* 1. The target variable ‘Intervention Urgency’ has been classified into three classes ‘None’, ‘Medium’ and ‘High’, which makes classification predictive modelling to be used for this problem. However, regression is used for real values like integer or float etcetera.
  2. **Gender dataset:** Male

**Healthcare Question:** Question number 3)

**Type of classification problem:** multi-class classification problem.

* 1. **Possible input variables**

Note: I have splitted the screenshot for clear visibility. Kindly zoom in for more clarity.

Graphical user interface

Description automatically generated with medium confidence

Graphical user interface, application

Description automatically generated

**Class variable** **Distribution of class variable**

Graphical user interface, application, Teams

Description automatically generated

Chart, pie chart

Description automatically generated  
**Find the attached cleaned dataset here.**



**Statistically description**

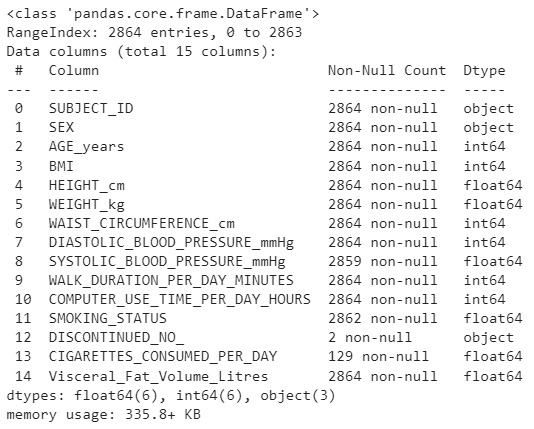
A picture containing table

Description automatically generated

Graphical user interface

Description automatically generated with low confidence

**Measurement Scale Type**



**a)**

|  |  |  |
| --- | --- | --- |
| **Dataset or Variable** | **Name of Variable** | **Issue Description** |
| Variable | SYSTOLIC\_BLOOD\_PRESSURE\_mmg | Found 5 null values. |
| Variable | SMOKING\_STATUS | Found 2 missing values. Some incorrect records Example- People who are smoking cigarettes have a smoking status of either 0 or undefined). |
| Variable | DISCONTINUED\_NO\_ | Almost, 2862 (99%) missing values are found. It is unknown missingness |
| Variable | CIGARETTES\_CONSUMED\_PER\_DAY | Almost, 95% of data is missing but it should be 0 because smoking status is 0. |
| Variable | AGE\_years | One record has age=190 which is not practically possible. |
| Variable | Height\_cm | Two records has height of 1.7cm and 1.8cm which is unreal/impossible. |
| Variable | COMPUTER\_USE\_TIME\_PER\_DAY\_HOURS | Found 12 records mentioning the use of computers more than 24 hours/day which is not possible at all in a single day. |
| Variable | Visceral\_Fat\_Vole\_Litres | Two records with visceral fat having values in minus like  -0.76liters, -0.59liters. |

**b)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset or Variable** | **Name of Variable** | **Issue Description** | **Solution** | **Justification** |
| Variable | SYSTOLIC\_BLOOD\_PRESSURE\_mmg | Found 5 null values. | NAN fillna(df[‘ ’]).mean() | The best practice to deal with NAN in a numerical variable is filling it with mean of the whole variable. |
| Variable | SMOKING\_STATUS | * Found 2 missing values. * Some incorrect records Example- People who are smoking cigarettes have a smoking status of 0. | NAN replaced with 1. Smoking status (0) is replaced with 1. | NAN 1  Because the count of cigarettes consumed per day is available.  (0 1)  Because person who is smoking cigarettes is a smoker (1) not a non-smoker (0). |
| Variable | DISCONTINUED\_NO\_ | Almost, 2862 (99%) missing values are found. It is unknown missingness. | Dropped the column | It is not a good practice to have 99% assumption in a column. So, I dropped it also it does not have much contribution for the class variable. |
| Variable | CIGARETTES\_CONSUMED\_PER\_DAY | Almost, 95% of data is missing. | Replaced  NAN 0 | It should be 0 because smoking status is 0. |
| Variable | AGE\_years | One record has age=190 | Dropped the record | It is not practically possible for humans to have age of 190 years. So, it is not useful for our predictions. |
| Variable | Height\_cm | Two records have height of 1.7cm and 1.8cm. | Dropped the records. | It is unreal data. |
| Variable | COMPUTER\_USE\_TIME\_PER\_DAY\_HOURS | Found 12 records mentioning the use of computers more than 24 hours/day. | Dropped the records. | It is not possible at all in a single day to spend more than 24 hours on computer. |
| Variable | Visceral\_Fat\_Vole\_Litres | Two records with visceral fat having values in minus like  -0.76liters, -0.59liters. | Replaced  Values 0 | Visceral fat cannot be in minus. However, the next possible value is 0 to be replaced with. Otherwise, if I would use mean () there is possibility that the person would be categorize in ‘Medium’ or ’High’ Intervention urgency and it is not appropriate. |

**c)**

Note: A sample of 5 records is displayed for showing problems and solutions instead of all the problematic records. You can refer to python notebook viewing complete changes.

**SYSTOLIC\_BLOOD\_PRESSURE\_mmg**

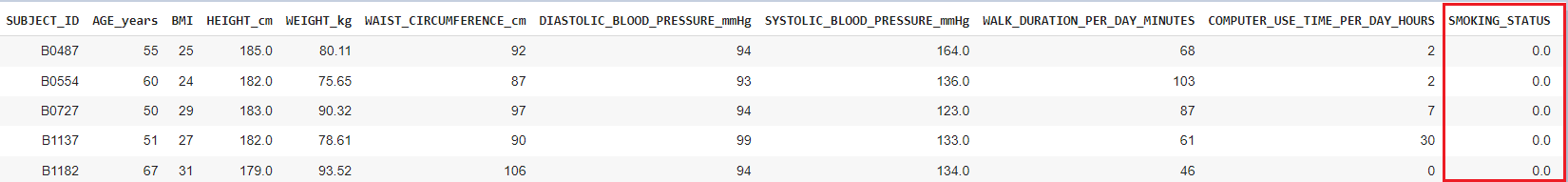
Graphical user interface, table

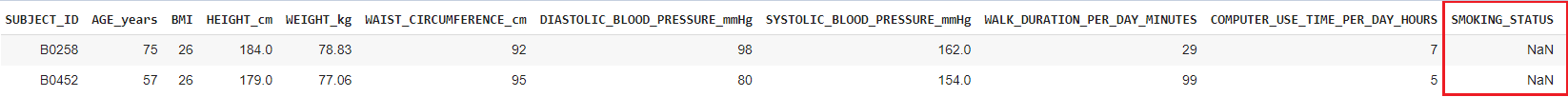
Description automatically generated

Graphical user interface, application

Description automatically generated

**SMOKING\_STATUS**







**DISCONTINUED\_NO\_**

Graphical user interface, application

Description automatically generated



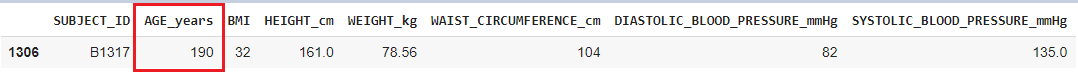
**CIGARETTES\_CONSUMED\_PER\_DAY**

A picture containing graphical user interface

Description automatically generated

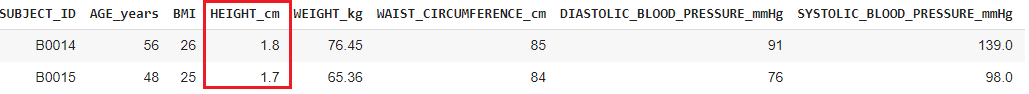


**AGE\_years**

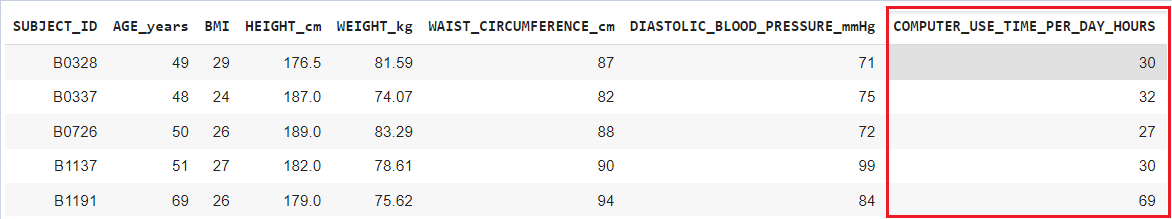




**Height\_cm**

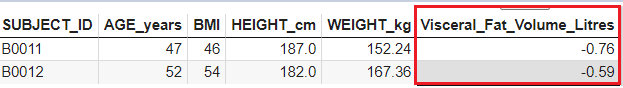


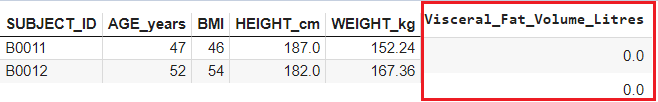
**COMPUTER\_USE\_TIME\_PER\_DAY\_HOURS**





**Visceral\_Fat\_Vole\_Litres**

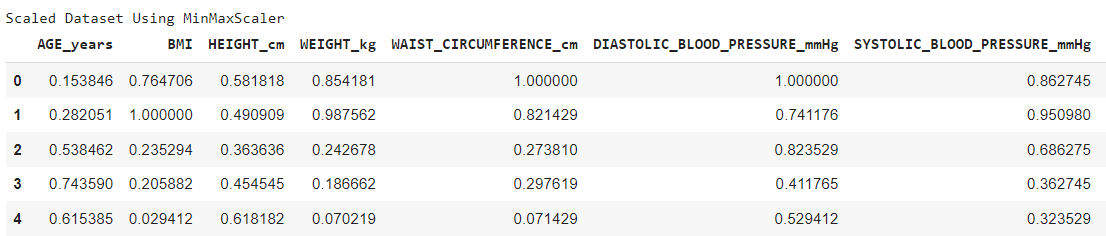


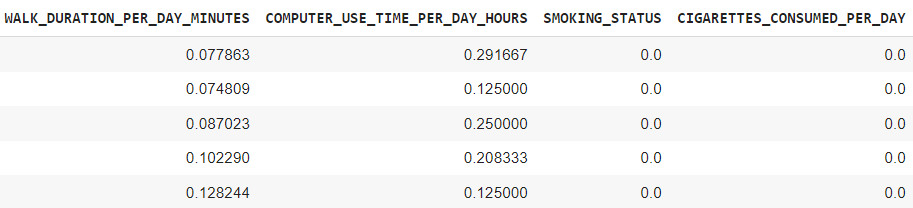


**a)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm Name** | **Type of Algorithm** | **Possible Hyper parameters** | **Python package source code to call the algorithm** |
| NB | Parametric | RepeatedStratifiedFold( n\_splits, n\_reeats, random\_state), GridSearchCV(estimator, param\_grid, verbose, cv, scoring) | from sklearn.naive\_bayes import GaussianNB |
| DT | Non-parametric | max\_depth, max\_leaf \_nodes, Class\_weight, criterion, max\_features, min\_impurity\_decrease, min\_impurity\_split, min\_samples\_leaf, min\_samples\_split, min\_weight\_fraction\_leaf, presort, random\_state, splitter | from sklearn.tree import DecisionTreeClassifier |
| KNN | Non-parametric | n\_neighbors, metric, weights, RepeatedStratifiedFold( n\_splits, n\_reeats, random\_state), GridSearchCV(estimator, param\_grid, n\_jobs, cv, scoring) | from sklearn.neighbors import KNeighborsClassifier |
| ANN | Parametric | GridSearchCV(MLPClassifier(random\_state), hidden\_layer\_size, param\_grid, n\_jobs, cv, verbose), activation, hidden\_layer\_sizes, learning\_rate, max\_iter, solver,neuron, optimizer, batch\_size, epochs | from sklearn.neural\_network import MLPClassifier |

**b)**





**Justification (Choice of train test split ratio):**

I divided the whole dataset into 80:20 so that I can evaluate the performance of an algorithm based on how it learned during the training process. There is high possibility that model would perform better if it is trained more. It is a good practice to test the model on same dataset to avoid overfitting rather than testing it on new data.

Graphical user interface, text, application

Description automatically generated

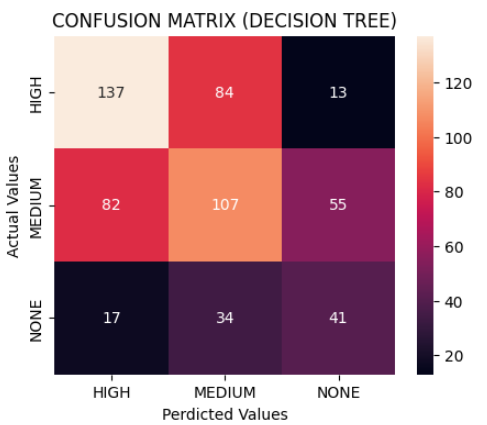
# **a)**

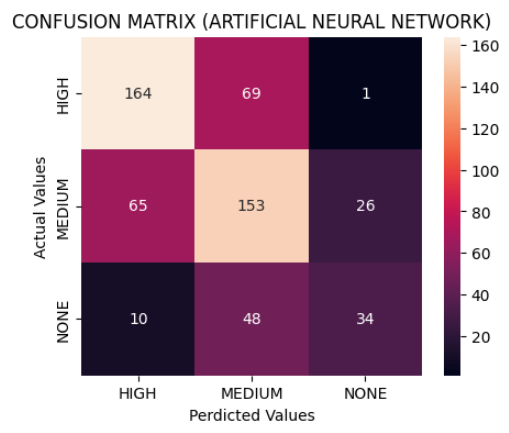
Chart, bar chart, treemap chart

Description automatically generated

Chart, treemap chart

Description automatically generated





**b)**

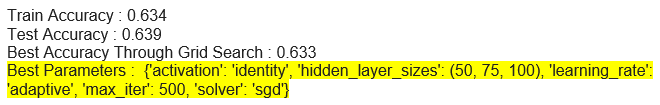
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metric Name** | **Related or Unrelated** | **Justification in relation to the success criteria** | **Model Name** | **Metric Score** |
| Accuracy | Not Related | Accuracy score matrix helps to evaluate the overall performance of a classifier. It reveals the fraction of correct predictions. Whereas only accuracy is not sufficient to evaluate the proportion of each class. | NB | **0.62** |
| DT | **0.50** |
| KNN | **0.60** |
| ANN | **0.62** |
| Recall | Strongly Related | Recall helps to identify true positive predictions.  It is good practice to prefer a higher recall rate rather than higher precision for meeting the success criteria in medical cases. Because it is better to make many people aware about the urgency for fast recovery rather than ignoring some positive instances. | NB | |  |  | | --- | --- | | **HIGH** | **0.62** | | **MEDIUM** | **0.69** | | **NONE** | **0.42** | |
| DT | |  |  | | --- | --- | | **HIGH** | **0.59** | | **MEDIUM** | **0.44** | | **NONE** | **0.45** | |
| KNN | |  |  | | --- | --- | | **HIGH** | **0.67** | | **MEDIUM** | **0.72** | | **NONE** | **0.13** | |
| ANN | |  |  | | --- | --- | | **HIGH** | **0.70** | | **MEDIUM** | **0.63** | | **NONE** | **0.37** | |
| Precision | Related | The matrix that tells us about the quality of positive prediction is precision. The highest percentage of positive rate would reduce the possibility of wrong predictions for each class. | NB | |  |  | | --- | --- | | **HIGH** | **0.75** | | **MEDIUM** | **0.55** | | **NONE** | **0.54** | |
| DT | |  |  | | --- | --- | | **HIGH** | **0.58** | | **MEDIUM** | **0.48** | | **NONE** | **0.38** | |
| KNN | |  |  | | --- | --- | | **HIGH** | **0.69** | | **MEDIUM** | **0.55** | | **NONE** | **0.55** | |
| ANN | |  |  | | --- | --- | | **HIGH** | **0.69** | | **MEDIUM** | **0.57** | | **NONE** | **0.56** | |
| F-Measure | Related | As compared to accuracy F-measure score is a better measure. As it depends on the harmonic mean of both precision and recall. | NB | |  |  | | --- | --- | | **HIGH** | **0.68** | | **MEDIUM** | **0.61** | | **NONE** | **0.48** | |
| DT | |  |  | | --- | --- | | **HIGH** | **0.58** | | **MEDIUM** | **0.46** | | **NONE** | **0.41** | |
| KNN | |  |  | | --- | --- | | **HIGH** | **0.68** | | **MEDIUM** | **0.62** | | **NONE** | **0.21** | |
| ANN | |  |  | | --- | --- | | **HIGH** | **0.69** | | **MEDIUM** | **0.60** | | **NONE** | **0.44** | |
| AUC-ROC | Strongly Related | As compared to other metrics the most important is AUC-ROC. It helps to measure and compare performance of different classifiers. The more AUC-ROC is close to 1.0, means a classifier is best. | NB | **0.77** |
| DT | **0.61** |
| KNN | **0.75** |
| ANN | **0.77** |

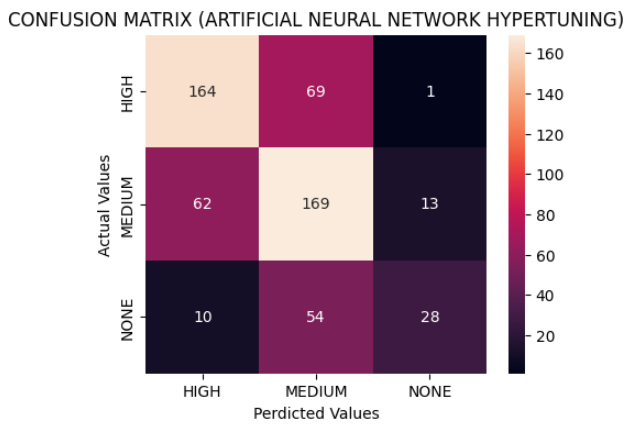
**c)**

Based on strongly related metrics the best classification model is Artificial Neural Network. This classifier has predicted the highest percentage of people for whom intervention urgency is HIGH and MEDIUM. Based on the results of ANN classifier 70% people will be informed to immediately make changes in their lifestyle and diet for leading healthy life. Moreover, 63% people will be made aware about the urgency so that they can reduce the detrimental health impacts (that visceral fat can cause) by taking appropriate resolutions into consideration. However, 37% masses will be informed to not initiate any lifestyle and diet changes. Overall, this model helps to cut down the medical cost (cost of MRI) for people who are healthy or they have MEDIUM intervention urgency. Only the subjects classified in HIGH urgency will be prescribed to have tests for beginning treatment. However, some classified in MEDIUM class can be tested as well based on their condition.

**d)**







|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Artificial Neural Network classification report after Hyper tuning** | | | | | |
|  | Accuracy | Recall | Precision | F-Measure | AUC-ROC |
| HIGH | 0.63 | 0.70 | 0.69 | 0.70 | 0.78 |
| MEDIUM | 0.69 | 0.58 | 0.63 |
| NONE | 0.30 | 0.67 | 0.42 |

A minor enhancement has been noticed in the generalization of base Artificial Neural Network. The proportion of people classified in HIGH intervention urgency are same as original ANN results. For MEDIUM class an incline of 6%, whereas for NONE class decline of 7% has been recorded. AUC-ROC score is inclined by 1% which still presents ANN as the best model among others. Conclusion, more expenses of MRI tests can be saved now as compared to the original ANN model. As, it has classified more people in MEDIUM and reasonably identified for NONE class for whom it is not that mandatory to spend on tests as they can overcome it by adapting healthy lifestyle and diet.

**e)**

It is possible that ANN will help healthcare professionals to find the level of intervention urgency for a subject without taking visceral fat into consideration. As, machine learning algorithm has learned from 80% of the historical dataset which includes visceral fat and other attributes. Now, it has learned the pattern of numbers and it will predict the ‘Intervention urgency’ by excluding visceral fat column. So, all the patients no more need to undergo for MRI scans and wear huge medical expenses.

ANN has not provided enough correct predictions as it should do to resolve the problem completely. It has mistakenly classified 70 HIGH urgency subjects into MEDIUM and NONE. It repeated the same mistakes for MEDIUM and NONE class. Such, wrong predictions can lead to serious effects like death or unnecessary burden for those who are actually health but asked to change lifestyle and diet.

ANN has a limitation that it does not significantly enhance its performance after hyper-tuning.

It is clear from the AUC-ROC, ANN is the best performer among other models. For healthcare/medical cases, the highest recall rate is considered to detect the decease/problem and I prioritised HIGH, MEDIUM and NONE in order for selecting the best model.

**References:**

[Visual representation of distribution of class variable](https://www.analyticsvidhya.com/blog/2020/07/univariate-analysis-visualization-with-illustrations-in-python/)

[Visual representation of Confusion Matrix](https://www.analyticsvidhya.com/blog/2021/06/confusion-matrix-for-multi-class-classification/)

[AUC ROC SCORE](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html)