showed that more employees have left or are leaving than the ones that are staying.

A screenshot of a graph

Description automatically generated

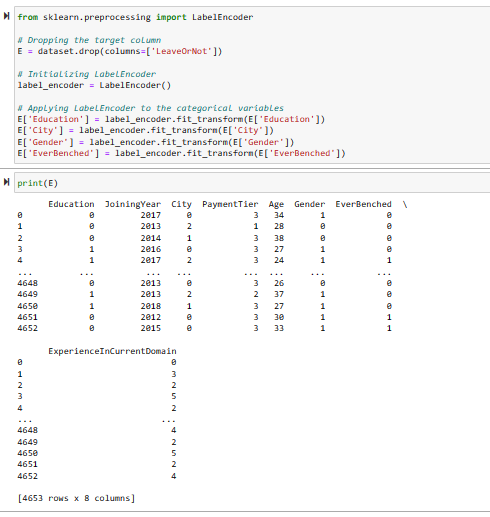
Based on age, a grouped bar chart was used to visualize how the decision to leave or stay varies across different age groups within the company. According to the graph, most of the employees had left or were leaving across all ages. The proportion of 24- to 28-year-olds leaving was higher compared to the other ages.

A screenshot of a graph

Description automatically generated

## 1.6: Implementation in Python (K-Nearest Neighbours – KNN and Decision Tree)

To adequately apply and improve the performance of the classification algorithms on the Employee format dataset they can completely understand, the categorical variables – Education, City, Gender and EverBenched were encoded into numerical variables. The target column, LeaveOrNot, was dropped.



The Employee dataset was then prepared for the classification tasks by slicing it into input and output features. The target column is the output features while the others are input features.

A screenshot of a computer code

Description automatically generated

The dataset was split into training and testing sets at 70% and 30% respectively. The features of the training and testing sets were standardized to have a mean of 0 and a unit variance ensuring that each feature contributed equally to the distance metrics to be used by the classification algorithms.

A screenshot of a computer program

Description automatically generated

### 1.6.1: KNN Classification

KNeighboursClassifier class was imported to carry out the classification task. It was fitted using the training data to learn the patterns and relationships in the training data which will then enable it to make predictions on the test data.

A screenshot of a computer program

Description automatically generated

The trained KNN classifier was then applied to the test data to predict the class labels of the testing set. Due to a few elements of the predicted class label being shown, printing options were set to display all elements. This was then compared with the actual labels, the testing set.



KNN classifier’s performance was then evaluated using different performance metrics.

* Accuracy score comparing the true labels (y\_test) with the predicted labels (y\_pred).
* Confusion matrix, summarizing the model’s performance by tabulating true positive, true negative, false positive and false negative predictions. The confusion matrix was also visualized using a heatmap.
* Classification report providing report on precision, recall, F1-score, and support.

A screenshot of a computer

Description automatically generated A screenshot of a computer screen

Description automatically generated

### 1.6.2: Decision Tree Classification

DecisionTreeClassifier was imported from scikit-learn library to create a decision tree classifier. The decision tree classifier was fitted to the training set to be able to predict the target variable, LeaveOrNot, based on the features provided.

A screenshot of a computer code

Description automatically generated

Then, the predicted outcome of the testing set was obtained based on the decision tree model already built and compared with the actual labels of the testing set..

A screenshot of a computer screen

Description automatically generated

A number grid with numbers

Description automatically generated with medium confidence

The performance of the Decision tree classification model was then evaluated by computing the following metrics – accuracy, confusion matrix (also depicted on a heat map) and classification report.

A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

## 

## 1.7: Implementation in Azure Machine Learning Designer (Two-Class Neural Network and Two-Class Decision Forest)

Before implementing the classification task on Azure Machine Learning Designer, the Employee dataset was preprocessed on Excel. The categorical variables were converted to numerical data to properly apply classification models. New numerical data columns – education, city, gender and Everbenched were created beside their corresponding categorical data columns.

Azure Machine Learning workspace, a new compute cluster and a pipeline were created. The Employee dataset was uploaded and explored in the Data page of the created workspace.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A graph with a bar and a couple of people

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

A blue rectangle with a white background

Description automatically generated

A blue rectangular object with a white background

Description automatically generated

A close-up of a blue box

Description automatically generated

A screenshot of a computer

Description automatically generated

A graph with a bar

Description automatically generated

A blue rectangle with black text

Description automatically generated

The dataset was then dragged and placed on the canvas to begin training the classification model.

To proceed, the dataset was normalized by dragging the Normalize Data module from Component tab in Asset library, to the canvas below the Employee dataset. The output of the Employee dataset was connected to the input of the Normalize Data module. The transformation method of the module was set to MinMax and all the numerical data columns were selected for normalization or standardisation to prevent bias towards features with larger scales during the training process.

A screenshot of a computer

Description automatically generated

This pipeline was then run as an experiment and viewed as a transformed dataset in Jobs panel

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

### 1.7.1: Two-Class Neural Network

After normalization, the dataset was split using the Split Data module in component tab in Data panel. The Split Data module was dragged and dropped on the canvas under the Normalize Data module. It was connected to the transformed dataset output of the Normalize Data module through its input. The fraction of rows in the first output dataset was set at 0.7.

To train a model, the Train Model module from the component tab in Data panel under Assets library was dragged and dropped to the left below the Split Data module. The results dataset1 output of the Split Data module was connected to the dataset input of the Train Model module. Since the model to be trained was to predict the LeaveOrNot value, it was set as the label column.

The Two-Class Neural Network module was then dragged from the component tab and dropped to the left of the Split Data module. Its untrained model output was connected to the untrained model input of the Train Model module.

A Score Model was also dragged and dropped below and to the right of the Train Model module to predict labels for the features in the normalized dataset. Its input was connected to the output of the Trained Model module. Its dataset input was connected to the results dataset2 output of the Split Data module.

This pipeline was then run, training the model, and predicting the labels.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

After training and predicting the LeaveOrNot label, the training model – Two-Class Neural Network was evaluated by dragging and dropping Evaluate Model module from component tab to the canvas just below Score Model module. The output of Score Model module was connected to the input of Evaluate Model module and ran as a pipeline.

A screenshot of a computer

Description automatically generated

When the experiment had finished running, the evaluation result was viewed.

A graph showing a curve

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

### 1.7.2: Two-Class Decision Forest

A Two-Class Decision Forest classification model was used to train the normalized dataset side by side.

To do this, a second Train Model module was dragged to the canvas at the far right of the first Training Model module. Its dataset input was connected to the results dataset1 output of Split Data module on the canvas.

Two-Class Decision Forest module was dragged from the component tab to the canvas and placed above the second Train Model module. Its untrained model output was connected to the untrained model input of the second Train Model module.

A second Score Model module was dragged to the canvas and placed to the right of the first Score Model module. Its trained model input was connected to the trained model output of the second Train Model module. Its dataset input was connected to the results dataset2 output of Split Data module.

This pipeline created was then ran as an experiment and viewed afterwards.

A screenshot of a computer

Description automatically generated

To evaluate the scored dataset, the scored dataset output of second Score Mode module was connected to the second scored dataset input of the Evaluate Model module already on the canvas. The pipeline was then run and the evaluation results viewed.

A screenshot of a computer

Description automatically generated

A graph showing a number of patients

Description automatically generated with medium confidence

A screen shot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

## 1.8: Result Analysis and Discussion

### 1.8.1: Results: Performance metrics of the classification models

#### **KNN**

Accuracy: 81% Confusion Matrix: [842 80]

**precision recall F1-score support** [190 284

0 0.82 0.91 0.86 922

1 0.78 0.60 0.68 474

#### **Decision Tree**

Accuracy:81% Confusion Matrix: [809 113]

**precision recall F1-score support** [153 321]

0 0.84 0.88 0.86 922

1 0.74 0.68 0.71 474

#### **Two-Class Neural Network**

Accuracy 79.1% Precision 0.795, Recall 0.516 F1 Score 0.626

AUC 0.815

Confusion Matrix: [244 63]

[229 860]

#### **Two\_Class Decision Forest**

Accuracy 82.9% Precision 0.785, Recall 0.681 F1 Score 0.729

AUC 0.84

Confusion Matrix: [322 88]

[151 835]

### 1.8.2: Discussion

* K-Nearest Neighbours (KNN) achieved an accuracy of 81% with a confusion matrix indicating 842 true negatives, 80 false positives, 190 false negatives, and 284 true positives. The classification report depicts proportionate good exactness and recall for class 0 (negative) but lower values for class 1 (positive).
* Decision Tree also achieved an accuracy of 81%, with a confusion matrix showing 809 true negatives, 113 false positives, 153 false negatives, and 321 true positives. The classification report specify slightly lower accuracy and recall compared to KNN.
* Two-class Neural Network achieved an accuracy of 79.1% with a precision of 79.5%, recall of 51.6%, and F1-score of 62.6%. The confusion matrix shows 244 true negatives, 63 false positives, 229 false negatives, and 860 true positives.
* Two-class Decision Forest achieved an accuracy of 82.9 with a precision of 78.5%, recall of 68.1%, and F1-score of 72.9%. The confusion matrix indicates 322 true negatives, 88 false positives, 151 false negatives, and 835 true positives.

These classification algorithms showed varying levels of performance in predicting employee attrition. Decision Forest outperformed other models with the highest accuracy, precision, recall, and F1-score, indicating its effectiveness in handling both classes. Two-class Neural Network performed reasonably well but had lower recall compared to Decision Forest. KNN and Decision Tree algorithms also performed satisfactorily but have limitations in accurately predicting positive cases of attrition.

## 1.9: Conclusion

The classification algorithms exhibited varying capabilities in predicting employee attrition based on factors such as education, tenure, location, payment tier, age, gender, benching status, and domain experience. The classification algorithms particularly Two-class Decision Forest and Two-class Neural Network, can predict employee attrition to a reasonable extent. While KNN and decision tree models performed well in terms of accuracy and precision, the two-way neural network and two-class decision forest models offer competitive performance with balanced precision and recall.

Employers can utilize these algorithms to identify at-risk employees and implement targeted retention strategies. However, further optimization and improvement of the models would be necessary to improve their predictive accuracy and generalizability in real-world scenarios.

## 

## 1.10: REFERENCES

Elmetwally, T. (2023). Employee dataset. *Kaggle*. <https://www.kaggle.com/datasets/tawfikelmetwally/employee-dataset>

Sinha, V., Abraham, A., Bhaskarna, B., Xavier, K., & Kariat, K. (2014). Role efficacy: Studying the impact on employee engagement, employee motivation and attrition. *International Journal of Human Capital and Information Technology Professionals (IJHCITP)*, *5*(4), 35-54. DOI: 10.4018/ijhcitp.2014100103

Ajit, P. (2016). Prediction of employee turnover in organizations using machine learning algorithms. *algorithms*, *4*(5), C5. DOI:[10.14569/IJARAI.2016.050904](https://doi.org/10.14569/IJARAI.2016.050904)Mishra, S., & Mishra, D. (2013). Review of literature on factors influencing attrition and retention. *International Journal of Organizational Behaviour & Management Perspectives*, *2*(3), 435-444.

Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2019). Employee turnover prediction with machine learning: A reliable approach. In *Intelligent Systems and Applications: Proceedings of the 2018 Intelligent Systems Conference (IntelliSys) Volume 2* (pp. 737-758). Springer International Publishing. https://doi.org/10.1007/978-3-030-01057-7\_56

# Task 2: Exploring Obesity Trends – Comparative Analysis of K-Means and Hierarchical Clustering Algorithms

## 2.1: Introduction

Obesity is a compound public health issue with intense involvement for individuals and societies worldwide. Understanding obesity trends and patterns is important for developing successful public health interventions, personalized treatment plans, and addressing the associated metabolic risks. Utilizing data analysis techniques such as k-means and hierarchical clustering algorithms provide a powerful viewpoint to gaining perception into obesity-related factors, recognizing clear subgroups within the overweight and obese population, and informing targeted interventions. (Li et al., 2019)

The shoot up prevalence of obesity has set off an alarming global health treat, necessitating a comprehensive understanding of its basic factors and trends. Across the world, obesity rates have witnessed a fundamental rise, posing notable challenges to public health systems. This surge in obesity is closely related to various risk factors, encompassing lifestyle changes, dietary habits, and sedentary behaviors prevalent in modern societies. Recognizing the complex web of contributors to global obesity is important for putting together successful plans that address the many-sided nature of this health crisis. Lawmakers must think about not only individual behaviors but also broader societal influences, such as urbanization and the availability of unhealthy food options, to develop policies that can limit the rising tide of obesity.

## 2.2: Research Question

How do K-Means and Hierarchical Clustering Algorithms uncover unique obesity patterns by considering individual’s health and lifestyle factors?

## 2.3: Literature Review

Obesity is a worldwide public health challenge signalized excessive body fat accumulation, posing notable threats to health. This literature review synthesizes discovery on obesity prevalence, inspect multiple risk factors putting up to its problem, and delves into the multifactorial health results connected with obesity. Many studies highlight the rising prevalence of obesity globally. The World Health Organization (WHO) call attention to escalating rates, with both developed and developing countries experiencing an alarming increase. Research consistently specify variations across demographics, emphasizing the significance of understanding obesity as a variability and evolving health concern. Dietary routine rich in energy-dense foods and sedentary way of life are key contributors to obesity. Studies highlight the effect of modern diets, marked by high-calorie, processed foods, and reduced physical activity, on the increasing prevalence of obesity. The interaction between genetic factors and environmental impact is a central theme in obesity research. Twin and family studies highlight genetic predispositions, while environmental factors such as obesogenic surroundings amplify the risk. The complexity arises from the intricate interplay between nature and nurture. (Frood et al., 2013).

This literature review examines the global prevalence of obesity, emphasizing its upward trajectory and variations across demographics. It explores the multifactorial nature of obesity, identifying key risk factors such as dietary habits, genetics, socioeconomic status, and psychosocial influences. The interaction between genetic predispositions and environmental factors highlight the problem of obesity etiology. The review highlight the complex relationships between obesity and various health results, including cardiometabolic risks, type 2 diabetes, mental health effect, cancer, and orthopedic and respiratory difficulties. Recognizing the interconnected essence of obesity is important for developing comprehensive plan to address its factors and abate associated health risks. (Sarma et al., 2021). The theoretical basis of obesity involves an intricate interaction of genetic, environmental, and behavioral elements. Genetic predispositions contribute to an individual's susceptibility, while environmental influences, such as obesogenic surroundings and access to unhealthy foods, amplify risk. Behavioral aspects, encompassing lifestyle choices and psychosocial factors, play a crucial role in shaping eating behaviors and physical activity patterns.

Grundy, S.M. (1998). The multifactorial nature of obesity acknowledge that its growth arises from a combination of genetic susceptibility interacting with a myriad of environmental and way of life factors. This complex web of factors highlight the need for holistic and personalized approaches in understanding and addressing the obesity outbreak.

Clustering algorithms, grounded in unsupervised learning principles, employ distance-based metrics and centroid-based approaches to identify inherent patterns within health data. Theoretical foundations involve concepts like Euclidean distance and density-based methods, allowing for the grouping of data points based on similarities. In practical applications, clustering algorithms play a crucial role in health data analytics by identifying distinct subgroups and patient profiles, facilitating precision medicine and personalized healthcare. Disease phenotyping and risk stratification benefit from clustering, enabling tailored interventions and efficient resource allocation in healthcare settings. Moreover, clustering supports public health initiatives by categorizing population subgroups, guiding targeted interventions for specific health issues within identified clusters, however, K-means clustering exhibits strength in efficiently handling large datasets and producing tight, well-defined clusters based on centroid distances. However, its limitation lies in sensitivity to initial cluster assignments and struggles with non-spherical or unevenly sized clusters. On the other hand, hierarchical clustering offers an advantage in revealing hierarchical relationships among clusters but may be computationally intensive for large datasets, and its outcomes are affected by the choice of linkage methods. Ultimately, the selection between k-means and hierarchical clustering depends on the dataset's characteristics and the desired insights from obesity-related data.

## 2.4: Dataset

The Obesity Dataset used was gotten from Kaggle (Palechor & Manotas, 2019a). It contains 17 features and 2111 samples, capturing obesity estimates among individuals across Mexico, Peru and Colombia between the ages of 14 and 61. The dataset reflects the individuals’ dietary patterns and physical lifestyle.

Palechor & Manotas, 2019b explained the meaning of each feature as follows:

1. Gender: Male or Female.
2. Age
3. Height: in metres.
4. Weight: in kilograms.
5. family\_history\_with\_overweight: Family member suffers or suffered from overweight – yes or no.
6. FAVC: Frequent consumption of high caloric food – yes or no.
7. FCVC: Frequency of consumption of vegetables. (1 = Never, 2 = Sometimes, 3 = Always)
8. NCP: Number of main meals. (1 = once, 2 = twice, 3 = thrice, 4 = more than three times)
9. CAEC: Consumption of food between meals. (No, Sometimes, Frequently or Always)
10. SMOKE: Smoker – yes or no.
11. CH20: Consumption of water daily. (1 = Less than a litre, 2 = Between 1 and 2 litres, 3 = More than 2 litres).
12. SCC: Calories consumption monitoring – yes or no.
13. FAF: Physical activity frequency. (0 = None, 1 = one to two days, 2 = two to four days, 3 = four to five days).
14. TUE: Time using technology devices. (0 = zero to two hours, 1 = three to five hours, 2 = more than five hours).
15. CALC: Consumption of alcohol – No, Sometimes, Frequently, Always.
16. MTRANS: Means of transportation – Walking, Bike, Motorbike, Public\_Transportation, Automobile.
17. NObeyesdad: Obesity level deducted – Insufficient\_Weight, Normal\_Weight, Overweight\_Level\_I, Overweight\_Level\_II, Obesity\_Type\_I, Obesity\_Type\_II, and Obesity\_Type\_III.

Of the 17 variables, nine were categorical and the others, numerical.

## 2.5: Explanation and Preparation of Dataset (Exploratory Data Analysis)

Necessary Python libraries needed to work with the dataset and to create visualisations, NumPy, Pandas, Matplotlib and Seaborn were imported. The dataset was loaded into a Pandas DataFrame called ‘dataset’.

The first and last five rows of the dataset were viewed using the ‘dataset.head()’ and ‘dataset.tail()’ codes respectively in order to quickly inspect the dataset and to ensure it had been loaded successfully.

A summary of the dataset was gotten showing the number of rows and columns, column names, column datatypes, count of non-missing values in each column and memory size. No columns had missing values.

A screenshot of a computer

Description automatically generated

The distribution of obesity levels across different age groups was visualized using a stacked bar chart. The Ages were grouped in 10 years intervals. Each bar represented a specific age group and the height of each section within the bar showed the number of individuals in that age group with a particular obesity level.

A screenshot of a computer

Description automatically generated

The distribution of transportation modes used by individuals in the dataset was visualized using a piechart where each segment represents the proportion of individuals using a distinct transportation mode.

A screen shot of a pie chart

Description automatically generated

The distributions of four variables (Age, CALC, FAF, TUE) were visualised shedding light on the age demographics, calorie consumption habits, physical activity levels, and technology usage patterns.

A screenshot of a computer program

Description automatically generated

A group of graphs showing different sizes of data

Description automatically generated with medium confidence

The Age Group column created for visualization was removed from the dataset as it would no longer be needed for further analysis.

The categorical variables of the dataset were converted to numerical variables (0, 1, 2, 3, 4, 5, 6) corresponding with the number of responses each categorical variable has. This was done to make the variables suitable to apply K-Means and Hierarchical clustering algorithms.

A screenshot of a computer

Description automatically generated

The relationships between the different variables were visualized using a pairplot to identify patterns and correlations. This pairplot showed a grid of scatterplots.

A screenshot of a computer screen

Description automatically generated

## 2.6: Implementation of Clustering Algorithms

Preprocessing steps for clustering were carried out. StandardScaler from ‘sklearn.preprocessing’ was imported to standardize the dataset ensuring that all the features have the same scale, a mean value of 0 and a standard deviation of 1 and can contribute equally to a proper clustering process.

A screenshot of a computer program

Description automatically generated

### 2.6.1: K-Means Clustering Algorithm

KMeans library was imported for clustering. The Elbow method was then used to determine the ideal number of clusters for KMeans clustering by analysing the within-cluster sum of squares (WCSS) across different K values of range 1 to 10. K value of 2 was noted as the elbow point in the graph where the decrease in the WCSS became less rapid.

A screen shot of a computer program

Description automatically generated

A graph with a line

Description automatically generated

Due to the high dimensionality of the dataset as it now has 17 numerical variables, Principal Component Analysis (PCA) was used to reduce the dimensionality allowing for better scatterplot visualization in a lower-dimensional space while retaining as much variance as possible. Two new dimensions with over 31% of the original variance were gotten.

A screenshot of a computer

Description automatically generated

The KMeans clusters were then visualized in a scatter plot using two distinct colours, red and blue. This enabled the visualisation of the clustering obesity patterns and relationships of the individuals based on their health and lifestyle features.

A screen shot of a computer screen

Description automatically generated

### 2.6.2: Hierarchical Clustering Algorithm

Necessary functions for hierarchical clustering analysis like AgglomeottedrativeClustering function ported from scikit-learn and SciPy libraries. Linkage matrix was computed using Ward’s method. Then the hierarchial clustering dendogram was plotted.

A screenshot of a computer screen

Description automatically generated

7 clusters were gotten by taking a horizontal line around halfway down the dendogram. Hierarchial clustering was fitted to the dataset same way as KMeans clustering algorithm. However, 7 clusters were specified as gotten from the dendogram.

A screenshot of a computer program

Description automatically generated

The 7 clusters formed by hierarchical clustering were visualized using different colours for clarity. Points belonging to the same cluster were marked with the same colour.

A screen shot of a computer

Description automatically generated

## 2.7: Result Analysis and Discussion

The analyses aimed to explore obesity trends using a comparative analysis of K-Means and Hierarchical Clustering algorithms. The research question sought to understand how these algorithms uncover unique obesity patterns based on individual health and lifestyle factors.

The elbow point for K-Means clustering was identified at 2, indicating the optimal number of clusters. In the K-Means plot, the data points of each cluster were closely related to each other, with one end of each cluster touching the other, suggesting relatively compact clusters.

For Hierarchical Clustering, 7 clusters were identified halfway down the dendrogram. Among these clusters, 1, 3, and 6 displayed closely related data points, indicating clear and distinct patterns. Clusters 2, 5, and 7 showed closely related data points but with some space between them, suggesting moderate cohesion. However, Cluster 4 exhibited more scattered data points, indicating greater variability within this group.

## 2.8: Conclusion

The study provides insights into obesity trends through the lens of clustering algorithms. K-Means and Hierarchical Clustering offer different perspectives on the underlying patterns within the dataset. K-Means clustering suggests relatively compact clusters, while Hierarchical Clustering reveals varying levels of cohesion among different clusters.

Understanding these obesity patterns can inform public health interventions, personalized healthcare strategies, and targeted wellness programs. Further research could explore additional clustering techniques, validate findings across diverse datasets, and integrate other variables to enhance the accuracy and applicability of obesity trend analysis.

## 2.9: REFERENCES

Palechor, F. M., & Manotas, A. de la H. (2019a). Obesity or CVD risk (Classify/Regressor/Cluster). *Kaggle.* <https://www.kaggle.com/datasets/aravindpcoder/obesity-or-cvd-risk-classifyregressorcluster/data>

Palechor, F. M., & Manotas, A. de la H. (2019b). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. *Data in Brief*, *25*, 104344. <https://doi.org/10.1016/j.dib.2019.104344>

Li, L., Song, Q., & Yang, X. (2019). K-means clustering of overweight and obese population using quantile-transformed metabolic data. *Diabetes, metabolic syndrome and obesity: targets and therapy*, 12:, 1573-1582, DOI: [10.2147/DMSO.S206640](https://doi.org/10.2147/DMSO.S206640)

Malik, V. S., Willett, W. C., & Hu, F. B. (2013). Global obesity: trends, risk factors and policy implications. *Nature reviews endocrinology*, *9*(1), 13-27. https://doi.org/10.1038/nrendo.2012.199

Sarma, S., Sockalingam, S., & Dash, S. (2021). Obesity as a multisystem disease: Trends in obesity rates and obesity‐related complications. *Diabetes, Obesity and Metabolism*, *23*, 3-16. <https://doi.org/10.1111/dom.14290>

Frood, S., Johnston, L. M., Matteson, C. L., & Finegood, D. T. (2013). Obesity, complexity, and the role of the health system. *Current obesity reports*, *2*, 320-326. https://doi.org/10.1007/s13679-013-0072-9

Grundy, S. M. (1998). Multifactorial causation of obesity: implications for prevention. *The American journal of clinical nutrition*, *67*(3), 563S-572S.

# Task 3: Analysing Twitter Sentiments towards Technological Products

## 3.1: Introduction

Social media platforms, particularly Twitter, have become integral channels for individuals to express opinions, engage in discussions, and share sentiments on a wide range of topics. The increasing significance of platforms like Twitter as rich sources of data stems from their ability to capture real-time, unfiltered expressions of public opinions. In the realm of technological products, analyzing Twitter data provides valuable insights into how users perceive, discuss, and react to innovations, contributing to a deeper understanding of consumer sentiments and preferences.

Twitter, with its vast user base and succinct communication style, serves as a dynamic environment for users to share viewpoints on various technological products. The platform's real-time nature allows researchers and businesses to tap into the pulse of public opinion, tracking trends, identifying influencers, and gauging sentiment shifts. Natural Language Processing (NLP) and sentiment analysis techniques are frequently employed to extract meaningful insights from the massive volume of tweets, providing a nuanced understanding of user sentiments towards emerging technologies, gadgets, and software. As a result, Twitter has evolved into a valuable resource for businesses, researchers, and policymakers alike, offering a direct and immediate connection to the collective voice of the public on matters related to technological advancements. (Murphy et al., 2014).

Twitter, among various social media platforms, is gaining significance as a valuable source for understanding public opinions and sentiments on diverse topics, especially technological products. Its real-time and unfiltered nature allows for capturing immediate and dynamic expressions of user perceptions. With a vast user base, Twitter serves as a dynamic environment where individuals discuss and share their views on emerging technologies. Leveraging Natural Language Processing and sentiment analysis techniques, researchers and businesses can extract valuable insights from the massive volume of tweets, providing a nuanced understanding of user sentiments towards technological innovations. As a result, Twitter has become an essential resource, offering direct access to the collective voice of the public and aiding businesses, researchers, and policymakers in comprehending and responding to technological trends.

The growing importance of sentiment analysis lies in its capacity to extract valuable insights from the vast volume of user-generated content, especially in the realms of business intelligence, market research, and customer relationship management (CRM), sentiment analysis plays a crucial role in transforming unstructured user-generated content into actionable insights, contributing significantly to informed decision-making across various business domains. Its applications in business intelligence, market research, and CRM underscore its growing importance as a strategic tool in today's data-driven and customer-centric business landscape. (Xu et al., 2019). Sentiment analysis is pivotal in business intelligence, providing organizations with a systematic approach to analyze customer opinions and preferences. By gauging sentiment, businesses can understand how their products or services are perceived in the market, helping in strategic decision-making and adapting to changing consumer sentiments.

In market research, sentiment analysis offers a powerful tool for monitoring public opinion, tracking trends, and evaluating the success of marketing campaigns, analyzing sentiment across various platforms, including social media and customer reviews, helps businesses stay attuned to market dynamics and consumer preferences. Sentiment analysis enhances CRM by enabling organizations to proactively respond to customer feedback and concerns. It helps in identifying areas for improvement, addressing customer issues promptly, and fostering positive customer experiences, thereby strengthening customer relationships.

Businesses utilize sentiment analysis to manage and monitor their brand reputation across online platforms. Identifying and addressing negative sentiment in real-time allows companies to protect their brand image and implement strategies to enhance overall brand perception. Sentiment analysis aids in product development by providing insights into customer expectations, preferences, and areas for improvement. Analyzing sentiment helps businesses stay innovative, aligning their products with evolving consumer needs and staying ahead in a competitive market.

Understanding sentiment extends to competitor analysis, helping businesses evaluate their standing in comparison to industry rivals, by monitoring sentiment towards competitors, organizations can identify market gaps, potential areas for differentiation, and refine their competitive strategies.

Sentiment analysis has implications for financial decision-making as well. Investors and financial analysts leverage sentiment data to gauge market sentiment towards specific stocks or sectors, influencing investment decisions. In the realm of digital marketing, sentiment analysis aids in optimizing social media strategies, analyzing user sentiments helps businesses tailor their content, campaigns, and messaging to resonate positively with their target audience.

## 3.2: Research Question

What insights can be derived from tweets about diverse technological products?

## 3.3: Literature Review

Research findings in sentiment analysis reveal temporal patterns in social media sentiments, with fluctuations tied to events and cultural moments. User demographics, including age and gender, influence sentiment expression on social media, adding a layer of complexity to understanding user opinions. Visual content, such as images and videos, significantly impacts sentiment expression, prompting the need for multimodal sentiment analysis. Fine-grained sentiment analysis in technological product reviews and aspect-based sentiment analysis provide detailed insights into user opinions on specific product features. Ethical considerations, including privacy concerns and biases, are recognized, and machine learning models, particularly deep learning, prove effective in sentiment analysis tasks across social media and product reviews. (Dey, Raktim Kumar, et al., 2020).

Mitra, A. (2020). Sentiment analysis employs Natural Language Processing (NLP) techniques, including tokenization and text preprocessing, alongside machine learning algorithms like SVM and deep learning models such as RNNs. Sentiment lexicons like SentiWordNet and machine learning tools like Scikit-Learn provide the basis for sentiment classification. Additional methodologies include rule-based systems, ensemble methods, and domain-specific adaptations, with evaluation metrics assessing model performance and commercially available APIs streamlining the integration of advanced sentiment analysis techniques.

Sentiment analysis faces challenges rooted in the complexity of language, including ambiguity arising from polysemy and homonymy. The context dependency of sentiments is a significant hurdle, as nuanced expressions, sarcasm, and tone variations can lead to misinterpretation. Issues with negation handling and contrast identification pose difficulties in accurately determining sentiment polarity, especially in short texts where context may be lacking. The domain-specific nature of language, including vocabulary variation, slang, and industry jargon, requires tailored sentiment analysis models for different sectors. Cultural and demographic variances, coupled with dynamic language evolution, add layers of complexity, emphasizing the need for ongoing refinement and adaptation in sentiment analysis approaches.

Mostafa, M. M. (2013). Sentiment analysis is widely applied across diverse industries, including marketing and advertising, where it optimizes campaign effectiveness and aids in brand perception management. In customer service, sentiment analysis helps businesses understand feedback, improve experiences, and swiftly address issues. E-commerce leverages sentiment analysis for product development, competitor analysis, and understanding customer opinions. Industries like healthcare use sentiment analysis to enhance patient experiences and monitor public perceptions of pharmaceutical products. Politics, tourism, human resources, education, financial services, and media also benefit from sentiment analysis, guiding decision-making, improving services, and managing public perceptions. Overall, sentiment analysis plays a pivotal role in shaping strategies and enhancing stakeholder experiences across a spectrum of industries.

## 3.4: Dataset

The Twitter Product Sentiment Analysis dataset was gotten from Kaggle (Densil, 2021). It is made up of two features and 9871 samples. The features were named id and tweet giving each tweet (text) a numerical value in the id column.

## 3.5: Explanation and Preparation of dataset (Exploratory Data Analysis)

Necessary Python libraries for data manipulation, visualization and natural language processing tasks were imported. The dataset was read into a pandas DataFrame called ‘data’.

The dataset was inspected to understand the structure of the dataset. ‘data.head()’ function showed the first five rows. ‘data.shape’ showed the dimensions of the dataset as 9873 rows and 2 columns. ‘data.dtypes’ showed the data types of each column, id as integer type and tweet as object/string type. ‘data.isnull().sum()’ function was used to check for A screenshot of a computer

Description automatically generated

## 3.6: Implementation of Text Analysis and Sentiment Analysis

### 3.6.1: Text Analysis

The text data in the tweet columns was preprocessed and cleaned by removing punctuations, converting it to lowercase and removing numbers. The cleaned text was stored in a new column named Newtweet.

A screenshot of a computer

Description automatically generated

The Newtweet column was tokenized to check if words in the column are stopwords and then were reconstructed without the stopwords to focus on the most relevant words. Stopwords are commonly occurring words that do not have significant meaning in text analysis. The reconstructed tweets were then stored in a StopRemovedtweet column.

A screen shot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

Lemmatization was carried out on the StopRemovedtweet column to reduce the words in the tweets to their base normalizing them for analysis. The lemmatized words were then joined back to form the Cleantweet column.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A word cloud was visualized displaying 2000 words where the size of each word represented the frequency of occurrence in the dataset. Words like iPhone, Apple and Samsung were more prominent showing that they are more common int the tweet variable compared to the less prominent words.

A screen shot of a computer screen

Description automatically generated

The cleaned tweets were processed, and unigrams were created. Unigrams are individual words. The 15 most common unigrams found in the tweets were shown in a bar plot. The height of each bar is the frequency count of each unigram. Iphone had the highest number of frequency count, followed by apple and samsung.

A screenshot of a computer

Description automatically generated

A screen shot of a graph

Description automatically generated

Bigrams, a combination of two words, were generated from the previously extracted unigrams and counted. The most frequent 15 of the bigrams in the tweet were plotted as a bar chart with the height of each bar corresponding to its frequency. Samsung galaxy, apple iphone and new phone had the highest number of counts.

A graph of a number of orange bars

Description automatically generated with medium confidence

### 3.6.2: Sentiment Analysis

The VADER sentiment analysis tool was installed and imported. The clean tweet was analysed using VADER and assigned a sentiment label of either positive, negative, or neutral based on the score provided by VADER. This sentiment labels were then used to form a column called sentimentLabel.



A screenshot of a computer program

Description automatically generated

This sentiment analysis classifies the tweets based on emotional tone towards the topics and products in the analysed tweets.

From the sentimentLabel column of the dataset, the sentiment labels were counted and visualized on a bar chart. The result revealed that among the analysed tweets, 5427 are neutral, 3634 were positive and 812 were negative.

A screenshot of a computer screen

Description automatically generated

## 3.7: Result Analysis and Discussion

The sentiment analysis conducted on tweets related to technological products yielded insightful findings. Among the sentiments identified—positive, negative, and neutral—neutral sentiments dominated the dataset, comprising 54.98% of the total tweets analyzed. Positive sentiments followed, representing 36.85% of the tweets, while negative sentiments were the least prevalent, accounting for only 8.17% of the dataset.

The dominance of neutral sentiments suggests that a significant portion of tweets related to technological products may contain general or non-opinionated content, such as news updates, product descriptions, or non-emotive discussions. Positive sentiments, on the other hand, indicate favorable opinions, satisfaction, or appreciation expressed by users towards various tech products. Conversely, negative sentiments highlight areas of concern, dissatisfaction, or criticism voiced by users, which could indicate potential areas for improvement or issues requiring attention from product developers and marketers.

## 3.8: Conclusion

Through sentiment analysis of Twitter data pertaining to technological products, valuable insights into public perceptions, opinions, and sentiments towards diverse tech products have been obtained. The predominance of neutral sentiments underscores the importance of distinguishing between informative and opinionated content in social media analysis. Positive sentiments reflect user satisfaction and endorsement of tech products, while negative sentiments offer opportunities for addressing concerns and enhancing product offerings. Overall, leveraging sentiment analysis enables a deeper understanding of consumer sentiment dynamics and can inform strategic decision-making processes in product development, marketing strategies, and customer engagement initiatives within the technology sector.

## 3.9: REFERENCES

Densil, B. (2021). Twitter Product Sentiment Analysis. *Kaggle*. <https://www.kaggle.com/datasets/blessondensil294/twitter-product-sentiment-analysis/data>

Murphy, J., Link, M. W., Childs, J. H., Tesfaye, C. L., Dean, E., Stern, M., ... & Harwood, P. (2014). Social media in public opinion research: Executive summary of the AAPOR task force on emerging technologies in public opinion research. *Public Opinion Quarterly*, *78*(4), 788-794. <https://doi.org/10.1093/poq/nfu053>

Xu, X., Wang, X., Li, Y., & Haghighi, M. (2017). Business intelligence in online customer textual reviews: Understanding consumer perceptions and influential factors. *International Journal of information management*, *37*(6), 673-683. <https://doi.org/10.1016/j.ijinfomgt.2017.06.004>

Dey, R. K., Sarddar, D., Sarkar, I., Bose, R., & Roy, S. (2020). A literature survey on sentiment analysis techniques involving social media and online platforms. *International Journal Of Scientific & Technology Research*, *1*(1), 166-173.

Mitra, A. (2020). Sentiment analysis using machine learning approaches (Lexicon based on movie review dataset). *Journal of Ubiquitous Computing and Communication Technologies (UCCT)*, *2*(03), 145-152. [10.36548/jucct.2020.3.004](https://doi.org/10.36548/jucct.2020.3.004)

Mostafa, M. M. (2013). More than words: Social networks’ text mining for consumer brand sentiments. *Expert systems with applications*, *40*(10), 4241-4251. <https://doi.org/10.1016/j.eswa.2013.01.019>