A picture containing text, businesscard, clipart

Description automatically generated

A picture containing text

Description automatically generated

Prt 565 – Machine learning and artificial intelligence

Assignment 1

**Lecturer: Asif Karim**

Student Name: Tai Phu Phan (s342489)

Table of Contents

[1. Problem Description 5](#_Toc115633123)

[2. Dataset Introduction 5](#_Toc115633124)

[3. Methodology 6](#_Toc115633125)

[3.1 Importing data 6](#_Toc115633126)

[3.2 Data pre-processing 7](#_Toc115633127)

[3.3 Exploratory Data Analysis (EDA) 10](#_Toc115633128)

[3.4 Machine Learning model 14](#_Toc115633129)

[3.4.1 Multinomial Logistic Regression 14](#_Toc115633130)

[3.4.2 Random Forest Classifier 16](#_Toc115633131)

[3.5 Performance Evaluation 17](#_Toc115633132)

[3.5.1 Multinomial Logistic Regression 17](#_Toc115633133)

[3.5.2 Random Forest Classifier 18](#_Toc115633134)

[4. Result and Findings 18](#_Toc115633135)

[REFERENCES 19](#_Toc115633136)

**List of Figures**

[Figure 1. Importing dataset 6](#_Toc115633137)

[Figure 2. Data types 7](#_Toc115633138)

[Figure 3. Missing data 7](#_Toc115633139)

[Figure 4. Descriptive characteristics of dataset 8](#_Toc115633140)

[Figure 5. Data after imputation 9](#_Toc115633141)

[Figure 6. Mapping response variable 9](#_Toc115633142)

[Figure 7. Barplot of osteo groups 10](#_Toc115633143)

[Figure 8. Distribution shape of features 11](#_Toc115633144)

[Figure 9. Outliers detection 12](#_Toc115633145)

[Figure 10. Correlation matrix of features 13](#_Toc115633146)

[Figure 11. Logistic Regression outcomes 14](#_Toc115633147)

[Figure 12. Parameters of the logistic regression model 14](#_Toc115633148)

[Figure 13. Statmodel outcome 15](#_Toc115633149)

[Figure 14. Random Forest Classifier outcomes 16](#_Toc115633150)

[Figure 15. Performance of Logistic Regression 16](#_Toc115633151)

[Figure 16. Performance of Random Forest Classifier 17](#_Toc115633152)

**List of Tables**

[Table 1. Dataset definition 6](#_Toc115627152)

# 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.

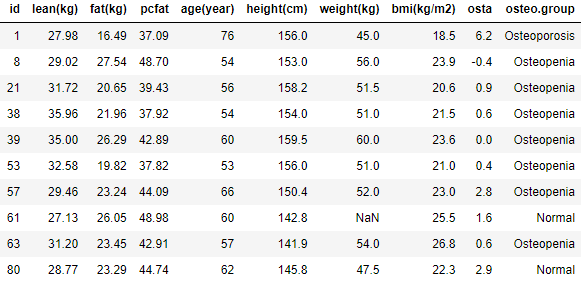


Figure 1. Importing dataset

## 3.2 Data pre-processing

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

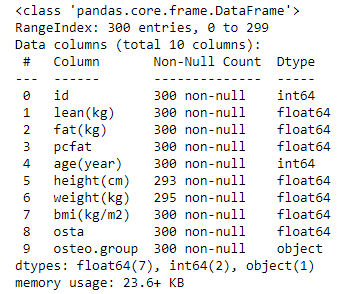


Figure 2. Data types

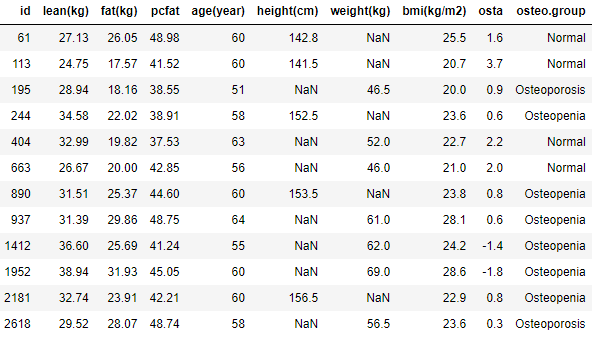


Figure 3. Missing data

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.

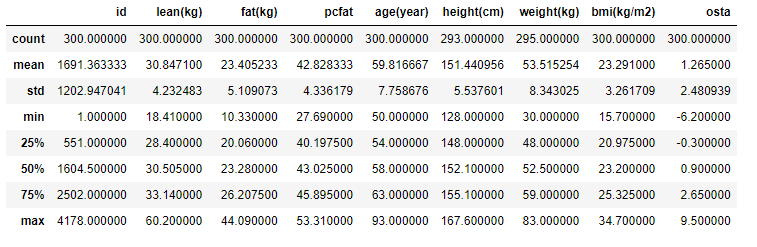


Figure 4. Descriptive characteristics of dataset

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.

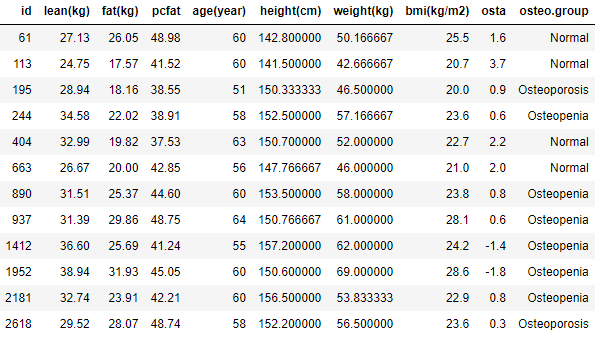


Figure 5. Data after imputation

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.

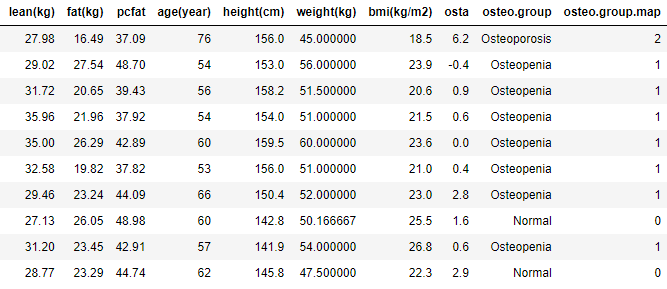


Figure 6. Mapping response variable

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%.

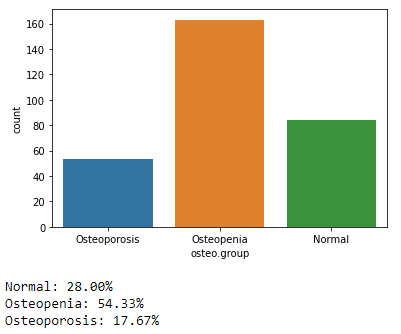


Figure 7. Barplot of osteo groups

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.

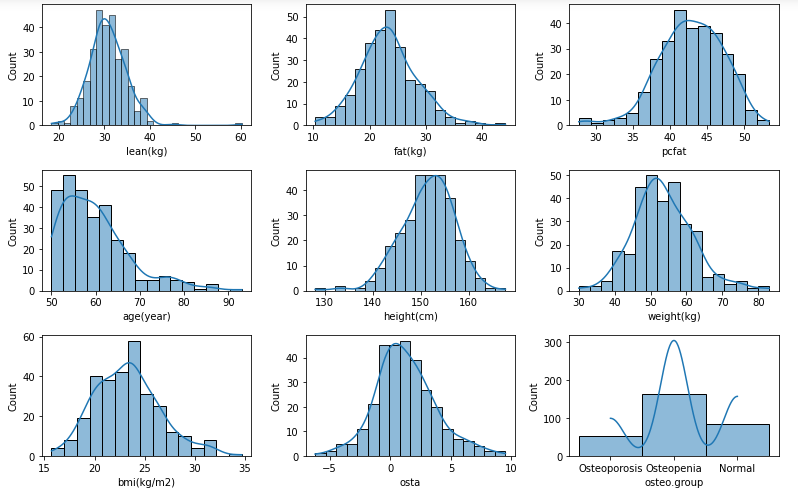


Figure 8. Distribution shape of features

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.

Chart, box and whisker chart

Description automatically generated

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.

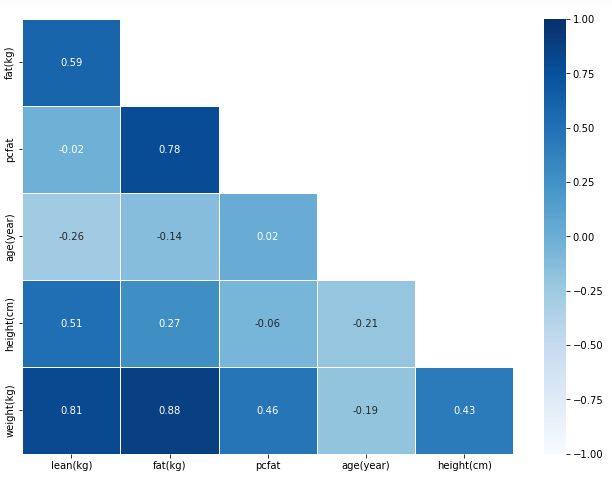


Figure 10. Correlation matrix of features

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.

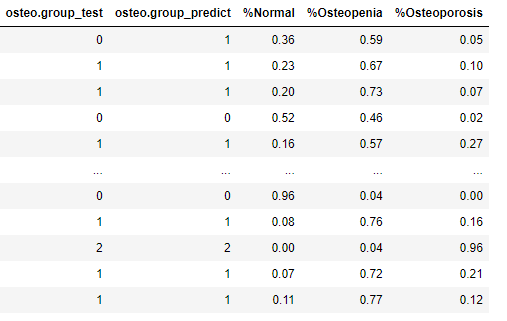


Figure 11. Logistic Regression outcomes

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.

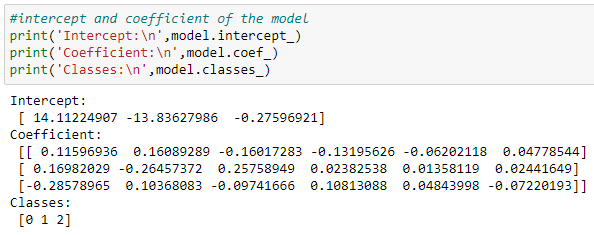


Figure 12. Parameters of the logistic regression model

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.

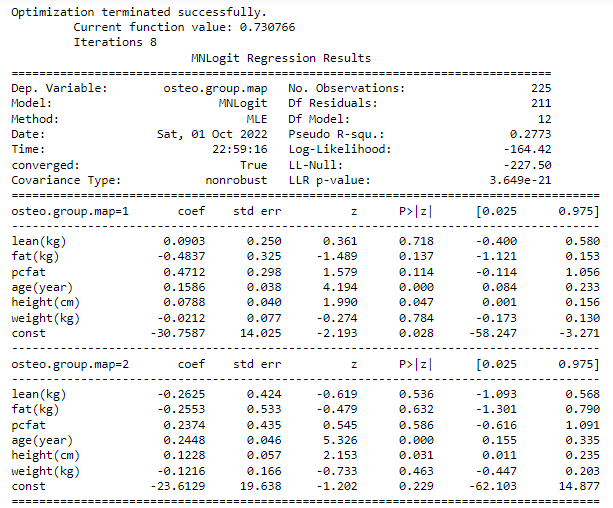


Figure 13. Statsmodel outcome

### 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.

Graphical user interface, application, Teams

Description automatically generated

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.

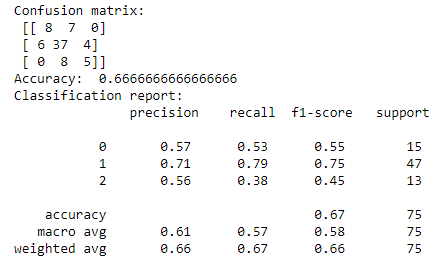


Figure 15. Performance of Logistic Regression

### 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.

Table

Description automatically generated with low confidence

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