**Hasan Enes Guray**

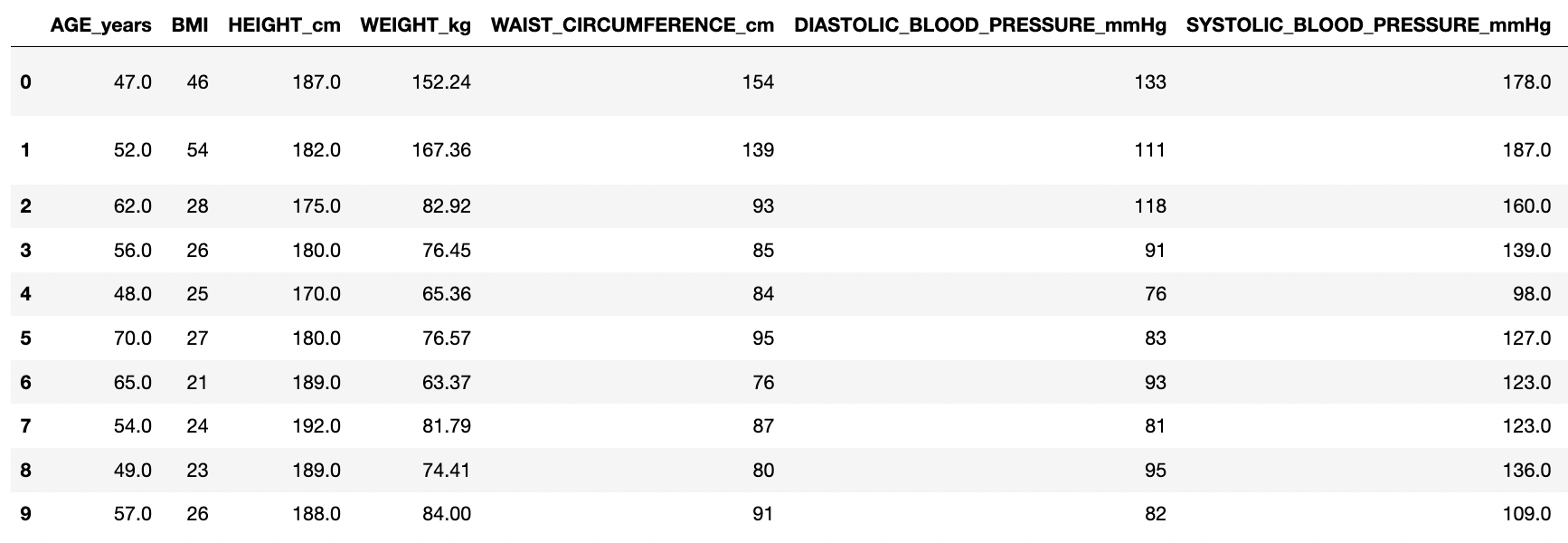
**19489124**

**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 modeling.

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

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

Table

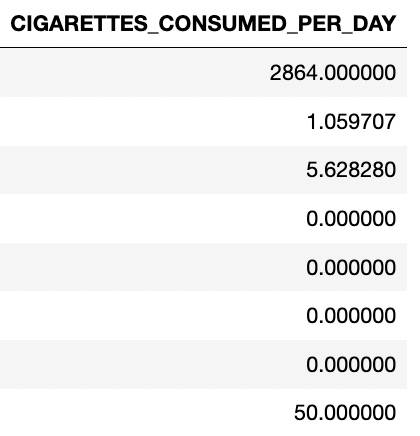
Description automatically generated

Statistical description:

Graphical user interface, application

Description automatically generated with medium confidenceTable

Description automatically generatedTable

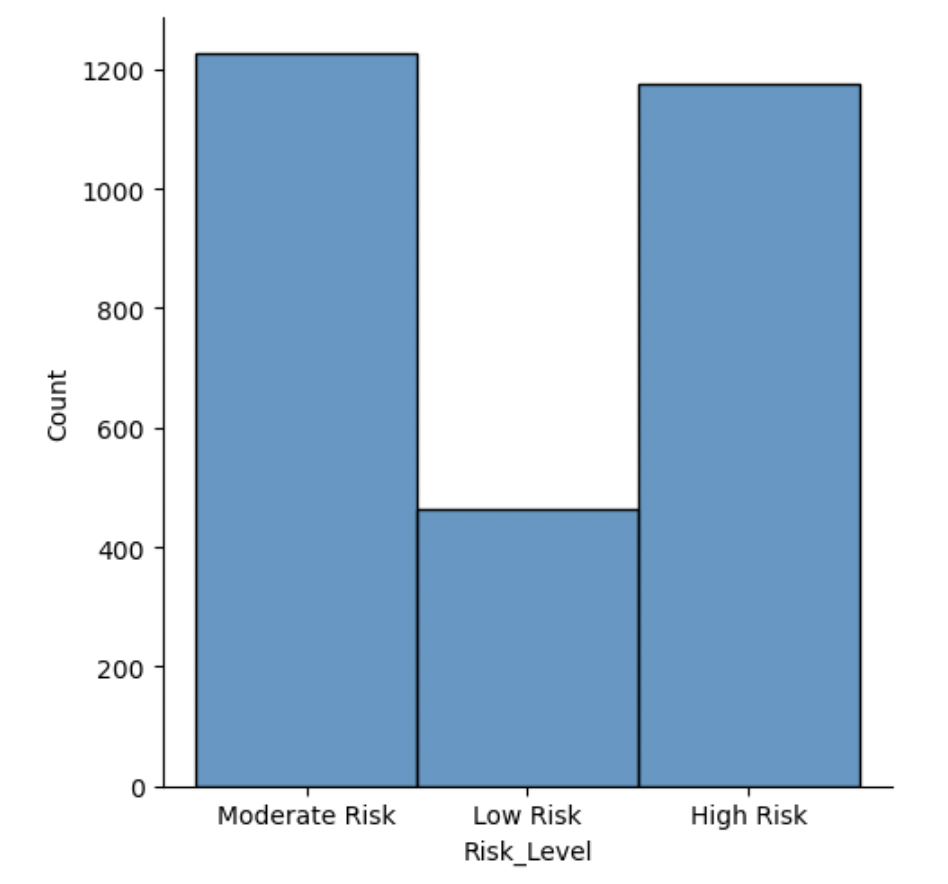
Description automatically generated

Measurement scale type:

Text, table

Description automatically generated with medium confidence

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

A picture containing text

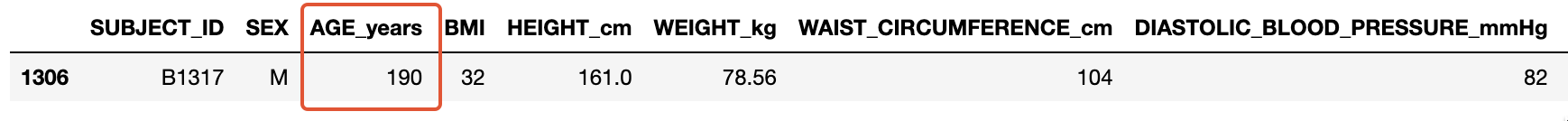
Description automatically generated

**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 unnecessary space. |
| Variable | HEIGHT\_cm | There are 2 persons, who are 1.8 and 1.7 cm, which is impossible. |
| Variable | SYSTOLIC\_BLOOD\_PRESSURE\_mmHg | It contains 5 missing values. |
| Variable | COMPUTER\_USE\_TIME\_PER\_DAY\_HOURS | There are values, which is larger than 24 hours. |
| Variable | SMOKING\_STATUS | There are records, in which smoke consumption is larger than 0, but seems nonsmokers. The column name has 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. |
|  |  | Other variables’ outliers will be added to increase the module with the highest |
| 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 | The presence of inadvertent or extraneous spaces within column names or datasets can trigger unintended errors or inconsistencies in the execution of the models, leading to suboptimal performance and the potential for erroneous outputs. (Dekanovsky, 2020) |
| Variable | HEIGHT\_cm | Containing unlogical values | Dropping them | This is an error value. |
| Variable | SYSTOLIC\_BLOOD\_PRESSURE\_mmHg | Containing missing values | Replacing them with the mean value | With the relatively symmetric nature of the distribution under consideration, it may be appropriate to employ mean value replacement as a more suitable approach for handling missing data. This approach can prove particularly useful in scenarios where the distribution is unimodal and displays 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. The computer programs are case-sensitive. The presence of inadvertent spaces within column names can trigger unintended errors in the 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 the input dataset have the potential to unbalance bias and variance, and also confound the learning process of machine learning models. (Brownlee, 2013) |
|  |  | Other variables’ outliers will be added to increase the module with the highest |  |  |
| Dataset |  | Having different magnitudes | Standardize them | Variables with different magnitudes contribute to the model in different proportions. (Loukas, 2020) |

1. AGE\_years:

Before:

After:

BMI:

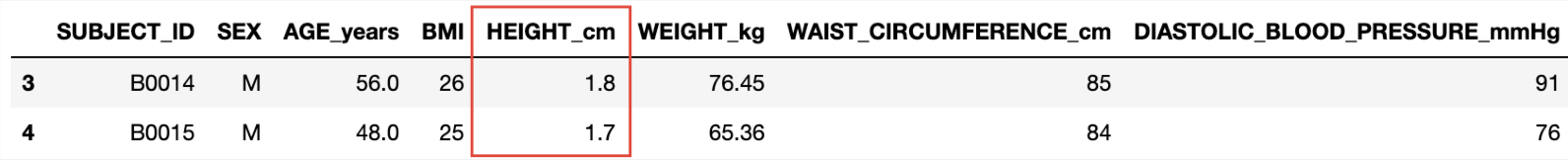
Before: Text

Description automatically generated

After: Text

Description automatically generated

HEIGHT\_cm:

Before: 

After: 

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

Chart

Description automatically generated

After: 

Chart, histogram

Description automatically generated

Visceral\_Fat\_Volume\_Litres(outliers):

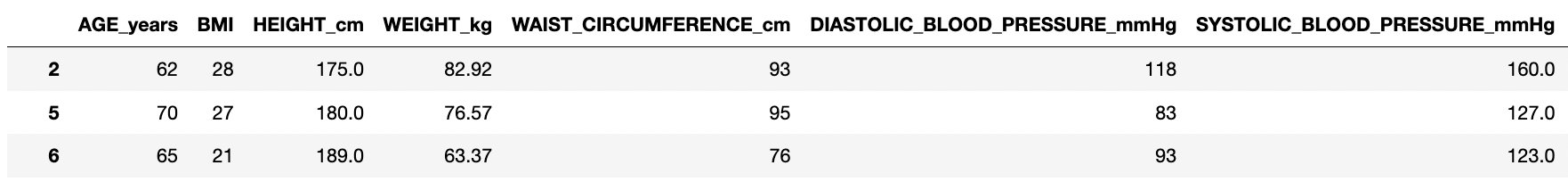
Before: Chart

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 | - | from sklearn.naive\_bayes import GaussianNB |
| DT | non-parametric | max\_depth, criterion, splitter, min\_samples\_split, min\_samples\_leaf, min\_weight\_fraction\_leaf, max\_features, random\_state, max\_leaf\_nodes, min\_impurity\_decrease, ccp\_alpha | from sklearn.tree import DecisionTreeClassifier |
| KNN | non-parametric | algorithm, leaf\_size, metric, n\_neighbors, p, weights | from sklearn.neighbors import KNeighborsClassifier |
| ANN(MLP) | non-parametric |  |  |

1. X\_train: Table

   Description automatically generated

Y\_train: Graphical user interface, text, application

Description automatically generated

The Pareto Principle, also commonly referred to as the 80-20 rule, has gained significant traction within the realm of business and data analytics due to its demonstrated effectiveness in identifying and prioritizing critical factors within complex datasets. Numerous empirical studies have attested to the utility of this approach in various business domains further cementing its widespread adoption as a preferred method 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 generatedANN

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric Name | Related or Unrelated | Justification in relation to the success criteria | Model Name | Metric Score |
| Accuracy | related | The success criterion of the model is accurately predicting the risk groups, a pertinent and correlated indicator in assessing its performance is the percentage of correct predictions across the entire sample that allows for a determination of the degree to which records in each risk group are estimated correctly. Nonetheless, it is important to note that this score is incapable to account for the accuracy of the risk groups' individual estimations, necessitating the examination of other metrics. | NB | 0.58 |
| DT | 0.54 |
| KNN | 0.54 |
| ANN |  |
| Recall | related | The utilization of the recall metric, which depicts the proportion of correct predictions in all predictions made for each risk group separately, offers critical insights when analyzing risk groups independently. Through prioritizing and scrutinizing the High and Moderate risk groups, which constitute the primary focus of healthcare practitioners, it is feasible to attain the desired model in terms of its performance. By identifying the proportion of accurate predictions made for each risk group, it is possible to strive toward their augmentation. | NB | High Risk: 0.62 Moderate Risk: 0.55 Low Risk: 0.52 |
| DT | High Risk: 0.62 Moderate Risk: 0.55 Low Risk: 0.30 |
| KNN | High Risk: 0.66 Moderate Risk: 0.51 Low Risk: 0.34 |
| ANN |  |
| Precision | related | The precision metric, which indicates the proportion of true positive predictions in actual records for each risk group, represents a vital analytical tool for the evaluation of risk groups in isolation. By affording priority and attention to the High and Moderate risk groups, which form the focal point of healthcare professionals, it is possible to attain the desired level of model performance. In this way, it can be aimed to increase by knowing the percentage of correct risk predictions in all records in the particular risk group. | NB | High Risk: 0.75 Moderate Risk: 0.52 Low Risk: 0.41 |
| DT | High Risk: 0.65 Moderate Risk: 0.52 Low Risk: 0.31 |
| KNN | High Risk: 0.61 Moderate Risk: 0.51 Low Risk: 0.44 |
| ANN |  |
| F-Measure | related | F-Measure, which entails the evaluation of the harmonic mean of the predictions for each risk group in relation to both the actual records and predictions made for the respective group, represents a pivotal metric for the development of the targeted model. By emphasizing the scrutiny of priority risk groups, this analytical metric facilitates the acquisition of valuable insights regarding the accurate prediction of individual risk groups. | NB | High Risk: 0.68 Moderate Risk: 0.54 Low Risk: 0.46 |
| DT | High Risk: 0.63 Moderate Risk: 0.54 Low Risk: 0.31 |
| KNN | High Risk: 0.63 Moderate Risk: 0.51 Low Risk: 0.38 |
| ANN |  |
| AUC-ROC | related | The ROC-AUC score presents a valuable metric in the analysis of risk group differentiation, which forms the central aspect of this problem. Although it is inadequate as a standalone analytical metric, when used in conjunction with other evaluation metrics, it affords insight into the separation of individual risk group predictions from other groups. Thus, it represents a crucial metric to be considered for the enhancement of model success. | NB | High Risk: 0.83 Moderate Risk: 0.65 Low Risk: 0.80 |
| DT | High Risk: 0.69 Moderate Risk: 0.59 Low Risk: 0.59 |
| KNN | High Risk: 0.75 Moderate Risk: 0.60 Low Risk: 0.71 |
| ANN |  |

1. Waiting for the ANN model to select the best one
2. Waiting for the ANN model to select the best one
3. Waiting for the ANN model to select the best one

**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).