**Hasan Enes Guray**

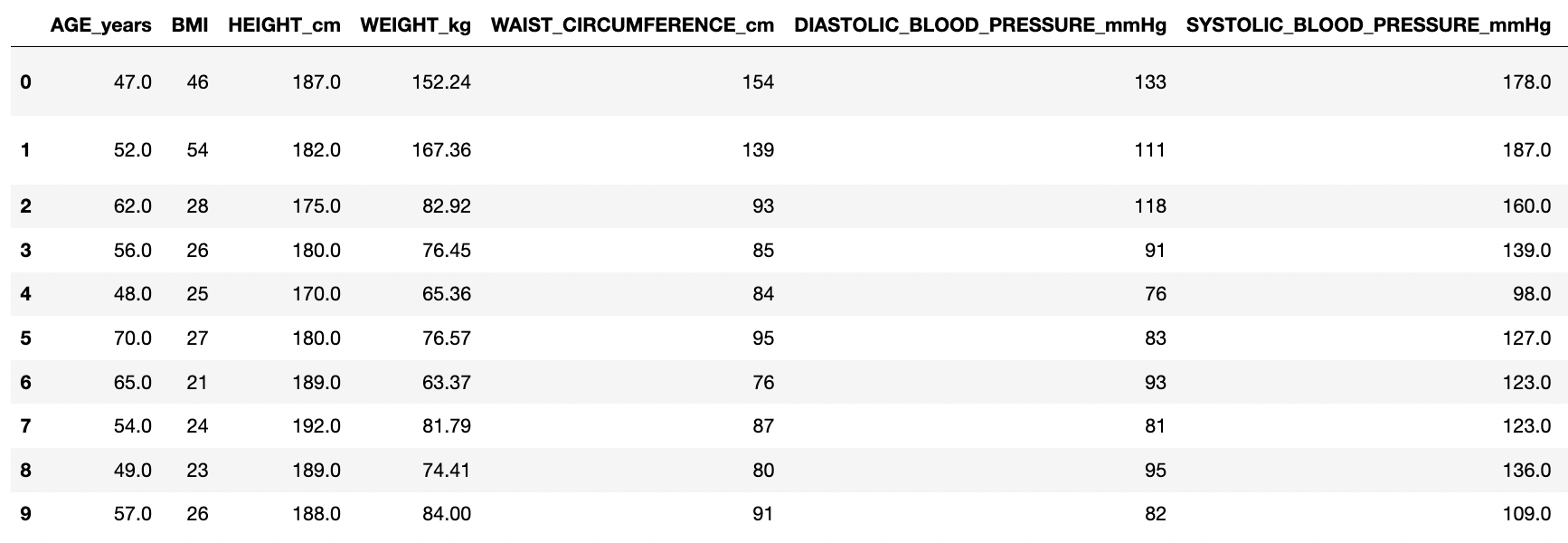
**19489124**

**Task (1) – Domain Understanding: Classification or Regression**

1. Given that the task pertains to the predicting classes of discrete variables, it appears rational to pursue predictive classification modelling.

**Task (2) – Data Understanding: Producing Your Experimental Designing**

1. Male dataset, 1. Question, multiclass classification problem.
2. Dataframe:

Table

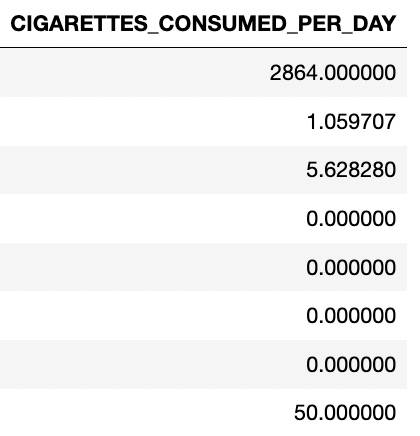
Description automatically generated

Statistical description:

Graphical user interface, application

Description automatically generated with medium confidenceTable

Description automatically generatedTable

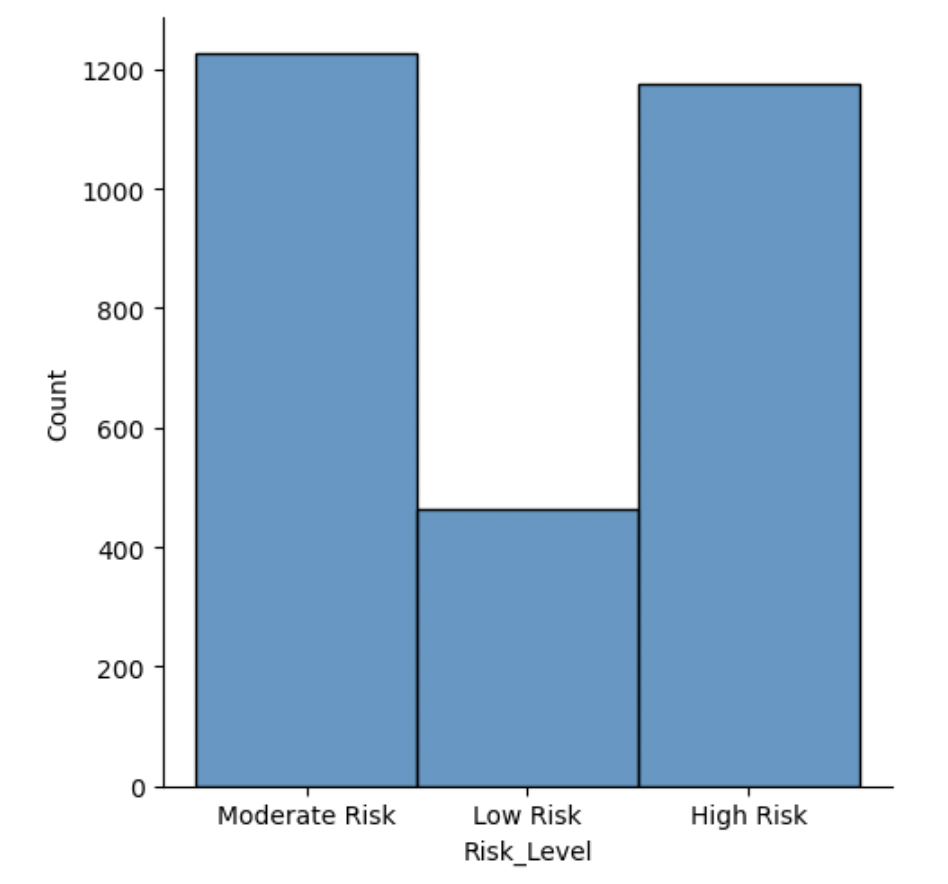
Description automatically generated

Measurement scale type:

Text, table

Description automatically generated with medium confidence

Distribution of the class variable(barplot and count percentages):

A picture containing text

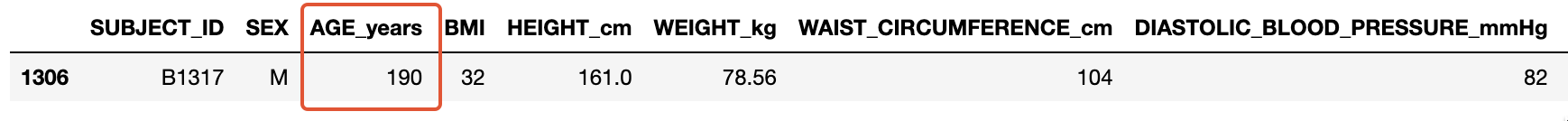
Description automatically generated

**Task (3) – Data Preparation: Cleaning and Transforming your data**

|  |  |  |
| --- | --- | --- |
| Dataset or Variable | Name of variable | Issue description |
| Variable | AGE\_years | There is a 190-year-old person, which is impossible. |
| Variable | BMI | The column name has an unnecessary space. |
| Variable | HEIGHT\_cm | There are 2 people, who are 1.8 and 1.7 cm, which is impossible. |
| Variable | HEIGHT\_cm | There are outliers. |
| Variable | SYSTOLIC\_BLOOD\_PRESSURE\_mmHg | It contains 5 missing values. |
| Variable | COMPUTER\_USE\_TIME\_PER\_DAY\_HOURS | Some values are larger than 24 hours. |
| Variable | SMOKING\_STATUS | There are records, in which smoke consumption is a positive integer, but seem nonsmokers. The column name has an unnecessary space. |
| Variable | CIGARETTES\_CONSUMED\_PER\_DAY | There are missing values for nonsmokers. |
| Variable | DISCONTINUED\_NO\_ | The meaning is unknown and there are just 2 non-missing values. |
| Variable | Visceral\_Fat\_Volume\_Litres | There are negative values. |
| Variable | Visceral\_Fat\_Volume\_Litres | There are outliers. |
| Dataset |  | The variables have different magnitudes. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset or Variable | Name of variable | The Issue | Solution | Justification |
| Variable | AGE\_years | Containing unlogical value | Dropping it | This is an error value. |
| Variable | BMI | Unnecessary space in the column name | Removing the space | Inadvertent or extraneous spaces in column names or datasets can lead to errors and inconsistencies during model execution, leading to suboptimal performance and inaccurate outputs. (Dekanovsky, 2020) |
| Variable | HEIGHT\_cm | Containing unlogical values | Dropping them | This is an error value. |
| Variable | HEIGHT\_cm | Containing outliers | Dropping them | Outliers are observations in a dataset that stand out significantly from the majority of data and could be indicative of errors or irregularities during data collection or processing. Outliers pose an enormous risk, potentially leading to inaccurate outcomes. (Castillo, 2021) |
| Variable | SYSTOLIC\_BLOOD\_PRESSURE\_mmHg | Containing missing values | Replacing them with the mean value | Given the relatively symmetric nature of this distribution, mean value replacement may be an appropriate approach for handling missing data. This technique works best when the distribution is unimodal with minimal skewness or kurtosis. (Kumar, 2021) |
| Variable | COMPUTER\_USE\_TIME\_PER\_DAY\_HOURS | Containing unlogical values | Dropping them | This is an error value. |
| Variable | SMOKING\_STATUS | Containing unlogical values and unnecessary space in the column name | Replacing the missing and 0 values with 1, if there is a cigarette consumption | These are error values. Computer programs are case-sensitive, so an accidental space within column names can lead to unexpected errors in a model. (Dekanovsky, 2020) |
| Variable | CIGARETTES\_CONSUMED\_PER\_DAY | Containing unlogical values | Replacing the null values with 0, if the person is a nonsmoker | These are error values. |
| Variable | DISCONTINUED\_NO\_ | Containing missing values | Dropping the column | There is 99% missing data. |
| Variable | Visceral\_Fat\_Volume\_Litres | Containing unlogical values | Dropping them | These are error values. |
| Variable | Visceral\_Fat\_Volume\_Litres | Containing outliers | Dropping them | Outlier data points present in an input dataset may cause bias and variance to shift, as well as impair machine learning models' learning process. (Brownlee, 2013) |
| Dataset |  | Having different magnitudes | Standardize them | Variables with varying magnitudes contribute to the model in different proportions. (Loukas, 2020) |

1. AGE\_years:

Before:

After:

BMI:

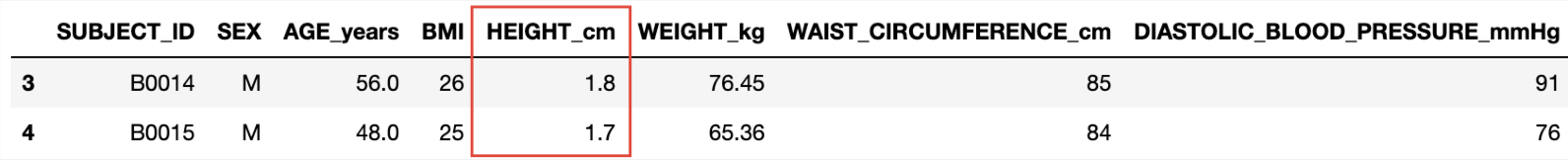
Before: Text

Description automatically generated

After: Text

Description automatically generated

HEIGHT\_cm(unlogical values):

Before: 

After: 

HEIGHT\_cm(outliers):

Before: Chart, box and whisker chart

Description automatically generated

After: Chart, histogram

Description automatically generated

SYSTOLIC\_BLOOD\_PRESSURE\_mmHg:

Before: Table

Description automatically generated

After: Table

Description automatically generated

COMPUTER\_USE\_TIME\_PER\_DAY\_HOURS:

Before: A picture containing text

Description automatically generated

After: 

SMOKING\_STATUS:

Before: Background pattern

Description automatically generated with medium confidence

After: Background pattern

Description automatically generated with low confidence

CIGARETTES\_CONSUMED\_PER\_DAY:

Before: A picture containing application

Description automatically generated

After: A picture containing background pattern

Description automatically generated

DISCONTINUED\_NO\_:

Before: Table

Description automatically generated

After: Table

Description automatically generated

Visceral\_Fat\_Volume\_Litres(unlogical values):

Before: A picture containing graphical user interface

Description automatically generated

After: 

Visceral\_Fat\_Volume\_Litres(outliers):

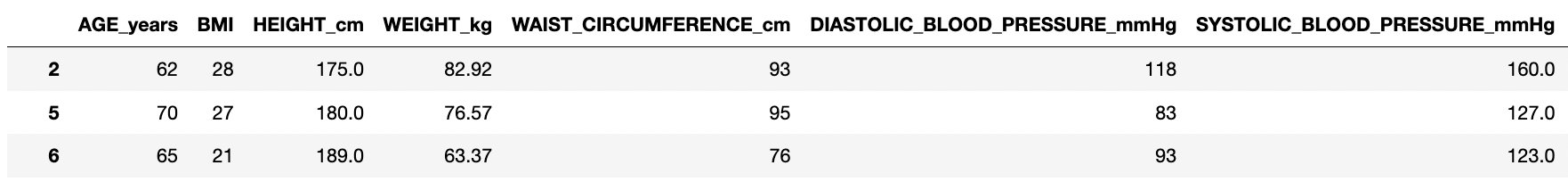
Before: Chart

Description automatically generated

After: Chart, histogram

Description automatically generated

Dataset:

Before: 

After: Table

Description automatically generated

**Task (4) – Modelling: Create Predictive Classification Models**

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm Name | Type of Algorithm | Possible Hyperparameters | Python package source code to call the algorithm |
| NB | parametric | priors, var\_smoothing | from sklearn.naive\_bayes import GaussianNB |
| DT | non-parametric | criterion, splitter, max\_depth, min\_samples\_split, min\_samples\_leaf, min\_weight\_fraction\_leaf, max\_features, random\_state, max\_leaf\_nodes, min\_impurity\_decrease, class\_weight, ccp\_alpha | from sklearn.tree import DecisionTreeClassifier |
| KNN | non-parametric | n\_neighbors, weights, algorithm, leaf\_size, p, metric, metric\_params, n\_jobs | from sklearn.neighbors import KNeighborsClassifier |
| ANN(MLP) | parametric | hidden\_layer\_size, activation, solver, alpha, batch\_size, learning\_rate, learning\_rate\_init, power\_t, max\_iter, shuffle, random\_state, tol, verbose, warm\_start, momentum, nesterovs\_momentum, early\_stopping, validation\_fraction, beta\_1, beta\_2, epsilon, n\_iter\_no\_change, max\_fun | from sklearn.neural\_network import MLPClassifier |

1. X\_train: Table

   Description automatically generated

Y\_train: Graphical user interface, text, application

Description automatically generated

The Pareto Principle, more commonly referred to as the 80-20 rule, has gained significant traction within business and data analytics due to its demonstrated success at identifying and prioritizing critical elements within complex datasets. Numerous empirical studies have attested to its utility across various business domains, further cementing its widespread adoption as a preferred approach for data analysis and modeling. (Tardi, 2023)

KNN: 

GaussianNB: 

Decision Tree: 

ANN(MLP): 

**Task (5) – Evaluation: How good are your models**

1. Chart

   Description automatically generated Chart, treemap chart

   Description automatically generated Chart, treemap chart

   Description automatically generated Chart

   Description automatically generated

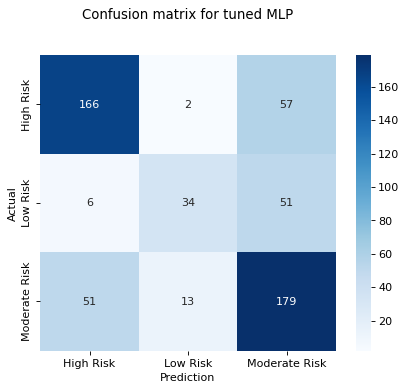
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric Name | Related or Unrelated | Justification in relation to the success criteria | Model Name | Metric Score |
| Accuracy | related | When testing the model's success at accurately predicting risk groups, one important and correlated indicator to evaluate its performance is the percentage of correct predictions across all samples. This score allows us to measure how accurately records in each risk group are estimated. Unfortunately, this score cannot account for individual estimation accuracy, necessitating further analysis using other metrics. | NB | 0.63 |
| DT | 0.51 |
| KNN | 0.56 |
| ANN | 0.66 |
| Recall | related | Healthcare professionals should prioritize identifying individuals classified as having high or moderate risks. While additional testing for some low-risk patients may result in additional expenses, misclassifying those at greater risk as healthy and dispensing them without treatment is far more concerning. Therefore, recall metrics play a vital role here since they increase the number of accurate predictions made. | NB | High Risk: 0.66 Moderate Risk: 0.63 Low Risk: 0.55 |
| DT | High Risk: 0.58 Moderate Risk: 0.48 Low Risk: 0.38 |
| KNN | High Risk: 0.72 Moderate Risk: 0.49 Low Risk: 0.37 |
| ANN | High Risk: 0.74 Moderate Risk: 0.72 Low Risk: 0.31 |
| Precision | related | The precision score suggests greater precision when making predictions within that risk group. Recall metrics are of primary importance, but reducing the number of individuals mistakenly classified as a high or moderate risk when they are Low risk can reduce the need for additional testing and provide cost savings by eliminating unnecessary testing. Though less critical than Recall, Precision is essential for assessing model success by measuring correct prediction probability. Therefore, it should be utilized in creating a more effective, successful, and comprehensive risk prediction model. | NB | High Risk: 0.76 Moderate Risk: 0.59 Low Risk: 0.49 |
| DT | High Risk: 0.60 Moderate Risk: 0.48 Low Risk: 0.36 |
| KNN | High Risk: 0.62 Moderate Risk: 0.53 Low Risk: 0.44 |
| ANN | High Risk: 0.72 Moderate Risk: 0.60 Low Risk: 0.67 |
| F-Measure | related | The F-Measure is the harmonic mean of Precision and Recall metrics, which helps to balance these measures when dealing with unbalanced classification problems, providing more precise outcomes. In contrast, not a primary metric in risk prediction models, striving for a maximum F-1 score is essential for developing an efficient yet well-balanced model. | NB | High Risk: 0.71 Moderate Risk: 0.61 Low Risk: 0.52 |
| DT | High Risk: 0.59 Moderate Risk: 0.48 Low Risk: 0.37 |
| KNN | High Risk: 0.66 Moderate Risk: 0.51 Low Risk: 0.40 |
| ANN | High Risk: 0.73 Moderate Risk: 0.66 Low Risk: 0.42 |
| AUC-ROC | related | A high ROC-AUC score indicates a greater probability of accurately identifying positive samples from a targeted class. However, this metric may not be enough when imbalances exist between classes like Low-Risk in the model; thus, class-specific scores must take priority over overall ROC-AUC score and efforts should be made to attain the highest possible probabilities. A higher ROC-AUC score implies a more precise classification between classes. | NB | High Risk: 0.85 Moderate Risk: 0.83 Low Risk: 0.70 |
| DT | High Risk: 0.66 Moderate Risk: 0.63 Low Risk: 0.54 |
| KNN | High Risk: 0.78 Moderate Risk: 0.73 Low Risk: 0.61 |
| ANN | High Risk: 0.85 Moderate Risk: 0.83 Low Risk: 0.71 |

1. Our model's primary objective is to accurately predict patients in High and Moderate risk groups. To do this, the recall score should be considered which indicates how accurately each group was predicted individually; the ANN model performed better than all others when taking into account this factor. Secondly, accuracy value - which measures how accurately all risk groups are predicted - also needs to be considered; again, the ANN scored higher here too. Considering all these, it can be concluded that the ANN model offers greater accuracy than other options and thus offers healthcare professionals more confidence when using it as their go-to tool for patient prediction needs.
2. The tuning process was carried out in 4 steps, as the capacity of the computer was not enough to examine all hyperparameters at once with GridSearchCV. First of all, the random\_state value, which creates the least error with the error rate graph, was selected as optimum, and at this stage, the max\_iter hyperparameters were decided as 1000 in order to get as few converge warnings as possible, while the capacity of the computer could be handled properly. Afterward, GridSearchCV was trained with activation, hidden\_layer\_sizes, solver, and learning\_rate hyperparameters, and optimum values ​​were obtained. Then, using the optimum values ​​as constants, GridSearchCV was trained a second time with alpha, batch\_size, shuffle, tol, and early\_stopping hyperparameters, and optimum results were obtained for these parameters. Finally, the warm\_start, verbose, beta\_1, beta\_2, epsilon, and n\_iter\_no\_change parameters were also trained and hyperparameter tuning was completed. 10 cross-validation K-folds were used on all GridSearchCVs. The hyperparameters values ​​we use in our newly created model are as follows:

random\_state=811, max\_iter=1000, activation="relu", hidden\_layer\_sizes=(100,), solver= 'adam', alpha=0.0001, batch\_size=200, early\_stopping=False, shuffle= True, tol= 0.0001, beta\_1=0.9, beta\_2=0.999, epsilon=1e-8, n\_iter\_no\_change=10, verbose=True, warm\_start=True

Since 'sgd' is not used as the solver, learning\_rate, power\_t, nesterovs\_momentum, and momentum were not used as parameters. Since ‘lbfgs’ is not used as the solver, max\_fun was not used as a parameter. Lastly, validation\_fraction was not used, as early\_stopping is not True.

|  |  |
| --- | --- |
| Accuracy | 0.68 |
| Recall | High Risk: 0.74 Moderate Risk: 0.74 Low Risk: 0.37 |
| Precision | High Risk: 0.74 Moderate Risk: 0.62 Low Risk: 0.69 |
| F-Measure | High Risk: 0.74 Moderate Risk: 0.68 Low Risk: 0.49 |
| AUC-ROC | High Risk: 0.85 Moderate Risk: 0.84 Low Risk: 0.71 |



After the Hyperparameter tuning procedure, additional 5 patients in the Moderate Risk group and 6 patients in the Low Risk group were accurately predicted. In particular, the random\_state parameter was decisive in increasing the scores. In general, Accuracy increased by 0.2, while the Recall value increased by 0.2 in Moderate Risk and 0.6 in Low Risk. The AUC-ROC score, which shows how successfully it is differentiated from other risk groups, showed an increase of 0.1 for the Moderate Risk group. Thus, after the hyperparameter tuning process, the scores of the ANN model increased, although not much, compared to the original model.

1. Machine learning algorithms have revolutionized the field of healthcare, giving healthcare professionals unprecedented insights into patient data. The Artificial Neural Network (ANN) Multi-Layer Perceptron (MLP) classification model has shown great promise in accurately predicting the risk of visceral fat accumulation, a major contributor to numerous health issues. Analyzing historical data has demonstrated that the ANN (MLP) classification model is more accurate and reliable than other models at predicting visceral fat accumulation, both in terms of recall and accuracy scores. In fact, this model has been shown to accurately detect 74% of patients in high and moderate risk groups with 74% accuracy and predict the risks for all groups with an accuracy rate of 68%, all without needing MRI scans. As with any machine learning algorithm, the accuracy and effectiveness of an ANN (MLP) classification model will improve with more data and processing power. In real-world scenarios, the model has the potential to further refine itself and provide even more precise predictions for patients at risk of visceral fat accumulation. This has the potential to significantly improve patient outcomes while alleviating strain on healthcare systems.

**References**

Kumar, A. (2021) Python – Replace Missing Values with Mean, Median & Mode, Vitaflux. Available at: https://vitalflux.com/pandas-impute-missing-values-mean-median-mode/ (Accessed: March 7, 2023).

Dekanovsky, V. (2020) Dealing with extra white spaces while reading CSV in Pandas, Towards Data Science. Medium. Available at: https://towardsdatascience.com/dealing-with-extra-white-spaces-while-reading-csv-in-pandas-67b0c2b71e6a (Accessed: March 7, 2023).

Loukas, P.D.S. (2020) How and why to Standardize your data: A python tutorial, Towards Data Science. Medium. Available at: https://towardsdatascience.com/how-and-why-to-standardize-your-data-996926c2c832 (Accessed: March 7, 2023).

Tardi, C. (2023) *The 80-20 Rule (aka Pareto Principle): What It Is, How It Works, Investopedia. Investopedia.* Available at: https://www.investopedia.com/terms/1/80-20-rule.asp (Accessed: March 7, 2023).

Brownlee, J. (2013) How to Identify Outliers in your Data, Machine Learning Mastery. Available at: https://machinelearningmastery.com/how-to-identify-outliers-in-your-data/ (Accessed: March 12, 2023).

Castillo, D. (2021) *Outlier Detection and Analysis Methods, Seldon*. Available at: https://www.seldon.io/outlier-detection-and-analysis-methods (Accessed: March 15, 2023).

sklearn.naive\_bayes.GaussianNB (no date) scikit-learn. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.naive\_bayes.GaussianNB.html (Accessed: March 15, 2023).

sklearn.tree.DecisionTreeClassifier (no date) scikit-learn. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html (Accessed: March 15, 2023).

sklearn.neighbors.KNeighborsClassifier (no date) scikit-learn. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html (Accessed: March 15, 2023).

sklearn.neural\_network.MLPClassifier (no date) scikit. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.neural\_network.MLPClassifier.html (Accessed: March 23, 2023).

Frost, J. (2017) *Multicollinearity in Regression Analysis: Problems, Detection, and Solutions, Statistics By Jim.* Available at: https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/ (Accessed: March 22, 2023).

Toth, G. (2020) *Multicollinearity, DataSklr.* DataSklr. Available at: https://www.datasklr.com/ols-least-squares-regression/multicollinearity (Accessed: March 22, 2023).

Tatler, J., Cassey, P. and Prowse, T.A.A. (2018) *High accuracy at low frequency: Detailed behavioural classification from Accelerometer Data, The Company of Biologists.* The Company of Biologists. Available at: https://journals.biologists.com/jeb/article/221/23/jeb184085/20503/High-accuracy-at-low-frequency-detailed (Accessed: April 3, 2023).