**Life- Style Analysis**

By

Oladipo Ayomide. O

Table of Contents

[What is data? 4](#_Toc155131644)

[Data preparation 4](#_Toc155131645)

[Machine learning 4](#_Toc155131646)

[Business/Data Understanding 4](#_Toc155131647)

[Data Preparation 5](#_Toc155131648)

[Early Data Analysis: 5](#_Toc155131649)

[Renaming Variables: 6](#_Toc155131650)

[Statistical exploration: 7](#_Toc155131651)

[Visualization: 7](#_Toc155131652)

[Distribution plot: 8](#_Toc155131653)

[EDA summary: 9](#_Toc155131654)

[Feature engineering 10](#_Toc155131655)

[Feature Normalisation: 10](#_Toc155131656)

[Robust scaler results: 11](#_Toc155131657)

[Result: 11](#_Toc155131658)

[PCA and LDA: 12](#_Toc155131659)

[MODELLING: 13](#_Toc155131660)

[Fitting Model: 13](#_Toc155131661)

[Hyperparameter tunning: 16](#_Toc155131662)

[Max- Depth tunning: 17](#_Toc155131663)

[Finding Important Feature using Scikit-learn: 18](#_Toc155131664)

[Grid Search to Find Optimal Hyperparameters: 19](#_Toc155131665)

[CONCLUSION: 19](#_Toc155131666)

[Reference list 20](#_Toc155131667)

[Figure 1: CRISP-DM Life Cycle 6](#_Toc155131673)

[Figure 2: Importing Libraries. 6](#_Toc155131674)

[Figure 3: Data dictionary (UCI repository, 2023). 7](#_Toc155131675)

[Figure 4: Renaming process. 8](#_Toc155131676)

[Figure 5: Box plot illustration of outliers (TrainDataHub, 2021) 9](#_Toc155131677)

[Figure 6: Boxplot results of numerical variables showing outliers and skewed distribution. 9](#_Toc155131678)

[Figure 7: Distribution plots of numerical variables showing highly skewed distribution. 11](#_Toc155131679)

[Figure 8: Correlation matrix dataframe for numerical variables in 'df'. 11](#_Toc155131680)

[Figure 9: Encoding age\_group variable. 12](#_Toc155131681)

[Figure 10: Results of Min-max scaled feature on classifying age\_group with random forest. 13](#_Toc155131682)

[Figure 11: Optimal Dimensionality illustrated. 14](#_Toc155131683)

[Figure 12: Independent variable feature selection. 16](#_Toc155131684)

[Figure 13: Defining dependent variable. 16](#_Toc155131685)

[Figure 14: Overfitted model. 16](#_Toc155131686)

[Figure 15: 30% and 20 % testing size after dropping age. 16](#_Toc155131687)

[Figure 16: Imbalance data illustration. 17](#_Toc155131688)

[Figure 17: Result after balancing dataset and splitting at (20 & 30) %. 17](#_Toc155131689)

[Figure 18: Max-depth graph and accuracy. 20](#_Toc155131690)

[Figure 19: Feature selection df 20](#_Toc155131691)

[Figure 20: GridSearchCV parameter search 21](#_Toc155131692)

# Introduction

## What is data?

Data are plain facts gathered within a defined context. Statisticians would refer to it as sets of observations contained in variables or columns of varying/unique characteristics. Data can also be referred to as a piece of information after it has been summarised to get its true meaning, and when subjected to analysis, data can be referred to as evidence. Before data can become information and ultimately evidence it must go through a process. Data analysis is the process in which raw data is ordered and organized, to be used in methods that help to explain the past and predict the future (Hector, 2013). Data analysis been a scientific process, there are not necessarily any methodology to conclude on every task, however generally, in other to achieve a quality result in any analysis, the data intended to be used for predictions with machine learning algorithm must be processed. This project is an integrated project of both data preparation and machine learning modelling.

## Data preparation

Data preparation also known as data preprocessing in a crucial step for getting a reliable data analysis and prediction results. This process involves early data analysis (EDA), data cleaning, transformation, handling imbalanced data, feature engineering, handling imbalanced data, visualisation, and feature selection and many more. DP takes over 80% of data analysis and modelling process because if we put ill-prepared data into our machine learning model, we would get bad results, hence having through DP is imperative to getting a reliable prediction from data.

## Machine learning

Machine learning (ML) is a branch of artificial intelligence (AI) that focuses on building computer systems that learn from data (Linda, 2021). Machine learning algorithms are trained to find relationships and patterns in data (Linda, 2021). Though different ML algorithms work in different ways, their goal is to make predictions about the future of a particular data context by learning the pattern in the present available data. The quality and reliability of the predictions made is dependent on the quality of the data used to train the ML algorithms.

Using CRISP-DM Methodology (Matsumoto and Carrinho, 2023)

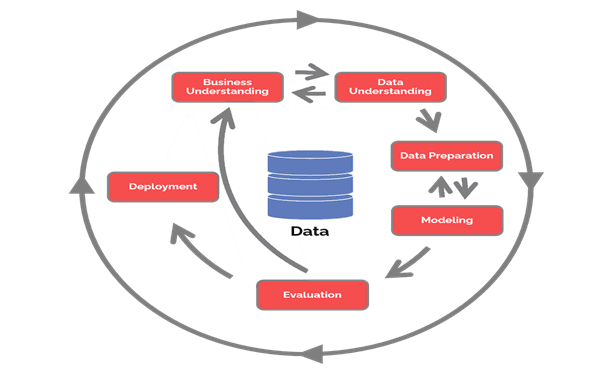


Figure 1: CRISP-DM Life Cycle

# Business/Data Understanding

Detailed in Jupiter notebook.

# Data Preparation

In other to facilitate manipulation of dataset, I imported relevant data preprocessing libraries pandas dataframe and Seaborn and matplotlib for visualisation and NumPy for number and statistical and mathematical manipulation.

A screenshot of a computer code

Description automatically generated

Figure 2: Importing Libraries.

Pandas(pd) has been used in the first line of code to load the 'NHANES\_age\_prediction.csv' into a dataframe. A dataframe is a tabular form, column-oriented data structure, It contains both rows and columns (McKenny, 2017). The columns in a data frame have corresponding names called features/ variables and the rows are the observation of each feature. Pandas dataframe makes working with the data in 'NHANES\_age\_prediction.csv' dataset easy, fast and it also enables the manipulation of this data which is why I converted csv file to dataframe (df).

## Early Data Analysis:

In other to manipulate data and explore data trueness, some methods in the panda’s library were used.

Using the ‘head ()’ and ‘tail ()’ method in pandas, I printed the first and last five rows of df, this enabled me to see the names of each specific column and a snapshot of the observations in them. Using ‘. shape ()’ method, Initial exploration of data size shows there are 2278 rows and 10 columns in df. Looking at the dataframe of respondents from NCHS data collection, I could not understand what each variable meant because the labels have been shorthand, and it makes no logical sent to me looking at it ‘. info ()’ method was used to get the dtypes of the variables and result shows there are 9 numerical and I categorical variable. In other to understand the variable names, I sort the data dictionary from UCI repository. Fig.3 is the data dictionary of the NHANES age-prediction.csv. Shown in the dictionary are the context in which each variable had been collected and the information were enlightening as it gives me information about the numerical observations in the variables.A screenshot of a medical report

Description automatically generated

Figure 3: Data dictionary (UCI repository, 2023).

## Renaming Variables:

Haven gotten a context for the variables I was able to rename the variables by creating a dictionary call new column, new\_column contains the variable name as read in csv and label are the new name I want to assign to each variable. After defining the dict, I reread the dataframe with the new\_column names as shown in fig 4.

A screenshot of a computer

Description automatically generated

Figure 4: Renaming process.

# Statistical exploration:

In other to accurately explore each variable, it is important that the variables are represented correctly in terms of their data type. (Niklas Donges, 2018) reported that having a good understanding of the different data types is a crucial prerequisite for doing Exploratory Data Analysis (EDA). This because having a true understanding of dtype of variables will inform for example the type of visualisation method to use to explore such variable, it will also help us to determine if a variable mathematical meaning or not hence guiding statistical interpretation of such variable.

The ’.describe()’ method was used to obtain the statistical analysis of all numerical variables, but I interpreted the true numerical variables only because the categorical variable numbers have no mathematical value they are categories. In other to statistically evaluate true numerical variables only, I explicitly converted the type of columns [Diabetic, Activity, Gender] into categorical variables because they are categorical and not numerical variables.

Overall, statistical evaluation shows that all the numerical variables are highly skewed, with ‘Age’ been less skewed compared to other numerical variables. Also, the numerical variables have a significantly large range, suggesting possible presence of outliers. It was important that I understand the statistical summary to understand what appropriate tool I need to process this data.

## Visualization:

Visualisation is a tool used to present patterns and finding to enhance understanding at a gaze. Box plot is a good visualisation tool used during EDA. It is used to visualise the distribution the distribution of numerical data and skewness by displaying the data quartiles (or percentiles) and averages (Mcleod, 2019). I plotted the box plots specifically to check for outliers in the variables. Outliers are observations that are numerically distant from the rest of the data as illustrated in fig 5. They are data points that are observed outside the whiskers of the box plot (Zach, 2022).

A diagram of a diagram

Description automatically generated

Figure 5: Box plot illustration of outliers (TrainDataHub, 2021)

In df, ['BMI', 'Glucose\_fasting', 'Oral', 'Blood\_Insulin'] variables showed significant outliers as labled in fig 6. This is also consistent with the large range of this variables that was discovered in the statistical summary.

A diagram of a blood glucose level

Description automatically generated A diagram of a body mass index

Description automatically generated A graph of a graph showing a bar graph

Description automatically generated with medium confidence

A diagram of glucose level

Description automatically generated

Figure 6: Boxplot results of numerical variables showing outliers and skewed distribution.

## Distribution plot:

The histogram plot in EDA is used to see the distribution of the numerical variables, and to visualise how they look under the curve. As shown by the plots fig. 7, all the numerical variables a very skewed. The [mode, median, mean] on a gaussian distribution should be at the centre on the plot. However, on all the numerical variables in df and mean median and mode are apart, this is also consistent with the finding on the statistical variable that show the minimum and the maximum observation in all numerical variables are far apart from each other.A graph of blood glucose distribution

Description automatically generated

A graph showing the age distribution

Description automatically generated A blue line graph with white text

Description automatically generated

A blue line graph with numbers

Description automatically generated

Figure 7: Distribution plots of numerical variables showing highly skewed distribution.

Pair plots show there are there are no linear relationship between the variables. To confirm this relationship visual outcome, I used the 'corr ()' method from pandas, and has shown, the numerical variables have a perfect linear relationship with themselves, which was outcomed as '1' on the diagonal of the correlation matrix table. However, majority of the values a closer to '0' than '1', suggesting a non-linear relationship between them. Variables like [Blood\_Insulin and BMI], [Oral and Glucose\_after\_fasting], showed a small to moderate linear relationship as they are closer to '1' approximately compared to the rest of the numerical variables. The corr () method was able to give me a more precise relationship value compared to what the pair plot visualised.

A table with numbers and symbols

Description automatically generated

Figure 8: Correlation matrix dataframe for numerical variables in 'df'.

## EDA summary:

After my in-depth early and exploratory data analysis, I have found some interesting insights about the dataset that would be the bases for the decisions I will make concerning the data in the next stages of analysis.

## Feature engineering

An important aspect of building a reliable and machine learning model is processing data to a form the model will understand and be less bias towards. Feature engineering is the process taking a raw data, extracting, and organising important feature from it, to fit model operation method (Sreemany, 2021).

I choose to use label encoding because model in this analysis is built to ultimately predict the age-group of a respondent based on their age and live styles for example, it is important for machine learning to be able to differentiate between a 75-year-old and 25-year-old respondent. One Hot encoder I believe is not suited for this task because there is ranking/ order in the age-group, because it is not equal if a person is a senior or an adult respondent. One Hot encoder will also create sparsity in the data, because lots of 0's are created hence creating more noise/complexity for model (Sreemany, 2021).

Following the transformation of the age\_group, a new column of encoded target variable was created {0: Adult, 1: Senior}.

A screenshot of a computer

Description automatically generated

Figure 9: Encoding age\_group variable.

# A screenshot of a computer Description automatically generatedFeature Normalisation:

Data normalisation is the process of structuring data according to normal form (Morris, 2022). We perform normalization a part of preprocessing technique to adjust feature values in a dataset to a common scale (Bhandari, 2020). Both the Min-max and log-transformation scaling techniques were implemented in my analysis because each of them has advantages that I am interested in (detailed In Jupyternote). To bring distribution to a gaussian or close to gaussian distribution, I also implement ed the robust scalar. Fig 10, 11 show results from fitting min-max and robust model in optimal random forest classifier respectively.

A screenshot of a computer screen

Description automatically generated A screenshot of a computer screen

Description automatically generated

Figure 10: Results of Min-max scaled feature on classifying age\_group with random forest.

## Robust scaler results:

Robust scaler brought the highly skewed dataset to gaussian distribution, and in some feature closely A screenshot of a computer

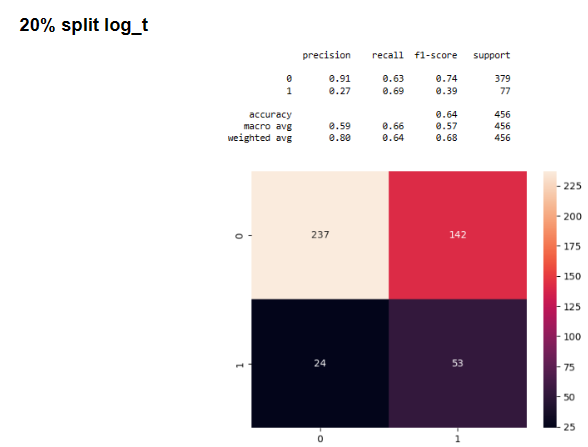
Description automatically generated A screenshot of a computer

Description automatically generated

normal distribution. Fig.11 show the result from fittingrobust scaled features into random forest classifier given the same circumstances.

**Log Transform Result:**

**A screenshot of a computer

Description automatically generated**

## Result:

As seen with min-max model, at testing size of (10, 20,30) %, model had accuracy of 70, 70, 72. The precision result shows that model predicted both classes considerably well at 30% testing size next at 20% and performed the least at classifying senior correctly at 10% testing size. Robust scaler model shows an accuracy of 72 % precision of 74 and 70 for 0 and 1 class respectively. This suggests it classifies the adult respondents correctly much more than the senior respondents. At 20% test, robust average accuracy was 68 and both precision for 0 and 1 classes where ≈ 67.5 average. Log-transformed model at 20% testing had an accuracy of 64% and precision of 91 and 27 for adult and senior respondent respectively. This suggests it did predict 28% correctly of 100% testing size given, suggesting a bad model. At 30% split, accuracy was 68% and precision of for both senior and adult respondents, meaning it predicted them at the same precision.

Overall, based on result we can conclude model fitted with 30% test size of min-max scaled dataset performed better at predicting age-group that other model evaluated.

## PCA and LDA:

These are techniques that are used to reduce dimension the name implies. They generally focus on finding the optimal number of features needed to preserve most of the variance in a dataset. This principle is based on 'HUGHES PHENOMENON'. As shown in fig.11, as the number of features increases, the classifier’s performance increases as well until we reach the optimal number of features, adding more features after this point without increasing the training set causes a downward performance of the classifier.

A graph of a number of features

Description automatically generated

Figure 11: Optimal Dimensionality illustrated.

In comparison, LDA and PCA are both robust dimensionality reduction technique their mode of operation however differs. pca aims to find the directions of maximum variance in the data, while LDA is a supervised method that aims to find the projection that best separates the classes in the data (Vungarala, 2023).

PCA and LDA for classification:

A blue and red dots

Description automatically generated A graph with numbers and lines

Description automatically generated

To the screen left, PCA attempted to classify the same target as LDA did on the right. Pca reduced the dimension of data to essentially what needed to preserve the variance, but class 1 is not distinctively separated from class 0 as it is the case for LDA classification. This was because LDA focuses on separating the classes away from each-other as much as possible.

# MODELLING:

Supervised learning technique involves using a dataset that has both the independent and dependent variable know while unsupervised does not have them defined, in this project I chose to use supervised learning because the variable labels are known and are those The dataset.

Model used:

In this stage of analysis, I would be utilising Random Forest classifier and its hyper parameter to produce a robust and optimal model that will classify the age-group of respondents given X variables. Random Forest classifier is supervised classification method. Decision Tree is collection of decision nodes, connected by branches, extending downward from root node to terminating leaf nodes. Beginning with root node, attributes tested at decision nodes, and each possible outcome results in branch. Each branch leads to decision node or leaf node (Random Forest, 2023). Machine learning deals with numbers only and to use RFC, the target variable must be in categorical format, this was why I encoded ‘age-group’ variable.

Importing libraries:

A screenshot of a computer

Description automatically generated

## Fitting Model:

In other to fit the RFC, we need to define the independent and dependent variable. Fig 12 shows the feature selections in the independent variable using the ‘log\_df’.

A screenshot of a computer

Description automatically generated

Figure 12: Independent variable feature selection.

Dependent variable is the encoded ‘age\_group’ as shown in fig.13.



Figure 13: Defining dependent variable.

As shown above my first attempt at classifying the age\_group of respondents, output an overfitted model. Overfitting is When machine learning model learns the details and noise in the training data such that it negatively affects the performance of the model on test data (Biswal, 2023). In the first attempt at 20% test size, I noticed model was not making any prediction it was only giving 100% the testing back. Researching from different authors, I found a concept “data leakage”. Prediction have been made available to model unintentionally in training data. To fix this, I will drop the age variable from the independent variable and refit my model and result is seen in fig. 15.

A screenshot of a computer screen

Description automatically generated

Figure 14: Overfitted model.

A screenshot of a computer

Description automatically generated A graph of a number

Description automatically generated with medium confidence

Figure 15: 30% and 20 % testing size after dropping age.

As shown in fig.15, the diagonal show that the model classified 364 adult respondents correctly and 15 incorrectly out of 379 testing size. Model classified 10 senior respondents correctly and 67 incorrectly out of 77 testing size. Overall testing size was (456 and 684) at 20% and 30% testing, out of which 379 and 77 are (0 and 1) class respectively at 20% testing and 556 and 107 are (0 and 1) class respectively at 30% testing. The model though had a good accuracy of 82% and 84% at (80 and 70) % training, it did perform better classifying adult respondent than senior respondent, this was because the target variable is imbalanced, there were more training data for adult than senior. The f-1 score can be used to asses model performance when classes are unbalanced, the f-1 score of 0 is very good while that of 1 is low, recall as well follows the same patter suggesting model is not performing well at classifying 1

Handling Imbalance:

I did a value count to see the proportion of each class in the target variable. As shown in output, result shows there are 364 senior respondents, approximately 5 times less than the adult class. This creates an imbalance in the data as illustrated below.

A blue and orange rectangles

Description automatically generated

Figure 16: Imbalance data illustration.

Synthetic Minority Oversampling Technique (SMOTE) is a technique in modelling that helps to create synthetic to balance the lesser portion or reduce the higher portion down the oversampling method was used because I did not want to reduce any variance by reducing sampling. After fitting the SMOTE model, artificial Null Values have been created in the 'Gender' and 'activity' columns of X this is because SMOTE are designed to traditionally handle numerical features (SATPATHY, 2020), I handled this problem by converting the two variables to numerical.

A screenshot of a test results

Description automatically generated

Figure 17: Result after balancing dataset and splitting at (20 & 30) %.

After balancing the dataset and removing the age column, the precision of the senior did increase but the overall accuracy of the dropped. The recall and f-1 score of 1 increased significantly compared to when the data was imbalanced suggesting a better performing model after balancing. The recall score means the model is now better at classifying true positives.

# Hyperparameter tunning:

Hyperparameters are built in parameters that when used can aid us in improving model performance. After two trials of tunning n-estimator at 3 max-depth, the accuracy of my model dropped significantly, in this next step of trail of building an optimal model to predict the age-group of respondents, I will try to find the optimum values for max\_depth and understand how the value of max\_depth impacts the overall accuracy of the ensemble. The overall cross validation score on both model was good suggesting a good performing model.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

## Max- Depth tunning:

As shown in ‘max\_ddf’, the max depths iterated are 2, 7, 12, 17 and generally as they increased, the mean\_test\_score should also increase because model will become more complex hence a better performing model. However, compared to the increment seen in max\_depth from 7 to 12, result show there are diminishing in the increment of mean\_test\_score of max\_depth 12 to 17. std\_test\_score also increased with max\_depth increament, indicating a wider variability in model performance. Overall, result shows max-depth 17 yielded the best mean\_test\_score ≈ 0.86, followed by 12, 7 and 2 with average of ≈ (0.85, 0.79, 0.71) respectively. This mean that with max-depth of 17 will give more accuracy at predicting age-group of respondents giving X as shown in the fig.18.

A graph with a line

Description automatically generated

Figure 18: Max-depth graph and accuracy.

## Finding Important Feature using Scikit-learn:

We are finding important features or selecting features in the Age-group prediction dataset. This step is part of an attempt to build an optimal random forest model that helps to predict the age class a respondent will belong given independent variables such as their life-style and activities.

A screenshot of a computer program

Description automatically generated

Figure 19: Feature selection df

Feature\_imp dataframe contains the results from importance feature selection to classify age\_group of respondents in dependent variable using pattern in independent variable with random forest classifier (Clf\_RF). Result in fig.19 shows, feature 4, 6, 3, 5 shows higher score with corresponding score of (0.28, 0.23, 0.22, 0.21) respectively in descending other. Feature 1, 0,2 have lesser importance according to result. This in conclusion shows feature 4, 6, 3,5 plays a more significant role at classifying the age-group class than features 1,0,2.

## Grid Search to Find Optimal Hyperparameters:

GridSearchCV is a technique used for hyperparameter optimization (Shah, 2021). I implemented it because it is a methodical way of iterating over various hyperparameter tunning at the same time, to find the optimal on for training and testing RFC model. After the GridSearchCV, result in fig. 20 shows the best accuracy we can get with this model is 71% using (max\_depth=8, min\_samples\_leaf=100, min\_samples\_split=200, max\_features=10, n\_estimators=300).

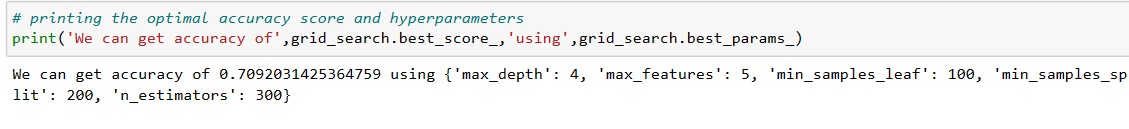


Figure 20: GridSearchCV parameter search

## CONCLUSION:

In conclusion, result of model prediction shows min-max scaler was the best scaling technique to use for age-predict data, the model precision increased after balancing the data and at 30%, with X variable, we can get an accuracy of 71% using the parameter given and min-max scaling of data. LDA classified the target variable better than PCA overall.

# Reference list

archive.ics.uci.edu. (2023). *UCI Machine Learning Repository*. [online] Available at: https://archive.ics.uci.edu/dataset/887/national+health+and+nutrition+health+survey+2013-2014+(nhanes [Accessed 1 Dec. 2023].

Biswal, A. (2023). *The Complete Guide on Overfitting and Underfitting in Machine Learning*. [online] Simplilearn.com. Available at: https://www.simplilearn.com/tutorials/machine-learning-tutorial/overfitting-and-underfitting#what\_is\_overfitting [Accessed 29 Dec. 2023].

Huilgol, P. (2020). *9 Feature Transformation & Scaling Techniques| Boost Model Performance*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2020/07/types-of-feature-transformation-and-scaling/#:~:text=The%20Log%20Transform%20is%20one [Accessed 28 Dec. 2023].

Linda, T. (2021). *What Is Machine Learning and Why Is It Important?* [online] SearchEnterpriseAI. Available at: https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML [Accessed 1 Dec. 2023].

Mcleod, S. (2019). *What Does a Box Plot Tell You?* [online] Simplypsychology.org. Available at: https://www.simplypsychology.org/boxplots.html [Accessed 25 Dec. 2023].

Mohajon, J. (2020). *Confusion Matrix for Your Multi-Class Machine Learning Model*. [online] Medium. Available at: https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826 [Accessed 29 Dec. 2023].

NA,NA. (2023). National Health and Nutrition Health Survey 2013-2014 (NHANES) Age Prediction Subset. UCI Machine Learning Repository. https://doi.org/10.24432/C5BS66.

Niklas Donges (2018). *Data Types in Statistics*. [online] Medium. Available at: https://towardsdatascience.com/data-types-in-statistics-347e152e8bee [Accessed 22 Dec. 2023].

SATPATHY, S. (2020). *SMOTE - A Common Technique to Overcome Class Imbalance Problem*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/ [Accessed 30 Dec. 2023].

Saxena, S. (2020). Here’s All you Need to Know About Encoding Categorical Data (with Python code). *Analytics Vidhya*. [online] 13 Aug. Available at: https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/ [Accessed 27 Dec. 2023].

Scikit-learn (2018). *sklearn.ensemble.RandomForestClassifier — scikit-learn 0.20.3 documentation*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html [Accessed 1 Jan. 2024].

Shah, R. (2021). *GridSearchCV |Tune Hyperparameters with GridSearchCV*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/06/tune-hyperparameters-with-gridsearchcv/ [Accessed 1 Jan. 2024].

Sreemany, T. (2021). *Introduction to Feature Engineering - Everything You Need to Know!* [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/10/a-beginners-guide-to-feature-engineering-everything-you-need-to-know/#h-what-is-feature-engineering [Accessed 27 Dec. 2023].

TrainDataHub (2021). *Everything You Need to Know About Boxplot and How to Construct Multiple Boxplots In Python*. [online] Medium. Available at: https://medium.com/@ooemma83/everything-you-need-to-know-about-boxplot-and-how-to-construct-multiple-boxplots-in-python-7726a1920cc [Accessed 26 Dec. 2023].

UCI reosatory (2023). *UCI Machine Learning Repository*. [online] archive.ics.uci.edu. Available at: https://archive.ics.uci.edu/dataset/887/national+health+and+nutrition+health+survey+2013-2014+(nhanes)+age+prediction+subset [Accessed 20 Dec. 2023].

Vungarala, S.K. (2023). *PCA vs LDA — No more confusion!* [online] Medium. Available at: https://medium.com/@seshu8hachi/pca-vs-lda-no-more-confusion-fc21fb8d06e9#:~:text=PCA%20is%20an%20unsupervised%20method%20that%20aims%20to%20find%20the [Accessed 28 Dec. 2023].

Zach (2022). *How to Read a Box Plot with Outliers (With Example)*. [online] Statology. Available at: https://www.statology.org/how-to-read-box-plot-with-outliers/ [Accessed 26 Dec. 2023].