# Task (1)

1. **Classification**: In this specific problem, we would be mapping the input data to different categories (Risk Level-Low, Moderate and High) based on Visceral fat levels. The target expected is purely classification as it has 3 labels within which data must be segregated.

# Task (2)

Classification Type: Multi-class

* 1. First 10 records of Data Frame: (Note: Please zoom in for clear view of the screenshot data.)

Table

Description automatically generated

Statistical Description:

Table

Description automatically generated with medium confidence

Measurement Scale Type:

Text

Description automatically generated with medium confidence

Distribution of Class Variable:

Chart, histogram

Description automatically generated

# Task (3)

|  |  |  |
| --- | --- | --- |
| Dataset or Variable | Name of variable | Issue description |
| Variable | AGE\_years | The Age is equal to 190 years |
| Variable | SYSTOLIC\_BLOOD\_PRESSURE\_mmHg | The column has 5 null Values |
| Variable | HEIGHT\_cm | Two records have Height as 1.7 and 1.8 cms. |
| Variable | COMPUTER\_USE\_TIME\_PER\_DAY\_HOURS | The computer usage is greater than 24 hours in a day. |
| Variable | CIGARETTES\_CONSUMED\_PER\_DAY | 95% missing values but dependent on Smoking Status – Missing values are for non-smokers which makes sense. |
| Variable | SMOKING\_STATUS | The column contains null values and few incorrect values.  Example: Smokers classified as non-smokers even though the no of cigarettes column has a value. |
| Variable | DISCONTINUED\_NO\_ | It’s an unknown column and has more than 99 percent null values |
| Variable | Visceral\_Fat\_Volume\_Litres | Variable contains negative values of Visceral fat (ex: -0.76, -0.59) |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset or Variable | Name of variable | The Issue | Solution | Justification |
| Variable | AGE\_years | Age=190 | Drop records where age>120 | It is not practically possible to have age as 190 in current situation therefore dropping the record. |
| Variable | SYSTOLIC\_BLOOD\_PRESSURE\_mmHg | 5 Null Values | Replace Null Values with Mean | Since it’s a numeric column it’s a good strategy to replace null values with mean. |
| Variable | HEIGHT\_cm | Height is equal to 1.7 and 1.8 cms | Drop records where height<50cms (1.5 feet) | Height cannot be less than 1.5 feet (Average born baby height) hence dropping the records. |
| Variable | COMPUTER\_USE\_TIME\_PER\_DAY\_HOURS | Computer Usage>24 Hours | Drop records where computer usage>24 hours in a day | We cannot feed incorrect input  data to train algorithms since dropping the records with computer usage>24 hours |
| Variable | CIGARETTES\_CONSUMED\_PER\_DAY | Missing Data | Replace null values with 0 for non-smokers | Non-smokers can have the number of cigarettes data as 0. |
| Variable | SMOKING\_STATUS | Null and Incorrect Values (Smokers are given status as Non-Smokers) | Setting smoking status=1 for smokers and 0 for non-smokers | Few columns have Smoking status as 0 and null even when the cigarettes consumed per day has a value. So, replacing those with 1. |
| Variable | DISCONTINUED\_NO\_ | 99% Missing data and Unknown Column | Dropping the Column | The column has no meaningful data and is not useful for target prediction. Therefore, we can eliminate it from our input variables. |
| Variable | Visceral\_Fat\_Volume\_Litres | Negative Values (-0.76, -0.59) | Replacing them with 0 as it is the nearest possible Value | Since we cannot have visceral fat volume as negative, it is considered as 0 which is nearest to the shown values. We can’t replace it with mean as there is a risk associated with higher visceral fat value. |

1. Finding and dropping the records with AGE\_years >120

Graphical user interface, application, Teams

Description automatically generated

Filling null Values with Mean for SYSTOLIC\_BLOOD\_PRESSURE\_mmHg

Table

Description automatically generated

Graphical user interface, application

Description automatically generated

Drop records where HEIGHT\_cm < 50 (1.5 feet)

Graphical user interface, application

Description automatically generated

Drop records where COMPUTER\_USE\_TIME\_PER\_DAY\_HOURS >24

Table

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Checking CIGARETTES\_CONSUMED\_PER\_DAY for Null Values and Replace with 0

Graphical user interface, application

Description automatically generated

Setting SMOKING\_STATUS to 1 for smokers

Table

Description automatically generated

Replacing negative Visceral\_Fat\_Volume\_Litres values to 0

Graphical user interface, table

Description automatically generated

Drop Columns (DISCONTINUED\_NO\_, SEX, SUBJECT\_ID and Visceral\_Fat\_Volume\_Litres) which are not used for Target Prediction and remove the Risk Level column from Input Data frame and consider it as a Target column to test train split the data.

Input Variables: 11

Target Variable: 1 (Risk Level)

# Task 4

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm Name | Type of Algorithm | Possible Hyper -parameters | Python package source code to call the algorithm |
| NB | parametric | priors,var\_smoothing | import sklearn  from sklearn.naive\_bayes import GaussianNB |
| DT | non-parametric | criterion, splitter, max\_depth, max\_leaf\_nodes, min\_impurity\_decrease, min\_samples\_split, min\_samples\_leaf,  max\_features | import sklearn  from sklearn.tree import DecisionTreeClassifier |
| KNN | non-parametric | n-neighbors, weights, metric, p, algorithm, leaf\_size | import sklearn  from sklearn.neighbors import KNeighborsClassifier |
| ANN(MLP) | parametric | hidden\_layer\_sizes, activation, solver, max\_iter, learning\_rate, random\_state | import sklearn  from sklearn.neural\_network import MLPClassifier |

1. Head of Data frame (Scaled using min max scaler) used to train the Algorithms.

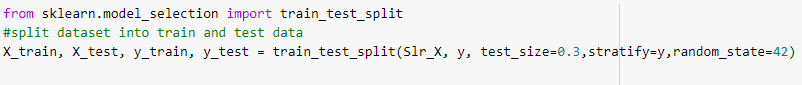
Graphical user interface

Description automatically generated

Graphical user interface, application

Description automatically generated

Since we have smaller dataset (<10000 records) I have chosen train test split ratio as 70:30 so that the algorithm is provided with the sufficient data to predict target. As we have 3 target classes I have set stratify=y (Risk\_Level) so that we can have the same proportion of records for each class in both test train data as compared to the whole dataset.



All algorithms have been trained and tested on the **same dataset which is attached below**.

Text, letter

Description automatically generated 

# Task 5



Chart, treemap chart

Description automatically generated

Chart, treemap chart

Description automatically generated

Chart, treemap chart

Description automatically generated

Chart, treemap chart

Description automatically generated



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric Name | Related or Unrelated | Justification in relation to the success criteria | Model Name | Metric Score |
| Accuracy | Unrelated | The Best accuracy gives us the greater number of correct predictions and helps to evaluate the overall performance of the classifier. However, Accuracy alone would not be a best choice in considering the best model to meet given success criteria. | NB | 0.63 |
| DT | 0.63 |
| KNN | 0.60 |
| ANN | 0.64 |
| Recall | Related | Recall calculates the positive predictions rate for each class.  It’s always preferred to inform as many subjects as possible about their urgency rather than ignoring some of them as it would cause serious impact on their health.  So, a higher recall for high and moderate risk and considerably good recall for low risk would give us a better model to meet the success criteria. | NB | |  |  | | --- | --- | | High Risk | 0.65 | | Moderate Risk | 0.67 | | Low Risk | 0.48 | |
| DT | |  |  | | --- | --- | | High Risk | 0.71 | | Moderate Risk | 0.64 | | Low Risk | 0.38 | |
| KNN | |  |  | | --- | --- | | High Risk | 0.71 | | Moderate Risk | 0.68 | | Low Risk | 0.14 | |
| ANN | |  |  | | --- | --- | | High Risk | 0.73 | | Moderate Risk | 0.69 | | Low Risk | 0.25 | |
| Precision | Related | Precision is the fraction of true positive predictions to the total number of positive predictions.  It’s important to have a good precision as we don’t want to falsely predict high and moderate risk as Low as it would cause someone to ignore further testing and medication that causes adverse effects and its vice versa to not predict low risk as high/moderate as it would cause unnecessary initiation of lifestyle. | NB | |  |  | | --- | --- | | High Risk | 0.77 | | Moderate Risk | 0.57 | | Low Risk | 0.53 | |
| DT | |  |  | | --- | --- | | High Risk | 0.72 | | Moderate Risk | 0.57 | | Low Risk | 0.54 | |
| KNN | |  |  | | --- | --- | | High Risk | 0.68 | | Moderate Risk | 0.55 | | Low Risk | 0.54 | |
| ANN | |  |  | | --- | --- | | High Risk | 0.73 | | Moderate Risk | 0.58 | | Low Risk | 0.55 | |
| F-Measure | Related | F1 score is harmonic mean of recall and precision. F1 measure is better metric in considering the best model as compared to accuracy as it gives us the performance score for each class. | NB | |  |  | | --- | --- | | High Risk | 0.71 | | Moderate Risk | 0.62 | | Low Risk | 0.50 | |
| DT | |  |  | | --- | --- | | High Risk | 0.72 | | Moderate Risk | 0.60 | | Low Risk | 0.44 | |
| KNN | |  |  | | --- | --- | | High Risk | 0.69 | | Moderate Risk | 0.61 | | Low Risk | 0.23 | |
| ANN | |  |  | | --- | --- | | High Risk | 0.73 | | Moderate Risk | 0.63 | | Low Risk | 0.35 | |
| AUC-ROC | Related | AUROC score is the area under ROC curve plotted against the true positive rate and false positive rate. The higher the AUROC the better is the classifier. | NB | **0.7903** |
| DT | **0.7671** |
| KNN | **0.7615** |
| ANN | **0.7979** |

Based on related performance metrics, I have chosen **MLP** as the Best Classification Model.

Considering the **recall and F1 score which are strongly related as compared to precision**,

The model has predicted higher percentage of high and medium risk cases and considerably good percentage of Low risk subjects.

It’s clear that 73 percent people can be intimated about the urgency to make necessary changes in their diet and lifestyle to lead a better healthy life and 69 percent subjects would be made aware about the urgency so that they can take necessary actions to reduce the adverse effects on their health that visceral fat may cause. However,25 percent of the people would not be impacted and required no initiation of lifestyle changes.

**AUROC** value is also higher for ANN(MLP) making it a more reliable classifier among the four.



TUNING USING GRIDSEARCHCV

Number of **Cross-validation** K folds Used: 5

**Best Parameters** : {'activation': 'relu', 'hidden\_layer\_sizes': (10, 20), 'learning\_rate': 'constant', 'max\_iter': 500, 'solver': 'adam'}

Chart, treemap chart

Description automatically generated

**A picture containing text, receipt

Description automatically generated**

After Hyperparameter tuning, the original best model has been slightly enhanced as there is an increase in related metrics (precision, recall, F1 scores and AUROC) for High, Moderate and Low risk Classes. So now there is an increase in recall and F1 score for moderate and low risk prediction by 4-5 percent on an average.

Our Best Model after tuning focused on generating larger true predictions for both High and Moderate risk and considerably good for Low Risk based on basis of which we can conclude that ANN(MLP) makes a reliable model to meet the success criteria as mentioned by the Healthcare professionals.



Yes, MLP Model has the potential to significantly classify high, moderate, and low risk classes without taking the visceral fat levels into consideration as the machine learning algorithm has been trained on the historical dataset and learnt the patterns based on Input features.

The algorithm can now predict the risk level for the new samples based on earlier learnt pattern which in turn reduces the high costs involved in MRI scans.

Although, the model has the greater prediction to classify High and moderate risk I see there is a limitation for MLP in predicting a higher number of Low-risk cases which means that it has a possibility of misclassifying them as either high or moderate risk. This would cause havoc for people who has no risk but classified as one.

Overall, MLP is very flexible and based on the metrics it’s proved to be learning more accurately as compared to the other 3 models from the input and predict higher number of HIGH and MODERATE risk cases which is the priority for healthcare professionals to help and save their life. People classified as High Risk and moderate risk can be diagnosed further to take proper medication and change their lifestyle.

References:

[KNN(Hyper Parameters)](https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html)

[KNN distance metrics](https://www.ibm.com/uk-en/topics/knn#:~:text=The%20k%2Dnearest%20neighbors%20algorithm%2C%20also%20known%20as%20KNN%20or,of%20an%20individual%20data%20point.)

[GaussianNB(Hyper Parameters)](https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html)

[Decision Tree(Hyper Parameters)](https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html)

[MLP Hyper Parameter Tuning](https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html)

[AUROC Score Calculation for Multiclass](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html) using Probabilities