PRT 565 – MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

ASSIGNMENT 1

Lecturer: Asif Karim

Student Name: Tai Phu Phan (s342489)

Table of Contents

List of Figures

List of Tables

1. Problem Description

Osteoporosis is the most prevalent chronic bone disease that is characterized by the loss of bone density, deterioration of bone tissue, and it can lead to an increase in the risk of bone fragility (Klibanski et al. 201). From recent studies, although osteoporosis has been seen in all age groups, gender, and ethnicities, it is more in Caucasians, older people, and women. The International Osteoporosis Foundation has recently released statistics showing that one in three women over the age of 50 and one in five men may have an osteoporosis fracture during their lifetime. Additionally, with the estimation that more than 200 million people are suffering from osteoporosis, it is believed that osteoporosis has increasingly become a global pandemic (Sözen et al. 2017). Therefore, increasing awareness about osteoporosis in the community and scientific research from health organizations worldwide is essential in preventing this pandemic. This also leads to the motivation of choosing this topic for the purpose of this assessment.

2. Dataset Introduction

For this assessment work, the dataset (osteoporosis disease) is extracted from a cross-sectional study involving 300 postmenopausal Vietnamese women aged more than 50 who were randomly sampled from different districts in Ho Chi Minh City, Vietnam. This research is led by Dr. Tuan V Nguyen and other researchers to determine the contributing factors to osteoporosis.

This dataset consists of 300 records representing all women who are chosen for the study. There are ten features, and only one feature is categorical; the rest are numeric. The definition of this dataset is as follows. The diagnosis of osteoporosis is based on bone density measured by a DXA scan. The dataset can be accessed via this link https://github.com/tuanvnguyen/Regression-Book

Feature Name	Data Type	Description	Unit	Example
Id	Numeric (Integer)	Unique number for all participants	N/A	10, 20
lean.mass	Numeric (Float)	Total body weight minus all the weight due to fat mass	kg	27.98, 32.58
fat.mass	Numeric (Float)	Total weight of body fat	kg	16.49, 20.65
Pcfat	Numeric (Float)	Percent body fat	N/A	37.09, 40.37

Age	Numeric (Integer)	Age	years	50, 70
Height	Numeric (Float)	Measure of height	cm	142.8, 156
Weight	Numeric (Float)	Measure of weight	kg	51.5, 47.5
Bmi	Numeric (Float)	Body Mass Index	kg/m2	18.5, 25.5
Osta	Numeric (Float)	Osteoporosis Self-Assessment Tool for Asians (OSTA) score - 0.2*(weight (kg) – age (year))	N/A	0.9, 6.2
osteo.group	Text	Categorized groups of participants	N/A	Normal, Osteopenia, Osteoporosis

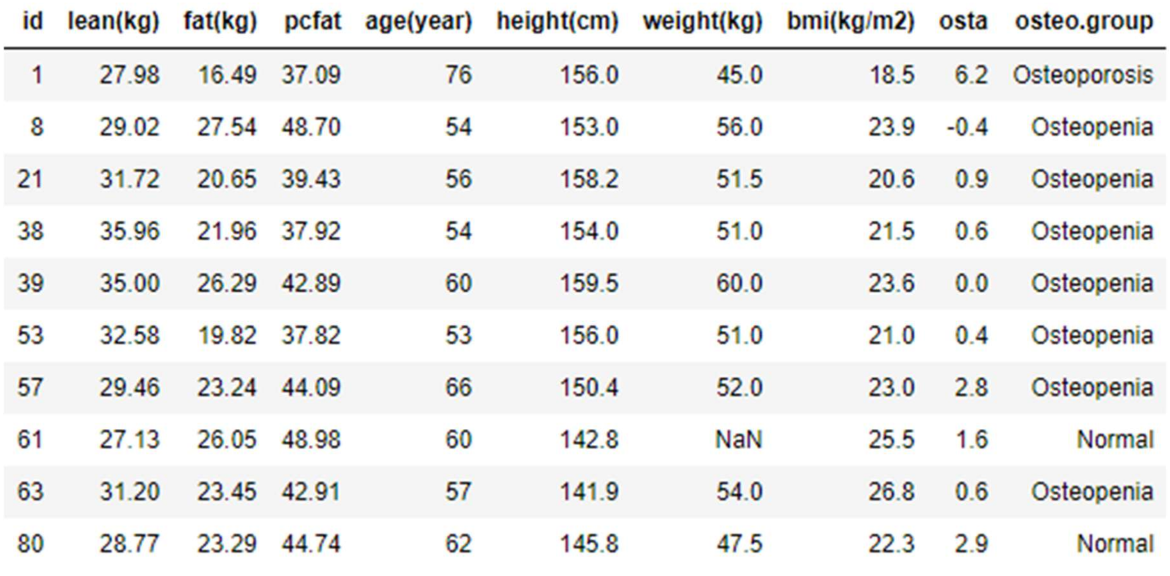
Table 1. Dataset definition

This assessment aims to build a machine learning model to predict the probability of being suffered from osteoporosis disease, either osteopenia (lower bone density) or osteoporosis (severe case of bone loss that can lead to fracture).

3. Methodology

3.1 Importing data

The dataset in csv format is imported using the Python Pandas package. Several columns are renamed to include the unit of measurement for easy reference.

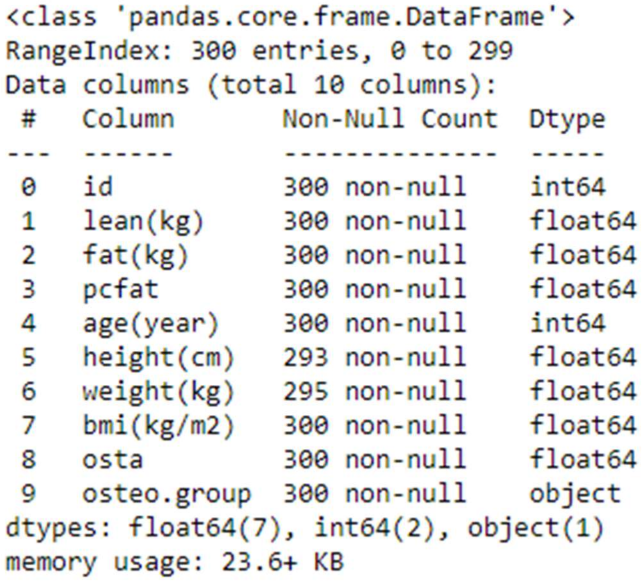


id	lean(kg)	fat(kg)	pcfat	age(year)	height(cm)	weight(kg)	bmi(kg/m2)	osta	osteo.group
1	27.98	16.49	37.09	76	156.0	45.0	18.5	6.2	Osteoporosis
8	29.02	27.54	48.70	54	153.0	56.0	23.9	-0.4	Osteopenia
21	31.72	20.65	39.43	56	158.2	51.5	20.6	0.9	Osteopenia
38	35.96	21.96	37.92	54	154.0	51.0	21.5	0.6	Osteopenia
39	35.00	26.29	42.89	60	159.5	60.0	23.6	0.0	Osteopenia
53	32.58	19.82	37.82	53	156.0	51.0	21.0	0.4	Osteopenia
57	29.46	23.24	44.09	66	150.4	52.0	23.0	2.8	Osteopenia
61	27.13	26.05	48.98	60	142.8	NaN	25.5	1.6	Normal
63	31.20	23.45	42.91	57	141.9	54.0	26.8	0.6	Osteopenia
80	28.77	23.29	44.74	62	145.8	47.5	22.3	2.9	Normal

Figure 1. Importing dataset

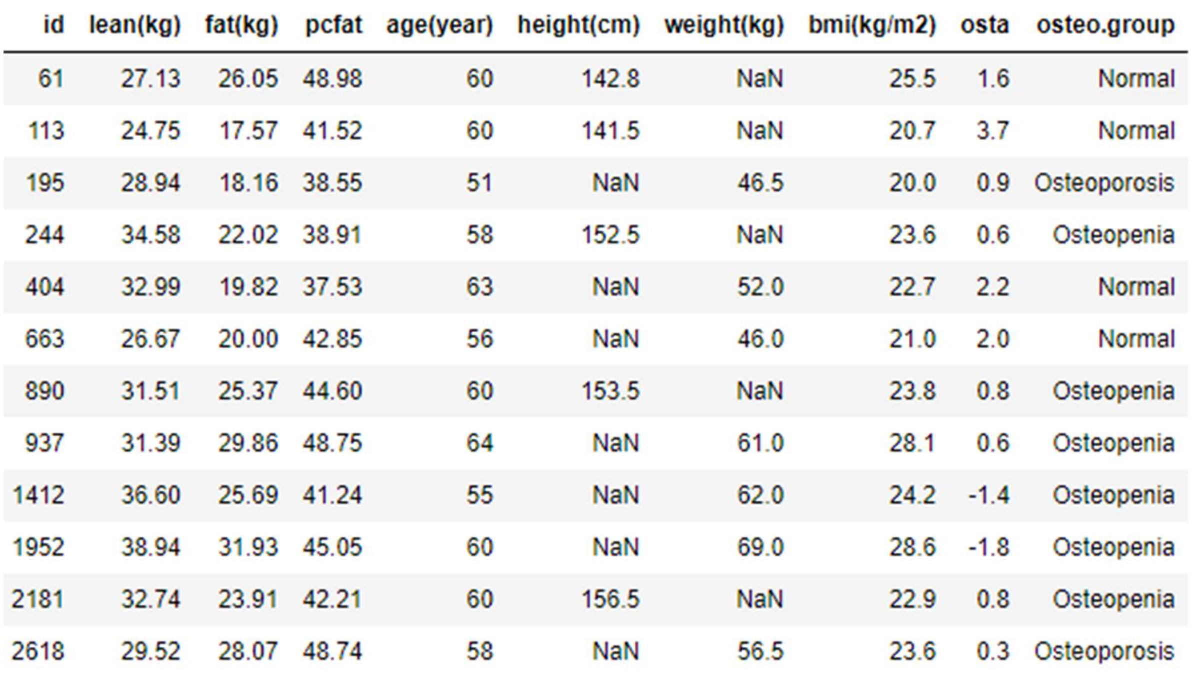
3.2 Data pre-processing

The dataset is first checked for data types and missing values.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 300 entries, 0 to 299
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   id            300 non-null    int64  
 1   lean(kg)      300 non-null    float64
 2   fat(kg)       300 non-null    float64
 3   pcfat         300 non-null    float64
 4   age(year)     300 non-null    int64  
 5   height(cm)    293 non-null    float64
 6   weight(kg)    295 non-null    float64
 7   bmi(kg/m2)    300 non-null    float64
 8   osta          300 non-null    float64
 9   osteo.group   300 non-null    object 
dtypes: float64(7), int64(2), object(1)
memory usage: 23.6+ KB
```

Figure 2. Data types

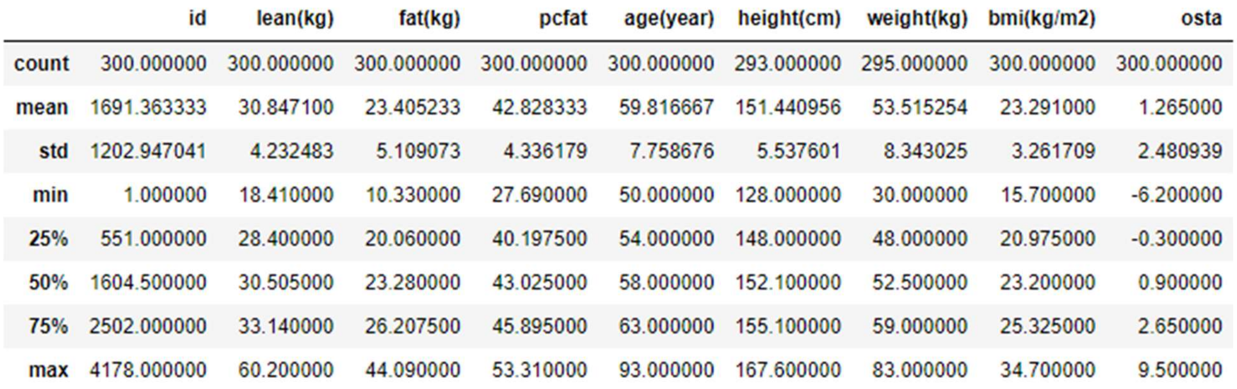


id	lean(kg)	fat(kg)	pcfat	age(year)	height(cm)	weight(kg)	bmi(kg/m2)	osta	osteo.group
61	27.13	26.05	48.98	60	142.8	NaN	25.5	1.6	Normal
113	24.75	17.57	41.52	60	141.5	NaN	20.7	3.7	Normal
195	28.94	18.16	38.55	51	NaN	46.5	20.0	0.9	Osteoporosis
244	34.58	22.02	38.91	58	152.5	NaN	23.6	0.6	Osteopenia
404	32.99	19.82	37.53	63	NaN	52.0	22.7	2.2	Normal
663	26.67	20.00	42.85	56	NaN	46.0	21.0	2.0	Normal
890	31.51	25.37	44.60	60	153.5	NaN	23.8	0.8	Osteopenia
937	31.39	29.86	48.75	64	NaN	61.0	28.1	0.6	Osteopenia
1412	36.60	25.69	41.24	55	NaN	62.0	24.2	-1.4	Osteopenia
1952	38.94	31.93	45.05	60	NaN	69.0	28.6	-1.8	Osteopenia
2181	32.74	23.91	42.21	60	156.5	NaN	22.9	0.8	Osteopenia
2618	29.52	28.07	48.74	58	NaN	56.5	23.6	0.3	Osteoporosis

Figure 3. Missing data

We can observe the data types of all features, and missing values appearing for two features, height and weight, in the figure 2 and figure 3. The proportions of missing data for these features are 2.3% and 1.7%, respectively, which are acceptable, but this assessment work will utilize the data imputation technique to make sure all the missing values are effectively used and will produce optimal results.

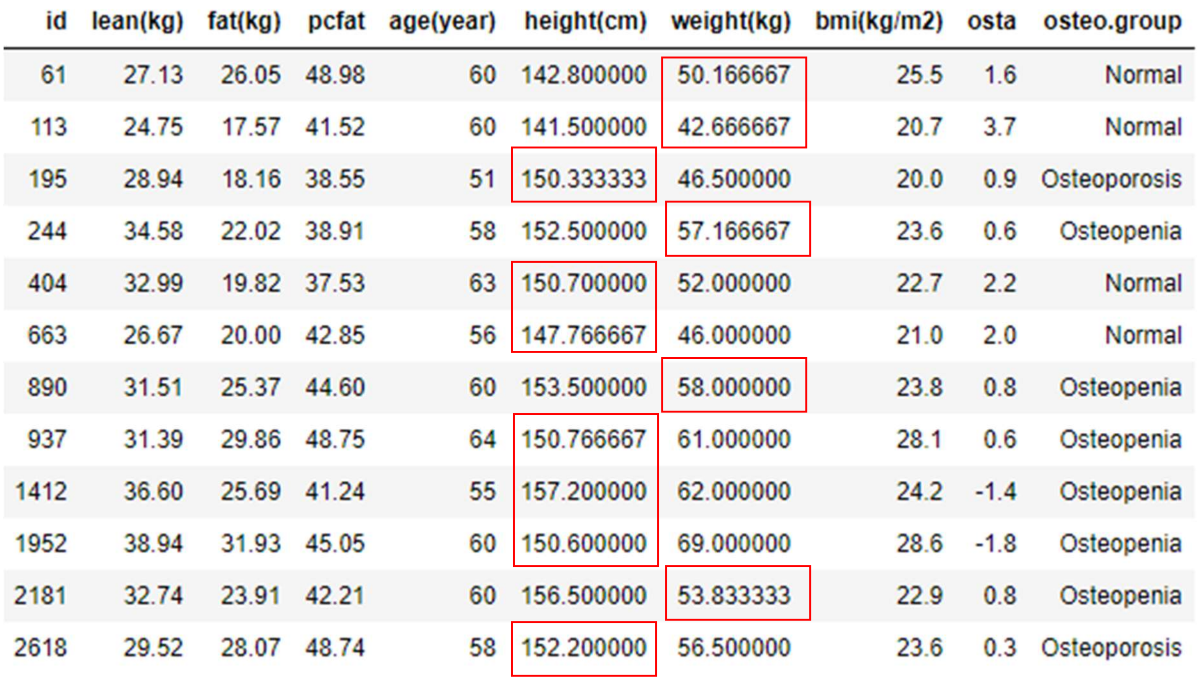
Figure 4 below shows the descriptive characteristics of the dataset. At first glance, except for missing values, all the data values seem to appear without errors or very rare observations; however, outliers may come up across the features. The following part will show the technique for detecting outliers.



	id	lean(kg)	fat(kg)	pcfat	age(year)	height(cm)	weight(kg)	bmi(kg/m2)	osta
count	300.000000	300.000000	300.000000	300.000000	300.000000	293.000000	295.000000	300.000000	300.000000
mean	1691.363333	30.847100	23.405233	42.828333	59.816667	151.440956	53.515254	23.291000	1.265000
std	1202.947041	4.232483	5.109073	4.336179	7.758676	5.537601	8.343025	3.261709	2.480939
min	1.000000	18.410000	10.330000	27.690000	50.000000	128.000000	30.000000	15.700000	-6.200000
25%	551.000000	28.400000	20.060000	40.197500	54.000000	148.000000	48.000000	20.975000	-0.300000
50%	1604.500000	30.505000	23.280000	43.025000	58.000000	152.100000	52.500000	23.200000	0.900000
75%	2502.000000	33.140000	26.207500	45.895000	63.000000	155.100000	59.000000	25.325000	2.650000
max	4178.000000	60.200000	44.090000	53.310000	93.000000	167.600000	83.000000	34.700000	9.500000

Figure 4. Descriptive characteristics of dataset

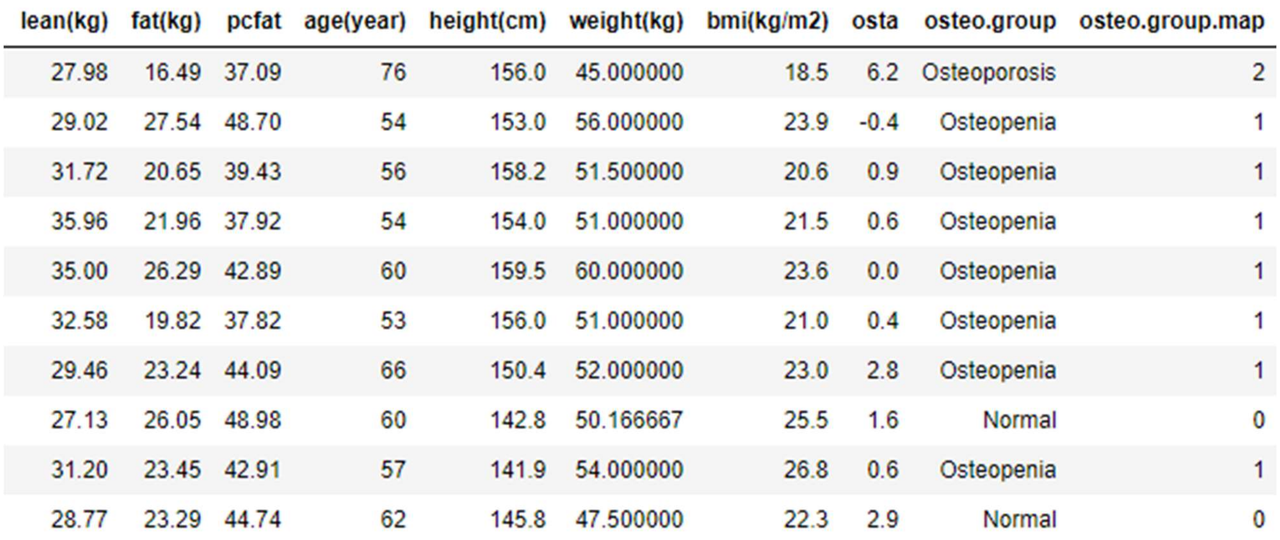
For handling missing values, the KNN imputation method is employed to predict those missing data. This method replaces missing values with the mean value of its neighbors with selected features such as lean, fat, pcfat, age. Table below represents the imputed value for all the cell highlighted in red; for example, weight of participant id equal to 61 is now 50.166 kg.



id	lean(kg)	fat(kg)	pcfat	age(year)	height(cm)	weight(kg)	bmi(kg/m2)	osta	osteo.group
61	27.13	26.05	48.98	60	142.800000	50.166667	25.5	1.6	Normal
113	24.75	17.57	41.52	60	141.500000	42.666667	20.7	3.7	Normal
195	28.94	18.16	38.55	51	150.333333	46.500000	20.0	0.9	Osteoporosis
244	34.58	22.02	38.91	58	152.500000	57.166667	23.6	0.6	Osteopenia
404	32.99	19.82	37.53	63	150.700000	52.000000	22.7	2.2	Normal
663	26.67	20.00	42.85	56	147.766667	46.000000	21.0	2.0	Normal
890	31.51	25.37	44.60	60	153.500000	58.000000	23.8	0.8	Osteopenia
937	31.39	29.86	48.75	64	150.766667	61.000000	28.1	0.6	Osteopenia
1412	36.60	25.69	41.24	55	157.200000	62.000000	24.2	-1.4	Osteopenia
1952	38.94	31.93	45.05	60	150.600000	69.000000	28.6	-1.8	Osteopenia
2181	32.74	23.91	42.21	60	156.500000	53.833333	22.9	0.8	Osteopenia
2618	29.52	28.07	48.74	58	152.200000	56.500000	23.6	0.3	Osteoporosis

Figure 5. Data after imputation

The next step as shown in figure 6 is to transform the response variable (osteo.group) into numeric values for feeding into the machine learning model.



lean(kg)	fat(kg)	pcfat	age(year)	height(cm)	weight(kg)	bmi(kg/m2)	osta	osteo.group	osteo.group.map
27.98	16.49	37.09	76	156.0	45.000000	18.5	6.2	Osteoporosis	2
29.02	27.54	48.70	54	153.0	56.000000	23.9	-0.4	Osteopenia	1
31.72	20.65	39.43	56	158.2	51.500000	20.6	0.9	Osteopenia	1
35.96	21.96	37.92	54	154.0	51.000000	21.5	0.6	Osteopenia	1
35.00	26.29	42.89	60	159.5	60.000000	23.6	0.0	Osteopenia	1
32.58	19.82	37.82	53	156.0	51.000000	21.0	0.4	Osteopenia	1
29.46	23.24	44.09	66	150.4	52.000000	23.0	2.8	Osteopenia	1
27.13	26.05	48.98	60	142.8	50.166667	25.5	1.6	Normal	0
31.20	23.45	42.91	57	141.9	54.000000	26.8	0.6	Osteopenia	1
28.77	23.29	44.74	62	145.8	47.500000	22.3	2.9	Normal	0

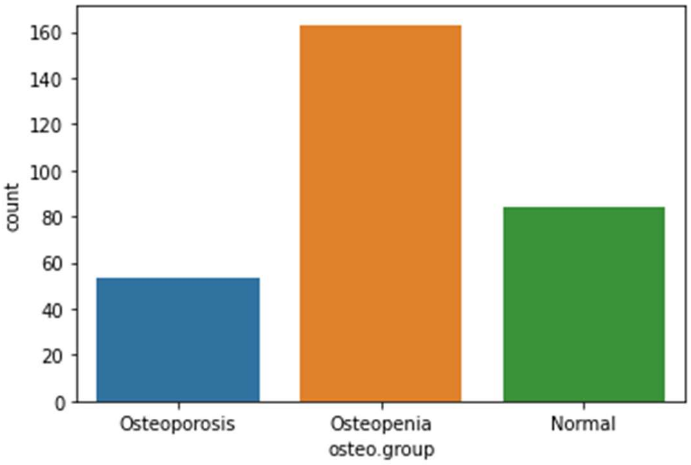
Figure 6. Mapping response variable

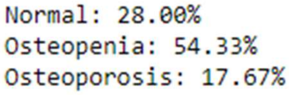
Now the new reponse variable (osteo.group.map) can be 0, 1, or 2 representing normal, osteopenia, or osteoporosis.

3.3 Exploratory Data Analysis (EDA)

The EDA process is conducted to analyze the dataset to summarize its characteristics with the visual elements before the modeling task.

The bar chart in figure 7 shows that only 28% participants out of 300 postmenopausal women are free from this bone disease, and over half of the sample (54.33%) are suffering from osteopenia, and the number of osteoporosis patients is still high, 17.67%.





```
Normal: 28.00%
Osteopenia: 54.33%
Osteoporosis: 17.67%
```

Figure 7. Barplot of osteo groups

Next, the histogram charts in figure 8 represent the visual description of the distribution shape for each contributing feature. All the features appear to be approximately normal distribution except for the age variable is right skewed. Outliers seem to appear across features.

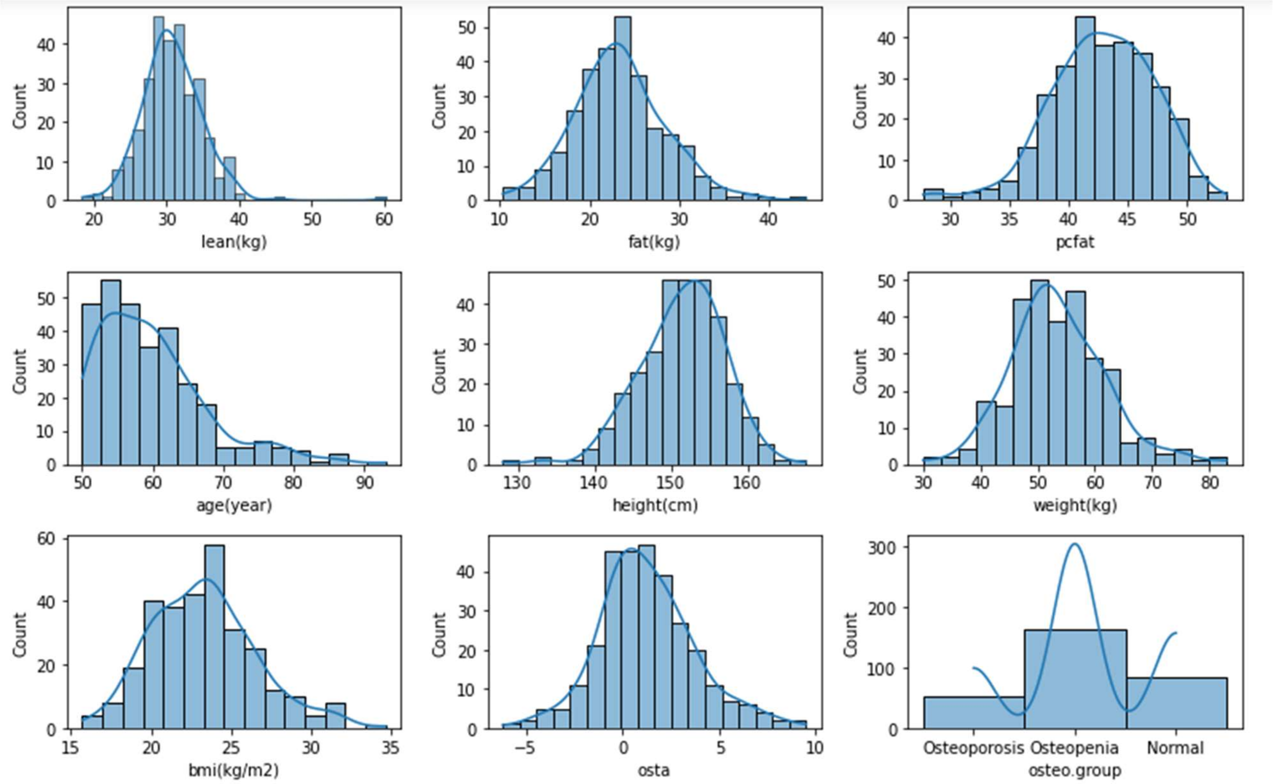


Figure 8. Distribution shape of features

Figure 9 consists of boxplot charts representing the outliers from all major contributing features. These outliers appear mostly outside of the 1.5 of the third quartile. Removing these outliers from the dataset is not an appropriate approach unless there is a strong/significant justification to remove these data points. Therefore, outliers are remained as a part of the selected sample.

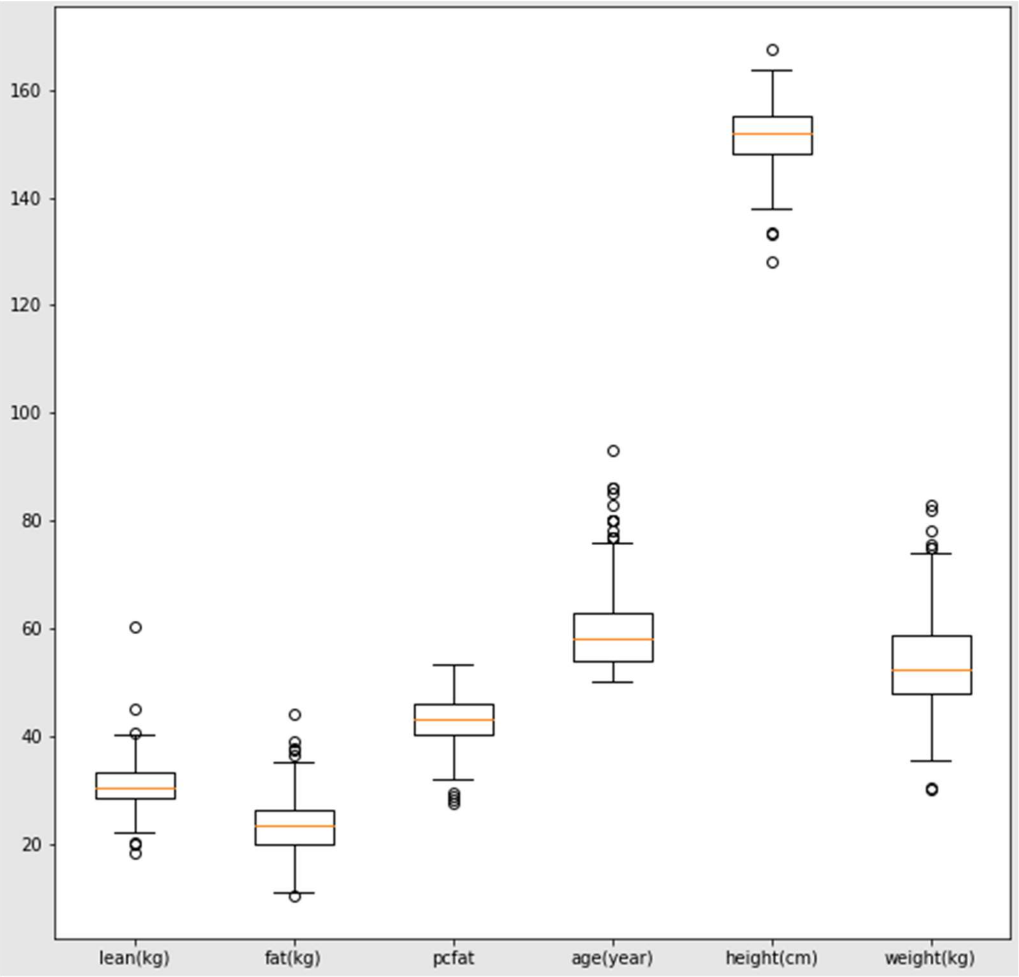


Figure 9. Outliers detection

The last step in this EDA process in this assessment is constructing a correlation matrix (figure 10) to evaluate the direction and strength of the relationship between selected features.

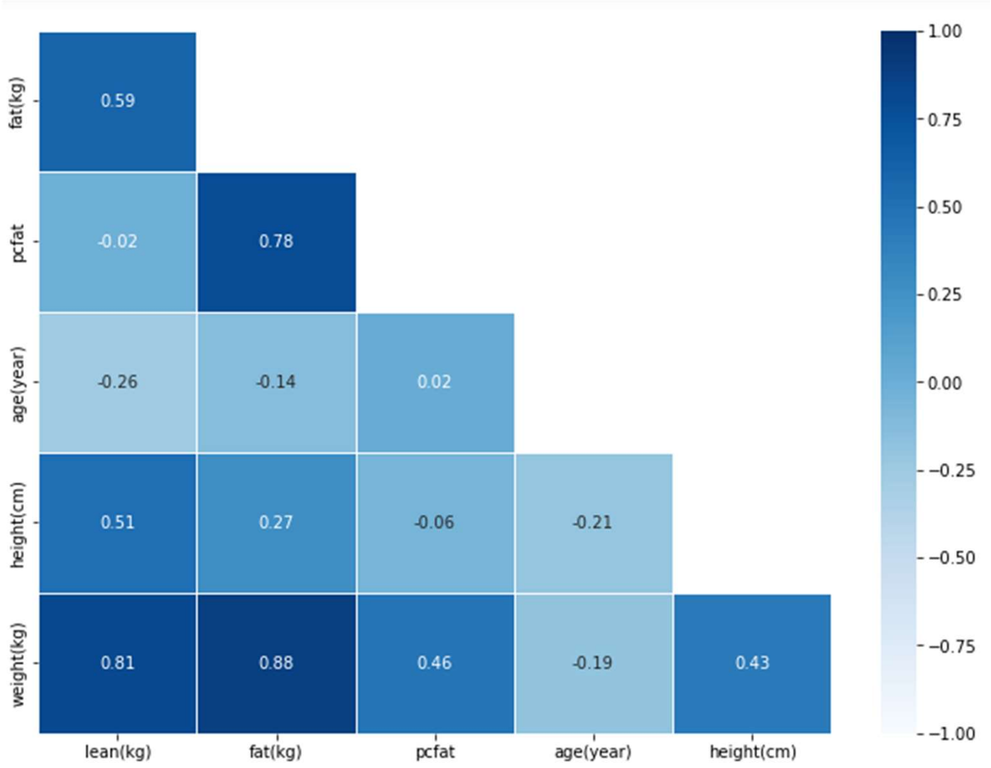


Figure 10. Correlation matrix of features

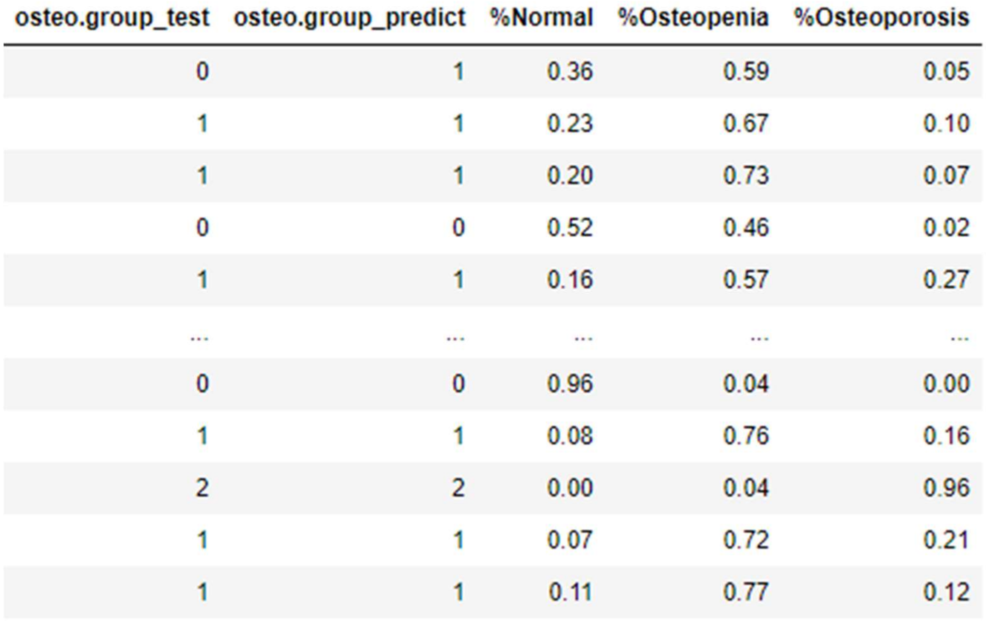
We can observe from the matrix that some variables are highly correlated with the others.

3.4 Machine Learning model

To achieve the aim of this assessment to predict the possibility of suffering from osteoporosis disease, which are multiple outcomes rather than binary, multinomial logistic regression and random forest are employed.

3.4.1 Multinomial Logistic Regression

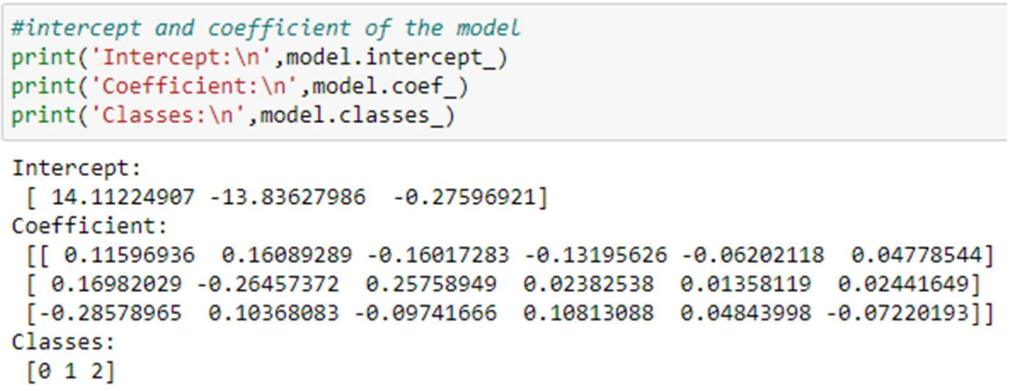
In this model, explanatory variables include “lean”, “fat”, “pcfat”, “age”, “height”, and “weight”. The response variable is “osteo.group.map”, which is encoded to numeric values representing normal or one of the diseases. The original dataset is split into train and test with a ratio of 25% for testing. Scikit-learn is used for splitting data and modeling. The probability of each predicted outcome will also be estimated to compare against the test dataset as in figure 11. For example, the first row shows that the model predicts the possibility of 59% that this participant is suffering osteopenia, 36% is normal and very low possibility of suffering osteoporosis. Therefore, the model classifies this participant in the group of osteopenia (column osteo.group_predit). However, this participant is normal (column osteo.group_test) as in the test dataset, so in this case the model produces a bad outcome. The other outcomes seem to be consistent with test dataset. The performance evaluation section will show the overal accuracy of the model.



osteo.group_test	osteo.group_predict	%Normal	%Osteopenia	%Osteoporosis
0	1	0.36	0.59	0.05
1	1	0.23	0.67	0.10
1	1	0.20	0.73	0.07
0	0	0.52	0.46	0.02
1	1	0.16	0.57	0.27
...	...	...	...	...
0	0	0.96	0.04	0.00
1	1	0.08	0.76	0.16
2	2	0.00	0.04	0.96
1	1	0.07	0.72	0.21
1	1	0.11	0.77	0.12

Figure 11. Logistic Regression outcomes

Besides prediction, the multinomial logistic regression model is also used to determine the impact of each explanatory variable on the odds ratio of the observed events of the disease. Figure 12 shows all the parameters of the model, such as intercept and respective coefficients for explanatory variables. These parameters can be used to build the regression equations for predicting purposes.

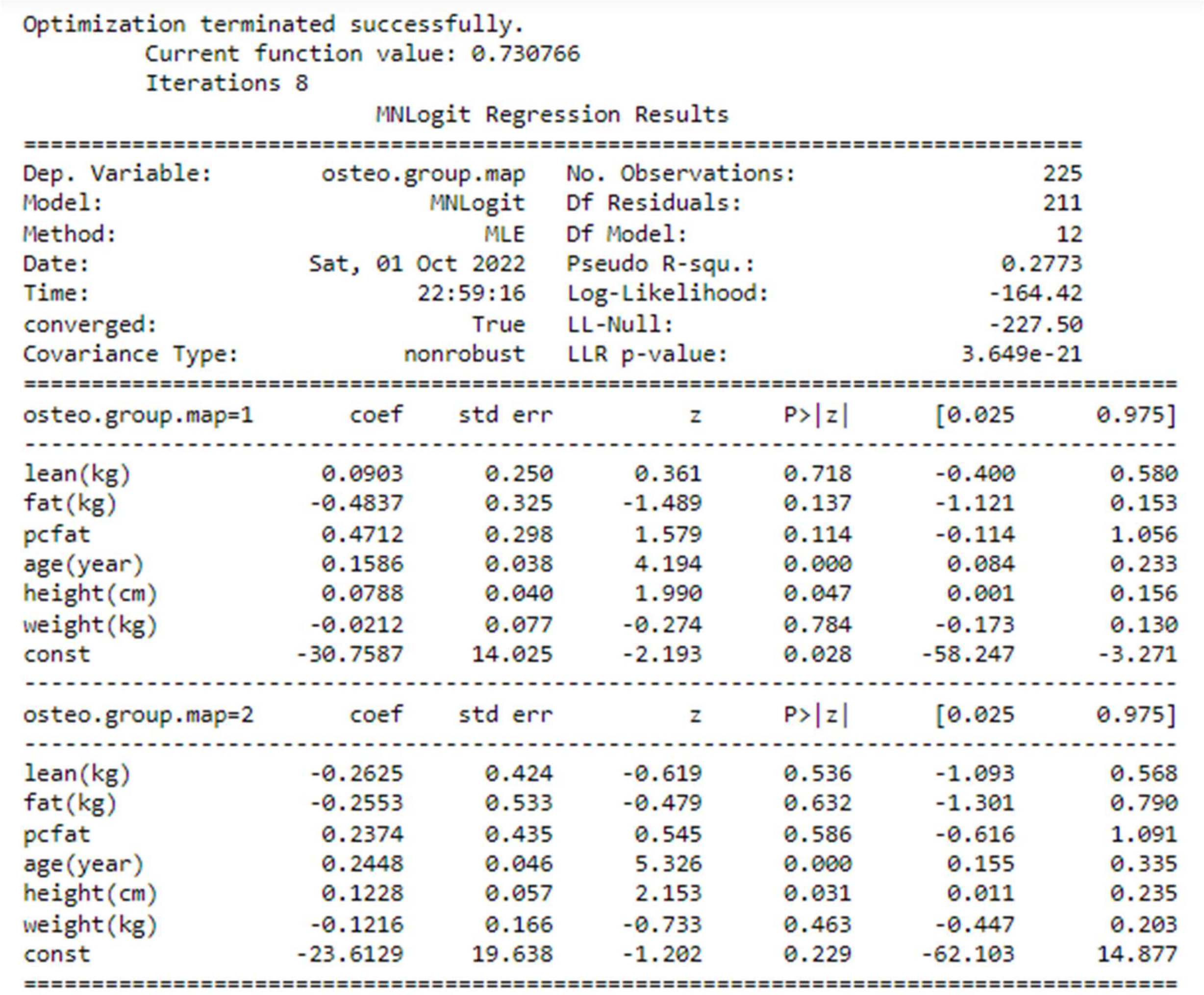


```
#intercept and coefficient of the model
print('Intercept:\n',model.intercept_)
print('Coefficient:\n',model.coef_)
print('Classes:\n',model.classes_)
```

```
Intercept:
 [ 14.11224907 -13.83627986  -0.27596921]
Coefficient:
 [[ 0.11596936  0.16089289 -0.16017283 -0.13195626 -0.06202118  0.04778544]
 [ 0.16982029 -0.26457372  0.25758949  0.02382538  0.01358119  0.02441649]
 [-0.28578965  0.10368083 -0.09741666  0.10813088  0.04843998 -0.07220193]]
Classes:
 [0 1 2]
```

Figure 12. Parameters of the logistic regression model

Another python package statsmodels is used to further assess these relationships stated above. This method’s output is different from Scikit-learn (figure 13), because it shows the coefficients against the reference group (reference group is a group with outcome is one or Normal). The coefficients represent the log of odds ratio between the probability of suffering diseases (either osteopenia or osteoporosis) vs. the probability of being normal.

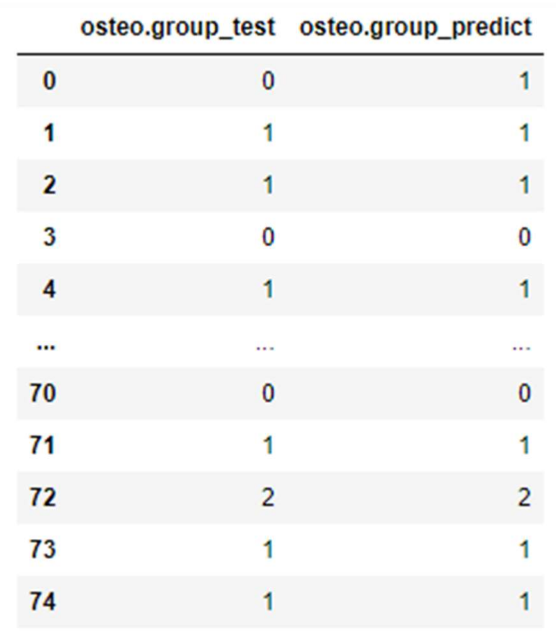


```
Optimization terminated successfully.
         Current function value: 0.730766
         Iterations 8
                          MNLogit Regression Results
==============================================================================
Dep. Variable:        osteo.group.map   No. Observations:                  225
Model:                        MNLogit   Df Residuals:                      211
Method:                           MLE   Df Model:                           12
Date:                Sat, 01 Oct 2022   Pseudo R-squ.:                  0.2773
Time:                        22:59:16   Log-Likelihood:                -164.42
converged:                       True   LL-Null:                       -227.50
Covariance Type:            nonrobust   LLR p-value:                 3.649e-21
=====================================================================================
osteo.group.map=1       coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
lean(kg)              0.0903      0.250      0.361      0.718      -0.400       0.580
fat(kg)              -0.4837      0.325     -1.489      0.137      -1.121       0.153
pcfat                 0.4712      0.298      1.579      0.114      -0.114       1.056
age(year)             0.1586      0.038      4.194      0.000       0.084       0.233
height(cm)            0.0788      0.040      1.990      0.047       0.001       0.156
weight(kg)           -0.0212      0.077     -0.274      0.784      -0.173       0.130
const               -30.7587     14.025     -2.193      0.028     -58.247      -3.271
-------------------------------------------------------------------------------------
osteo.group.map=2       coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
lean(kg)             -0.2625      0.424     -0.619      0.536      -1.093       0.568
fat(kg)              -0.2553      0.533     -0.479      0.632      -1.301       0.790
pcfat                 0.2374      0.435      0.545      0.586      -0.616       1.091
age(year)             0.2448      0.046      5.326      0.000       0.155       0.335
height(cm)            0.1228      0.057      2.153      0.031       0.011       0.235
weight(kg)           -0.1216      0.166     -0.733      0.463      -0.447       0.203
const               -23.6129     19.638     -1.202      0.229     -62.103      14.877
=====================================================================================
```

Figure 13. Statsmodel outcome

3.4.2 Random Forest Classifier

The random forest classifier is another suitable tool for predicting a multi-class response variable. As for the parameters, the number of trees in the forest is set to 50, and the function used to measure the quality of the tree split is entropy. Figure 14 shows the predicted outcomes of this model against the test set.



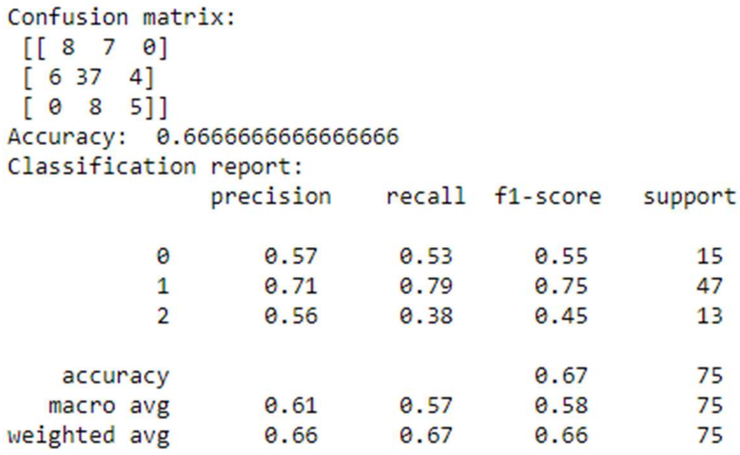
	osteo.group_test	osteo.group_predict
0	0	1
1	1	1
2	1	1
3	0	0
4	1	1
...	...	...
70	0	0
71	1	1
72	2	2
73	1	1
74	1	1

Figure 14. Random Forest Classifier outcomes

3.5 Performance Evaluation

3.5.1 Multinomial Logistic Regression

Figure 15 shows the confusion matrix and the metric for evaluating the performance of the multinomial logistic regression. The overall accuracy comes out to be 66.67% representing 50 cases out of 75 that are correctly predicted. Both precision and recall of the osteopenia group are relatively higher than the other two groups. It suggests that, given the dataset, this logistic model is highly relevant in predicting the osteopenia group.

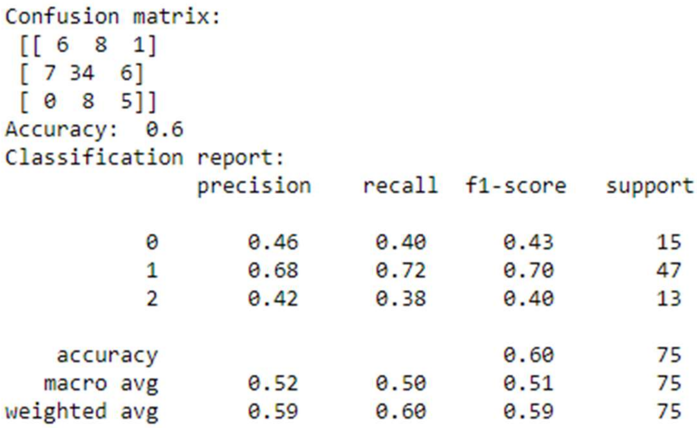


```
Confusion matrix:
 [[ 8  7  0]
 [ 6 37  4]
 [ 0  8  5]]
Accuracy:  0.6666666666666666
Classification report:
               precision    recall  f1-score   support

           0       0.57      0.53      0.55        15
           1       0.71      0.79      0.75        47
           2       0.56      0.38      0.45        13

    accuracy                           0.67        75
   macro avg       0.61      0.57      0.58        75
weighted avg       0.66      0.67      0.66        75
```

Figure 15. Performance of Logistic Regression

3.5.2 Random Forest Classifier

The evaluation result in Figure 16 shows that the overall accuracy of the model is 60% which is not significantly different from the multinomial logistic model. Precision and recall are a bit lower but follow a similar pattern of each group of participants.



```
Confusion matrix:
 [[ 6  8  1]
 [ 7 34  6]
 [ 0  8  5]]
Accuracy:  0.6
Classification report:
               precision    recall  f1-score   support

           0       0.46      0.40      0.43        15
           1       0.68      0.72      0.70        47
           2       0.42      0.38      0.40        13

    accuracy                           0.60        75
   macro avg       0.52      0.50      0.51        75
weighted avg       0.59      0.60      0.59        75
```

Figure 16. Performance of Random Forest Classifier

4. Result and Findings

Given the small dataset with 300 observations, both machine learning approaches suggest two models with the overal accuracy scores between 60% and 70%. These are really not good and effective models to predict all disease groups. However, the precision and recall of these models for the osteopenia group are still greater than 70% compared to the other groups.

One reason causing low accuracy could be the high correlation between the explanatory variables, as shown in figure 10. This is also known as the multicollinearity issue. Additionally, the result from figure 13 indicates that only p-values of age and height are less than 0.05, which is statistically significant rather than the other features. The next work is to try modeling with these two features or utilize the feature selection techniques could help to improve the accuracy.

REFERENCES

Klibanski, A, Adams-Campbell, L, Bassford, T, Blair, SN, Boden, SD, Dickersin, K, Gifford, DR, Glasse, L, Goldring, SR, Hruska, K & Johnson, SR 2001, 'Osteoporosis prevention, diagnosis, and therapy', *Journal of the American Medical Association*, 285(6): 785-795.

Sözen, T, Özışık, L & Başaran, NÇ 2017, 'An overview and management of osteoporosis', *European journal of rheumatology*, 4(1):46.

Tuan V. Nguyen, Garvan Institute of Medical Research · Osteoporosis and Bone Biology Program, DSc, PhD, https://www.researchgate.net/profile/Tuan-Nguyen-41