MACHINE LEARNING IN A DEXA DATABASE OF HIV PATIENTS.

Jordi Piqué Villorbina
Master in Bioinformatic and Biostatistics UOC-UB
Area 2. Data Analysis

Nuria Perez Alvarez
Carles Ventura Royo

December 2021

FITXA DEL TREBALL FINAL

Títol del treball:	*Machine learning in a dexa data base of HIV patients*
Nom de l'autor:	*Jordi Piqué Villorbina*
Nom del consultor/a:	*Nuria Perez Alvarez*
Nom del PRA:	*Carles Ventura Royo*
Data de lliurament :	*12/2021*
Titulació o programa:	*Master Bioinformatic and Biostatistics*
Àrea del Treball Final:	*Area 2-subarea2-Data Analysis*
Idioma del treball:	*English*
Paraules clau	*DEXA,HIV,MACHINE LEARNING*

Resum del Treball

El treball parteix d'una base de dades DEXA amb informació de tres malalties, Sarcopenia, Lipodistrofia i Osteoporosis.

La informació de la base de dades és sobre pacients amb SIDA, malaltia que avui en dia encara te una alta incidència en la població. Els tractaments han millorat molt l'esperança de vida dels pacients però han augmentat també el risc de tenir alguna de les tres patologies mencionades.

Les variables de la base de dades son relacionades amb aquestes tres malalties.

El que es farà és un anàlisis descriptiu de la base de dades i una predicció de les malalties, però fen la predicció d'una malaltia mitjançant les variables de les altres dues malalties.

La predicció es farà amb els algoritmes mes coneguts de Machine Learning(ML) i es farà de manera categòrica i numèrica.

També es considerarà si la variable densitat mineral total de l'os serveix per a predir el nivell d'osteoporosis.

Es crearà un informe dinàmic amb Rmarkdown que serveixi per a fer prediccions amb altres bases de dades.

Abstract

The work is based on a DEXA database with information on three diseases, Sarcopenia, Lipodystrophy and Osteoporosis.

The information in the database is about patients with AIDS, a disease that today still has a high incidence in the population. The treatments have greatly improved the life expectancy of patients but have also increased the risk of having some of the three pathologies mentioned.

The variables in the database are related to these three diseases.

What will be done is a descriptive analysis of the database and a prediction of the diseases, but doing the prediction of one disease through the variables of the other two diseases.

The prediction will be made with the best known Machine Learning(ML) algorithms and will be done categorically and numerically.
It will also be considered whether the variable total bone mineral density is used to predict the level of osteoporosis.

A dynamic report will be created with Rmarkdown that can be used to make predictions with other databases.

Contents

1 INTRODUCTION

1.1 Context and justification of the Work

The AIDS disease (HIV) remains today one of the most important diseases for the entire world population.

Some current data on the disease are[22]:

37.7 million (30.2 million – 45.1 million) people were living with HIV worldwide in 2020.

1.5 million (1.0 million – 2.0 million) people became infected with HIV in 2020.

680,000 (480,000–1.0 million) people died of AIDS-related illnesses in 2020.

27.5 million (26.5 million – 27.7 million) people had access to antiretroviral therapy in 2020.

79.3 million (55.9 million – 110 million) people have been infected with HIV since the beginning of the epidemic.

36.3 million (27.2 million – 47.8 million) people have died from AIDS-related illnesses since the beginning of the epidemic.

Today it is a chronic disease, that in most cases can be managed perfectly with the right combination of medications. Antiretroviral treatments keep the disease controlled and, in addition, patients who receive this medical care do not spread the infection.

Although current anti-HIV drugs have fewer side effects than those used initially, they still cause some problems in patients. Some of these long-term problems are osteoporosis, lipodystrophy, and sarcopenia.

Osteoporosis is a disease that develops due to loss of bone mass. The bones of people with osteoporosis become weak and are more likely to fracture. This disease increases the risk of hip, spine and wrist fractures.

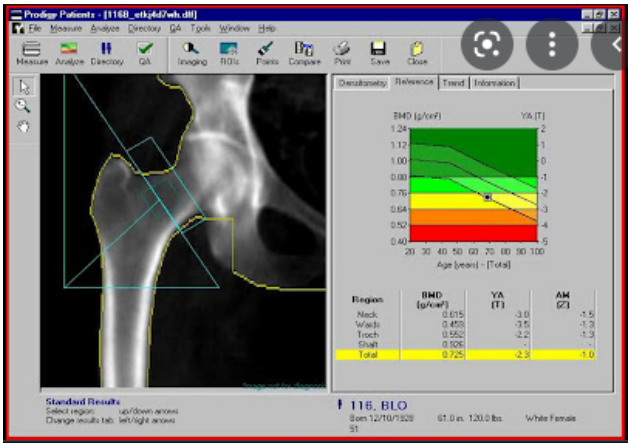
Lipodystrophy refers to changes in body fat that can affect some people with HIV. Lipodystrophy can be caused by HIV infection or by medicines used to treat HIV, but its true cause is unknown.

Sarcopenia is a progressive and generalized disease of the skeletal muscle, characterized by a decrease in muscle strength, muscle mass and finally physical performance.

To measure the risk of having a fracture, Osteoporosis, the best technique to use is DEXA (dual-energy x-ray absorptiometry).

DEXA uses a very small dose of ionizing radiation to produce images of the inside of the body, usually the lower spine (lumbar) and hips, to measure bone loss. Generally, it is used to diagnose osteoporosis, and assess an individual's risk of developing fractures due to osteoporosis.

DEXA is simple, fast, and non-invasive. It is the most commonly used and standard method for diagnosing osteoporosis.



DEXA can also distinguish the fat and muscle parts of the body. It is a very appropriate test to diagnose lipodystrophy and sarcopenia.

The objective of this work is from a DEXA database to make a descriptive analysis of the data and make qualitative and quantitative predictions with machine learning algorithms.

1.2 Objectives

- Make a descriptive and multivariate analysis of the database.
- Make first a categorical and then quantitative prediction of the three diseases from the anthropometric data of the DEXA database.
- Make a dynamic report for next results.

1.3 Approach and method followed

The diagnosis of these three diseases is made from some of the variables in the DEXA database.

- Osteoporosis ——— minTscore
- Lipodystrophy ———- FMR (Fat Mass Ratio)
- Sarcopenia ———— Apendicularleanmas

There is a lot of literature on the relationship between variables in a DEXA database and the prediction of osteoporosis.

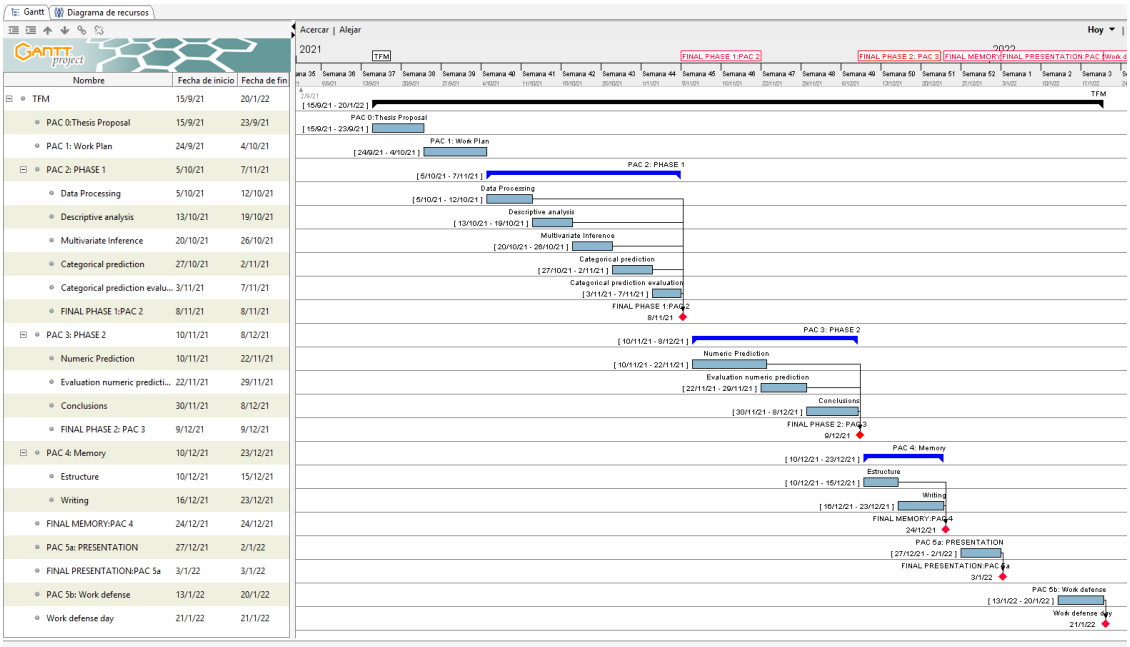
There is less work on the prediction of Lipodystrophy and Sarcopenia.

What we want to look for in this work is the relationship of each disease with the variables of the other two diseases.

We also want to study whether whole-body bone density analysis "TotalBMD" can serve as a predictor of Osteoporosis.

1.4 Task planning and timing

Work	**Time**
Bibliographic search.	7 days.
Work plan.	10 days.
Data processing.Outliers and omitted values.	8 days.
Descriptive analysis.	7 days.
Multivariate inference. PCA.	7 days.
Categorical prediction.	7 days.
Evaluation of the results of categorical prediction models.	6 days.
Numerical prediction.	13 days.
Evaluation of the results of numerical prediction models.	8 days.
Conclusions.	10 days.
Preparation of the report.	15 days.
Preparation of the presentation.	8 days.
Defense of work.	1 day.



PAC 0: Deadline 23/09/2021

- Proposal + bibliographic search.

PAC 1: Deadline 04/10/2021

- Work plan.

PAC 2: Deadline 08/11/2021

- Data processing.
- Outliers and omitted values.
- Descriptive analysis.
- Multivariate inference.
- PCA.
- Categorical prediction.
- Evaluation of the results of categorical prediction models.

PAC 3: Deadline 09/12/2021

- Numerical prediction.
- Evaluation of the results of numerical prediction models.
- Conclusions.

PAC 4: Deadline 24/12/2021

- Elaboration of the memory.

PAC 5a: Deadline 03/01/2022

- Elaboration of the presentation.

PAC 5b: Deadline 21/01/2022

- Defense of work.

1.5 Summary of products obtained

- A memory with all the information with the categorical and numerical prediction.
- A presentation of the results.
- A dynamic report with R Markdown

1.6 Description of the other chapters of the report

- Matherials&Methods
- Database Analisis
- Results
- Conclusions

2 MATHERIAL&METHODS

2.1 STADISTICAL PROGRAM

All the work is made with the program R studio and R markdown VERSION:

R version 4.1.2 (2021-11-01) – "Bird Hippie"

2.2 THE DEXA DATABASE

The database is a DEXA(Dual Energy X-ray Absorptiometry) database that comes from real patients from one hospital located in the area of Barcelona.

The name is "Totes DEXES completes_09-01-18_English_selection2", and we convert that in the "raw" data.

This data is used for diagnose Osteoporosis, Sarcopenia and Lipodistrophy.

The dataset **raw** have 1480 observations and 82 variables.

The variables of the dataset are values related to the three diseases.

As the dataset raw will be cleaned and converted to DEXA database, the information of the variables is explained later in the DEXA database information.

We mention here the three outcomes variable.

The 3 outcomes variables are Tscore_3cat, sarcopenia and Lipodistrophy.

Tscore_3cat,determines the level of bone disease in osteoporosis, osteopenia or normal.

Tscore_3cat it comes from the numeric variable minTscore, that comes from the minimum Tscore find in a patient in all the Tscore measures. Tscore measures are the bone mineral density measures.It shows how much higher or lower your bone density is than that of a healthy 30-year-old.

The cuts to diagnose the disease are:

Diagnostic	**Threshold MinTscore**
Healthy	>-1
Osteopenia	<-1 -2.5>
Osteoporosis	<-2.5

Sarcopenia determines the presence or absence of this disease.It comes from the threshold of the apendicularleanmas. This threshold is different for men and women.

Sarcopenia	**apendicularleanmas**
Men	<7
Women	<6

Lipodistrophy determines the presence or absence of this disease..It comes from the threshold of the fat mass ratio(FMR). This threshold is different for men and women.

Lipodistrophy	**FMR**
Men	>1.961
Women	>1.329

2.3 APPLIED STATISTICS

2.3.1 Correlation

Covariance is a measure of the common variability of two variables (growth of both at the time or growth of one and decrease of the other), but it is affected by the units in which each variable it is measured.

Thus, it is necessary to define a measure of the relationship between two variables, and that it is not affected by changes in the unit of measure.

The way to achieve this goal is to divide the covariance by the product of the standard deviations of each variable, since this gives a coefficient dimensionless, r, which is called the linear correlation coefficient of Pearson r.

$$\mathrm{r} = \frac{Sxy}{SxSy}$$

r is dimensionless, invariant to linear transformations. It takes values between -1 and 1.

2.3.2 PCA

Principal components are unrelated composite variables such that a few explain most of the variability of X.

PCA is used in exploratory data analysis and for making predictive models. It is commonly used for dimensionality reduction by projecting each data point into only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible. The first principal component can equivalently be defined as a direction that maximizes the variance of the projected data.

To construct this linear transformation, the covariance matrix or matrix of correlation coefficients must first be constructed.

Principal components are eigenvectors of the data's covariance matrix.

In applications it is expected that the first components explain a high percentage of the total variability.

There are ways to decide how many pca to keep. The most common are:

- **Percentage criterion:**take the number that explain more than 80%.
- **Kaiser criterion:** exclude components whose associated eigenvalues are less than 0.7.
- **Elbow criterion:** graphically, where the graph of the components makes an elbow.

It depends on the type of data, but normally we will stick with a few components.

2.3.3 Cluster Analisis

Cluster analysis is a generic term for a wide range of numerical methods for examining multivariate data with a view to uncovering or discovering groups or clusters of observations that are homogeneous and separated from other groups.

The three most common types of clustering procedures are:

Agglomerative hierarchical methods.

The classification consist of a series of partitions.This partitions are made by a series of successive fusions of the n individuals into groups.

One important thing to decide is how many groups/cluster are the best partition.

Hierarchic classifications may be represented by a two-dimensional diagram known as a dendrogram, which illustrates the fusions made a each stage of the analysis

K-means type methods.

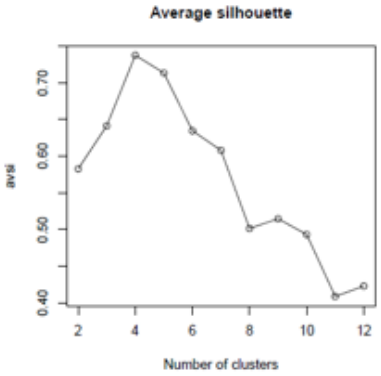
The k-means clustering technique seeks to partition a set of data into a specifiednumber of groups, k, by minimizing some numerical criterion, low values of whichare considered indicative of a "good" solution. The most commonly used approach,for example, is to try to find the partition of the n individuals into k groups, which minimizes the within-group sum of squares over all variables.

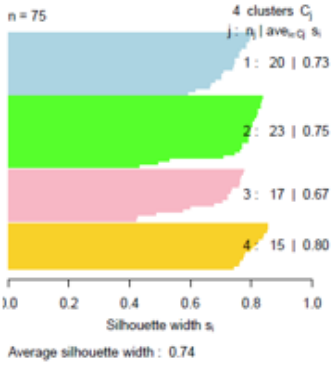
Classification maximum likelihood methods.

It is based on criteria of maximum likelihood.Assume the population consists of c subpopulations, each corresponding to a cluster of observation.

In all types of clustering it is important to decide the number of clusters. This must be calculated by specialists in each subject and by statistical information.

One method for making this decision is known as the Silhouette method. A Silhouette (value), moves between -1 and 1. If it is close to 1 it means that it is well matched, it is close to -1 it means that it is not well matched, if it is close to 0 it is halfway between two groups.





2.4 MACHINE LEARNING

Machine learning (ML) is the study of computer algorithms that can improve automatically through experience and by the use of data.

It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so.

Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

Currently, the computational improvement and data handling, together with a constant improvement of the algorithms has made it a booming topic in such a way that many investors bet on this issue, obtaining even better models.

There are a large number of algorithms depending on the type of data and the results that are to be obtained, here are the ones used in this work.

2.4.1 The k-NN algorithm

The nearest neighbors approach to classification is utilized by the kNN algorithm.

The system works from the fact that similar things can have similar properties.

The algorithm measures the distances of the nearest individuals. Once the distances are calculated, classify the individuals.

The most common distance used is Euclidean. There are others like Manhattan.

Table 5: Strengths and weaknesses of k-NN alghoritm[12]

Strengths	Weaknesses
• Simple, fast and effective • Makes no assumptions about the underlying data distribution • Fast training phase	• Does not produce a model, which limits the ability to find novel insights in relationships among features • Requires selection of an appropriate k • Slow classification phase

2.4.1.1 Choosing k The better the hyperparameter k is chosen, the better the model will become generalized.

s-variance tradeoff **: Balance between overfitting and underfitting the model**

- k **large**: More underfitting, reducing model variance, but increasing bias of ignoring important patterns.
- k **small**: More overfitting, increasing model variance (outlayers create errors), but decreasing bias.

The choice of k depends on the difficulty of the problem and the number of training data. Common practice is to initialize for $\sqrt{n}$. The larger the dataset is, less important is k.

2.4.2 Naive Bayes

Bayessian classifiers use training data (variables) to calculate the probability of output.

Used in:

- Text classification (spam / jam).
- Detection of problems in computer networks.
- Medical diagnoses given some symptoms.

Used in problems where the information of certain variables must be taken into account at the same time to calculate the probability of the output.

Some ML algorithms ignore variables with small effects. In Bayesian methods, the set of many variables with small effects can influence the prediction.

2.4.2.1 Basic concepts (Bayesian methods) Bayesian probability: The estimated probability (likelihood) of an event must be based on the evidence that an event occurs, given multiple attempts.

Bayesian methods help to estimate this probability.

2.4.2.2 Probability Probability of an event:

$$P(A) = \frac{\text{nº times of } A}{\text{nº totals times}}$$

Sum of probabilities of all events is 1.

2.4.2.3 Joint probability

For non-exclusive events (events can happen at the same time).

$$P(A \cap B) = P(A) \times P(B)$$

2.4.2.4 Conditional probability, Bayes theorem

The relationship between dependent events is calculated with conditional probability.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

Probability of A given that B has passed.

$$\text{Post Prob} = \frac{\text{Likelihood} \times \text{Prior Prob}}{\text{Marginal Likelihood}}$$

Table 6: Strengths and weaknesses of Naive Bayes alghoritm[12]

Strengths	Weaknesses
• Simple, fast and effective • Can work with missing and noisy data • You don't need a lot of data, and it works well with many • Easy obtaining of the probability of the event	• It is assumed that the features are of equal importance, and are independent • If we have many numerical variables, it is not ideal • The estimated probability is less accurate than the predicted class

2.4.3 Artificial Neuronal Network Algorithm

Black Box Models: The process of passing inputs to outputs is mathematically so complicated, that it is difficult to gain insights into how predictions are made.

An ANN creates relationships between inputs and output similar to how a network of biological neurons does.

Examples where ANNs are used:

- Voice recognition, writing.
- IoT (internet of things) automation such as autopilot drones, cars …
- Sophisticated, scientific, social or economic models.

They can be applied to many types of data and models, such as classification, regression, and unsupervised systems.

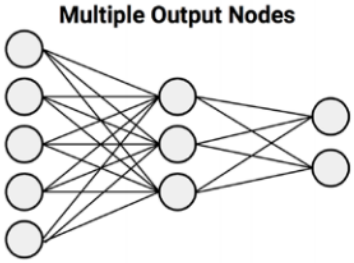
The best results are obtained for problems where the input and output data are well defined, but their relationship is complex.

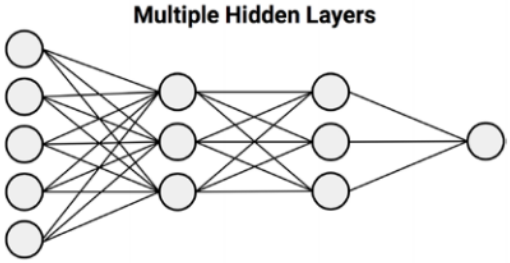
Neural networks can be defined by the following characteristics:

- **Activation function**: Transforms the signals of the neuron inputs into a single output.
- **Network architecture**: Number of neurons per layer, number of layers and connections.

- **Training Algorithm**: How to train the model to learn the parameters.

If there is more than one hidden layer, the network is called the deep neural network DNN.





Thanks to backpropagation, DNN feedforward is currently used in data mining:

Table 7: Strengths and weaknesses of feedforward NN[12]

Strengths	Weaknesses
• Used for classification and prediction • Able to model more complex patterns than any other algorithm • Makes few assumptions about how the data relates	• Very expensive computationally, slow to train, especially if the topology is complex • Easy to create overfitting in training data • Black box model, very difficult to interpret

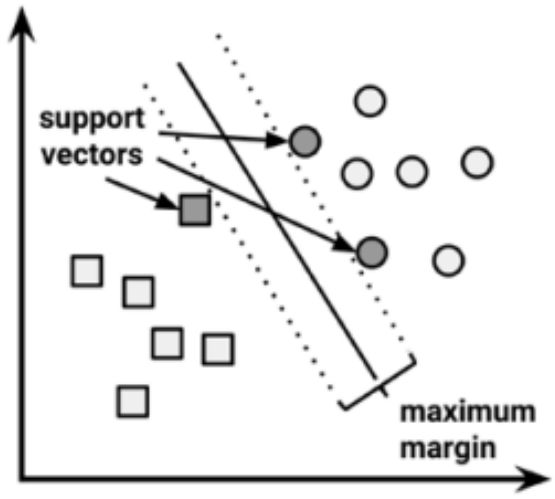
2.4.4 Vector Machine Support Algorithms

Divide the features into the hyperplane created by the inputs. Combination regression and NN.

Used in:

- Classification of genes (bioinformatics) to identify genetic diseases.
- Text categorization.
- Detection of infrequent events (earthquakes, security …).

2.4.4.1 Classification with hyperplanes The separation between classes is made from **Maximum Margin Hyperplane** (MMH). The line that creates the greatest separation will be the one that best generalizes.



2.4.4.2 Linear data If the classes can be separated linearly, the MMH is the farthest line at the *convex hull* (geometric limit of the set of points). It can be solved from quadratic optimization.

2.4.4.3 Data that cannot be separated linearly **Slack variable**: Creation of a margin that allows some points to pass. A cost C is applied to the crossing points, with the distance of the erroneous points ξ to the MMH, and it is wanted to minimize the cost:

C is a hyperparameter that must be adjusted:

- The larger C, the more separation will be attempted (more overfitting).
- The smaller C, the higher the margin, accepting more points on the wrong sides (more underfitting).

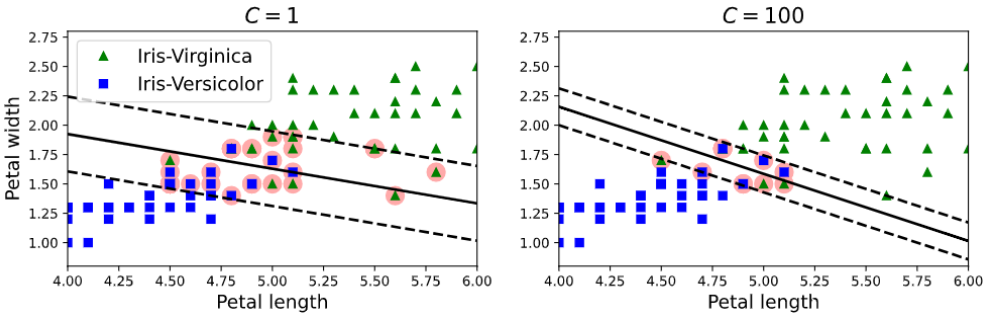


Table 8: Strengths and weaknesses SVMs[12]

Strengths	Weaknesses
• Classification and regression • It is not heavily influenced by noisy data, and does not usually create overfitting • Easier to use than ANN • Popular for good accuracy, used in competitions	• Finding the best model requires testing hyperparameters and kernels • It can be slow to train, especially if you have a lot of great features • Complex black model difficult or impossible to interpret

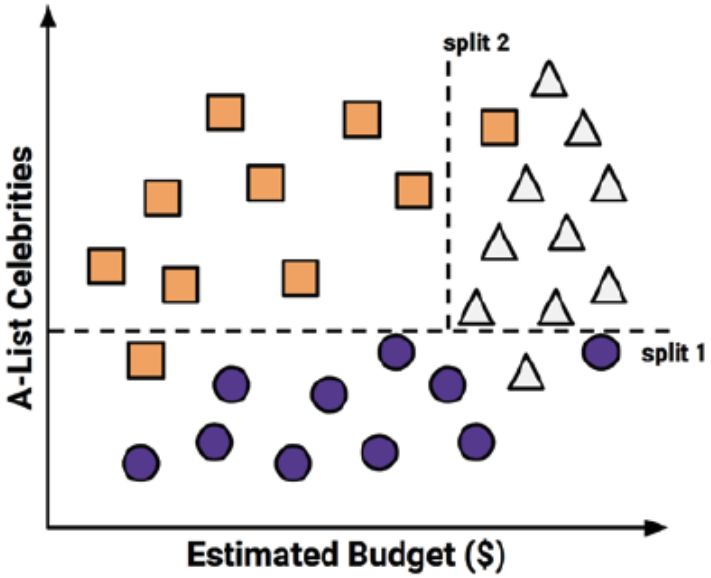
2.4.5 Decision Trees

Decision tree learners are powerful classifiers, which utilize a tree structure to model the relationships among the features and the potential outcomes.

A decision tree classifier uses a structure of branching decisions, which channel examples into a final predicted class value.

This provides tremendous insight into how and why the model works or doesn't work well for a particular task.

Decision trees are built using a heuristic called recursive partitioning. This approach is also commonly known as divide and conquer because it splits the data into subsets, which are then split repeatedly into even smaller subsets, and so on and so forth until the process stops when the algorithm determines the data within the subsets are sufficiently homogenous, or another stopping criterion has been met.



2.4.5.1 The C5.0 decision tree algorithm The first challenge that a decision tree will face is to identify which feature to split upon. The degree to which a subset of examples contains only a single class is known as purity, and any subset composed of only a single class is called pure

There are various measurements of purity that can be used to identify the best decision tree splitting candidate. C5.0 uses entropy,

$$\text{Entropy}(S) = \sum_{i=1}^{c} -p_i log_2(p_i)$$

A decision tree can continue grow indefinitely, choosing splitting features and dividing the data into smaller and smaller partitions until each example is perfectly classified or the algorithm runs out of features to split on. For that problem, solutions of pre-pruning and post-pruning can be made.

One of the benefits of the C5.0 algorithm is that it is opinionated about pruning— it takes care of many decisions automatically using fairly reasonable defaults.

Table 9: Strengths and weaknesses Decision Trees[12]

Strengths	Weaknesses
• An all-purpose classifier that does wellon most problems • Highly automatic learning process,which can handle numeric or nominal features, as well as missing data • Excludes unimportant features • Can be used on both small and large datasets • Results in a model that can be interpreted without a mathematical background (for relatively small trees • More efficient than other complex	• Decision tree models are often biased toward splits on features having a large number of levels • It is easy to overfit or underfit the model • Can have trouble modeling some relationships due to reliance on axis-parallel splits • Small changes in the training data can result in large changes to decision logic • Large trees can be difficult to interpret and the decisions they make may seem counterintuitive

2.4.6 Ensemble Methods

Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model.

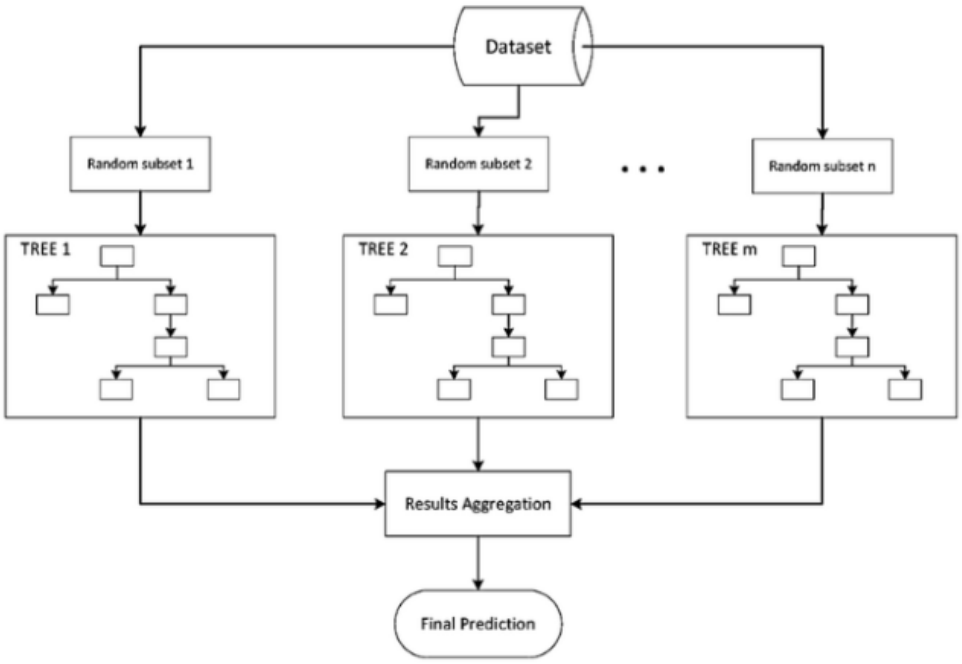
2.4.6.1 Boosting Boosting is an ensemble modeling technique that attempts to build a strong classifier from the number of weak classifiers. It is done by building a model by using weak models in series. Firstly, a model is built from the training data. Then the second model is built which tries to correct the errors present in the first model. This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models are added.

First, the resampled datasets in boosting are constructed specifically to generate complementary learners. Second, rather than giving each learner an equal vote, boosting gives each learner's vote a weight based on its past performance. Models that perform better have greater influence over the ensemble's final prediction.

Boosting will result in performance that is often quite better and certainly no worse than the best of the models in the ensemble.

A boosting algorithm called AdaBoost or adaptive boosting was proposed by Freund and Schapire in 1997

2.4.6.2 Bagging BAGGing, or Bootstrap AGGregating. BAGGing gets its name because it combines Bootstrapping and Aggregation to form one ensemble model. Given a sample of data, multiple bootstrapped subsamples are pulled. A Decision Tree is formed on each of the bootstrapped subsamples. After each subsample Decision Tree has been formed, an algorithm is used to aggregate over the Decision Trees to form the most efficient predictor.



2.4.6.3 Random Forest Combines the base principles of bagging with random feature selection to add additional diversity to the decision tree models.

Random forests combine versatility and power into a single machine learning approach. As the ensemble uses only a small, random portion of the full feature set, random forests can handle extremely large datasets, where the so-called "curse of dimensionality" might cause other models to fail. At the same time, its error rates for most learning tasks are on par with nearly any other method.

Table 10: Strengths and weaknesses Random Forest[12]

Strengths	Weaknesses
• An all-purpose model that performs well on most problems • Can handle noisy or missing data as well as categorical or continuous features • Selects only the most important features • Can be used on data with an extremely large number of features	• Unlike a decision tree, the model is not easily interpretable • May require some work to tune the model to the data

2.4.7 Linear Regression

It uses linear regression to predict models. Normally is used multiple linear regression.

To make the calculations, the least squares estimate is used

$$\sum(y_i - \hat{y}_i) = \sum -e_i^2$$

and then it's possible to predict the values of a from the values of b.

$$a = y - bx$$

Table 11: Strengths and weaknesses Regression Methods[12]

Strengths	Weaknesses
• By far the most common approach for modeling numeric data • Can be adapted to model almost any modeling task • Provides estimates of both the strength and size of the relationships among features and the outcome	• Makes strong assumptions about the data • The model's form must be specified by the user in advance • Does not handle missing data • Only works with numeric features,so categorical data requires extra processing • Requires some knowledge of statistics to understand the model

Also the correlation can be used to measure the results of the predictions.

Starting from linear regression there are many different types of algorithms with different types of regression.

2.4.8 Regularised Linear Models

Regularize the model is a good way to reduce overfitting

2.4.8.1 Ridge Regression Ridge Regression (also called Tikhonov regularization) is a regularized version of Linear Regression: a regularization term equal to $\alpha\Sigma i = 1n\theta i2$ is added to the cost function.

This forces the learning algorithm to not only fit the data but also keep the model weights as small as possible. Note that the regularization term should only be added to the cost function during training. Once the model is trained, you want to evaluate the model's performance using the un regularized performance measure.

The hyperparameter controls how much you want to regularize the model. If = 0 then Ridge Regression is just Linear Regression

2.4.8.2 Lasso Regression Least Absolute Shrinkage and Selection Operator Regression (simply called Lasso Regression) is another regularized version of Linear Regression: just like Ridge Regression, it adds a regularization term to the cost function, but it uses the 1 norm of the weight vector instead of half the square of the 2 norm.

An important characteristic of Lasso Regression is that it tends to completely eliminate the weights of the least important features.

2.4.9 Caret Package

The caret package (short for classification and regression training) contains functions to streamline the model training process for complex regression and classification problems.[24]

The package utilizes a number of R packages but tries not to load them all at package start-up.

One of the primary tools in the package is the train function which can be used to:

- evaluate, using resampling, the effect of model tuning parameters on performance
- choose the optimal model across these parameters
- estimate model performance from a training set

To split the data in two groups there is the function **createDataPartition**.

To change the candidate values of the tuning parameter, the **tuneLength** or **tuneGrid** arguments can be used.

The tuneLength argument controls how many parameters are evaluated.

The tuneGrid argument is used when specific values are desired.

To modify the resampling method, a **trainControl** function is used.

Most of the algorithms can be made and you can insert the metric for measure the model and you can find most of the results automatically.

There are a lot of options for predictions, in classification and regression, see (*caretpackage*)

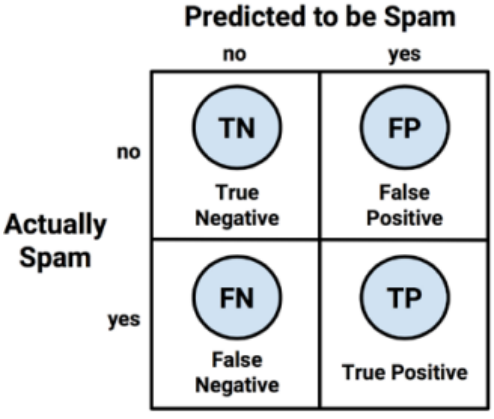
2.5 EVALUATING MODEL PERFORMANCE

2.5.1 Confusion matrix

Dimension tables k *timesk* classes, where rows is number of current classes, columns number of predicted cases. Correct predictions are found on the main diagonal.

	P1	P2	P3
A1	O	x	x
A2	x	O	x
A3	x	x	O

The interest rate is called **positive**, the rest **negative**.



2.5.1.1 Confusion matrix performance

Accuracy is the number of correct predictions among the number of predictions made:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

2.5.1.2 Sensitivity and specificity

Measures that control the aggressiveness of the algorithm.

- **Sensitivity** (true positive rate): Proportion of correctly classified positive examples.

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- **Specificity** (true negative rate): Proportion of correctly classified negative examples.

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

2.5.2 Kappa statistic

Adjusts the accuracy by counting the possibility of a correct prediction by randomness.

Important for data with **class imbalance**: avoids the problem of choosing the most common class to increase accuracy.

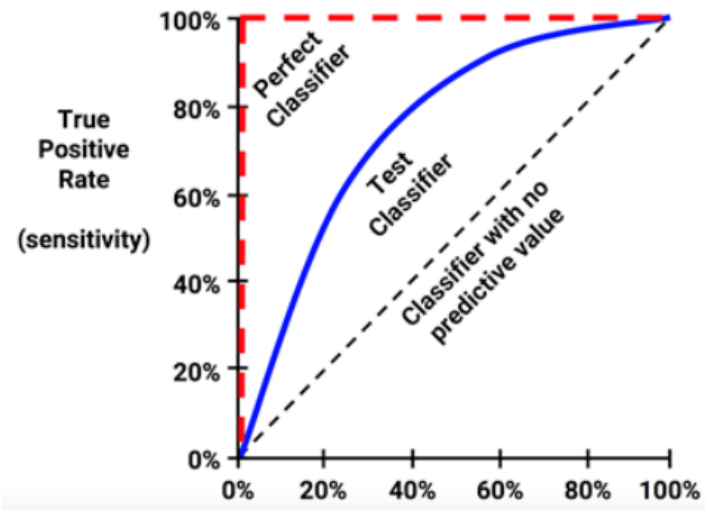
Formula (Kappa of Cohen):where $P(a)$ is the agreement ratio, and $P(b)$ is the expected agreement between the classifier and current values.

$$k = \frac{P(a) - P(e)}{1 - P(e)}$$

- Poor agreement: $k < 0.20$
- Fair agreement: $0.20 < k < 0.40$
- Moderate agreement: $0.40 < k < 0.60$
- Good agreement: $0.60 < k < 0.80$
- Very good agreement: $0.80 < k < 1.00$

2.5.3 Roc Curves and AUC

Receiver Operating Characteristic (ROC) curve: Trade-off between TP detection, avoiding FP. Equivalent to make sensitivity vs (1 - specificity).



The area under the ROC curve (AUC) is a statistic, which can be used to compare models. However, keep in mind that very different curves can have the same AUC. General convention of interpretation (although it varies according to task):

- Outstanding: 0.9 to 1.0
- Good: 0.8 to 0.9
- Fair: 0.7 to 0.8
- Poor: 0.6 to 0.7
- No discr: 0.5 to 0.6

2.5.4 Root Mean Squared Error(RMSE)

When the outcome is a number, the most common method for characterizing a model's predictive capabilities is to use the root mean squared error(RMSE).

The mean squared error (MSE) is calculated by squaring the residuals and summing them.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|(y_i - \hat{y}_i)^2}$$

The RMSE is then calculated by taking the square root of the MSE so that it is in the same units as the original data.

The value is usually interpreted as either how far(on average) the residuals are from zero or as the average distance between the observed values and the model predictions.

2.5.5 Mean Absolute Error(MAE)

For the numerical predictions it is also used this statitical to measure the different algorithms.

This measurement is called the mean absolute error (MAE). The equation for MAE is as follows, where n indicates the number of predictions and e_i indicates the error for prediction i:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|e_i$$

The error is just the difference between the predicted and actual values.

3 DATABASE ANALISIS

3.1 DEXA DATABASE

As we have seen this results comes from a DEXA (dual energy x-ray absorptiometry) analisis and are for the study of three pathologies Osteoporosis(Bone disease),Lipodistrophy(fat disease) and Sarcopenia(lean disease).

We have seen that presence or absence of each of the three pathologies it comes from one related variable.

DIAGNOSTIC	NUMERIC RELATED VARIABLE
OSTEOPOROSIS	minTscore
LIPODISTROPHY	FMR
SARCOPENIA	apendicularleanmas

Then the presence or absence of this pathologies comes from the cut-off of this variables and we have the three qualitative variables, Tscore_3cat, Lipodistrophy and Sarcopenia.

The firts thing we see is that minTscore and Tscore_3cat doesn't have values.

The mintscore is the variable used for classify the patients in Osteoporosis/Osteopenia/normal. It's the minimum value of all the Tscores.

We created the values for both variables of bone diseases.

The other two diseases that contains the raw data are Sarcopenia and Lipodistrophy.

Sarcopenia is classified in illness or not from the variable apendicularleanmas, which formula is:

$$\text{apendicularleanmas} = \frac{BothAlg + BothLLg}{Height^2 * 1000}$$

Lipodistrophy is classified in illness or not from the variable Fat mass ratio(FMR), which formula is:

$$\text{FMR} = \frac{TFp}{BothLFp}$$

The variables appendicularleanmas and FMR are verified to be correct.

As the difference is very slow we keep the same data.

Three patients have some problems with the data, errors, etc. They are patients in row 378, 385 and 500. They are put out of the database.

We look for the individuals that have missing values.There are 15 males and 7 females.

As in this group of patients the main variables follow a normal distribution we put out this group. The summary data of this group is in appendix 1.

We put the name "dexa" to the database.

We change the number of the patient to ID.

We put out the individuals with duplicated ID.

The order of the variables have been changed.

We create a new variable related to osteoporosis,"Tscore_2cat" which is "Pathology" that includes osteoporosis, and "Healthy" that is without disease, and includes normal and Osteopenia, because Osteopenia is not considered a disease.

Regarding Sarcopenia and Lipodistrophy we put the cut points for both diseases and classify them in two groups, “Healthy” or “Pathology”.

We transform some variables to numeric:Age,RAFg,RALg,LAFg,LALg,BothAFg,BothAlg,RLFg,RLLg,LLFg,LLLg, BothLFg,BothLLg,TFg,TLg,TotalFg,TotalLg.

The DEXA database has these groups variables:

General information and anthropometric variables

Are the variables related to gender,age,weight and height.

Fat/lean variables

Are variables that measure the quantity of fat and lean in arms, legs and in both.

Bone measures with T and Z values

Is the bone mineral density in different parts of the body, with T and Z score. The more importants parts are the neck, wards and trochanter of femur, and the lumbar spine values.

Disease and Calculated variables

Are the three diseases variables, with the related numerical variables and other variables calculated with a formula but coming from these.

All the variables can be seen in appendix 2.

We can see a summary of all the numeric variables in the DEXA database.

	Min	X25.	Mean	Median	X75.	Max	SD
ID	1417	200252.8	2879504	353519.5	573142.5	18922959	5397966
gender_num	1	1	1.24	1	1	2	0.43
Age	17	39	45.93	46	53	81	10.53
Age_cat	0	0	0.37	0	1	1	0.48
Height	1.4	1.64	1.7	1.71	1.77	1.93	0.09
Weight	34.6	60.51	69.61	69	78.04	120.5	12.59
RAFp	3.7	9.93	18.41	17	24.8	65.6	10.64
RAFg	49	353.25	697.86	614	940.75	4374	458.56
RALg	1125	2350.25	2890.28	2930	3411.5	9316	801.54
LAFp	2.2	9.83	18.52	17	24.8	63	10.71
LAFg	45	351	689.82	609	923.75	3532	458.45
LALg	299	2295.5	2820.45	2847.5	3321.75	6659	767.56
BothAFp	3.6	9.8	18.48	17.05	24.8	64.4	10.67
BothAFg	95	690.25	1385.05	1215.5	1865.5	7907	916.84
BothALg	2271	4650.75	5709.22	5792.5	6719	13317	1539.24
RLFp	3.8	11.6	20.47	18.8	27.4	63	10.96
RLFg	244	1174	2261.47	1989.5	3032	12570	1438.87
RLLg	3719	6841.75	8056.28	8190	9272.75	14476	1754.03
LLFp	3.8	11.53	20.42	18.8	27.3	63.8	10.94
LLFg	239	1176.75	2260.73	2014.5	3026.25	12570	1431.48
LLLg	3815	6737.75	8060.39	8207.5	9272.25	14428	1762.91
BothLFp	3.8	11.5	20.41	18.8	27.3	63.3	10.91
BothLFg	484	2350.5	4516.51	3993.5	6041.25	25140	2860.27
BothLLg	7776	13590.25	16115.67	16394	18525.75	28904	3493.45
TFp	4.6	20.8	28.66	29.15	36.5	59.4	10.58
TFg	1006	6585.25	10530.59	10096.5	13836.25	34163	5085.45
TLg	12471	21258.75	24343.96	24646	27381.5	59172	4658.89
TotalFp	4.2	16.7	24.08	24	30.3	56.6	9.61

	Min	X25.	Mean	Median	X75.	Max	SD
TotalFg	2023	10680.5	16742.65	15884	21384.5	52915	8089.18
TotalLg	25056	43390	49917	50920	56361	88914	9725.76
L1BMD	0.58	0.94	1.05	1.04	1.14	1.79	0.15
L1T	-4.5	-1.7	-0.85	-0.9	-0.1	3.4	1.27
L1Z	-3.7	-1.4	-0.56	-0.6	0.2	4.6	1.21
L2BMD	0.12	1.01	1.12	1.11	1.23	1.68	0.17
L2T	-5.5	-1.8	-0.91	-1	0	3.6	1.35
L2Z	-4.8	-1.5	-0.62	-0.7	0.2	4.1	1.29
L3BMD	0.34	1.02	1.13	1.13	1.24	1.76	0.17
L3T	-4.9	-1.8	-0.81	-0.9	0.1	4.4	1.39
L3Z	-4.2	-1.5	-0.53	-0.6	0.3	4.6	1.36
L4BMD	0.62	0.99	1.1	1.09	1.21	1.76	0.17
L4T	-4.8	-2	-1.05	-1.2	-0.2	4.3	1.39
L4Z	-4.3	-1.7	-0.77	-0.9	0.1	4.5	1.35
L1L4BMD	0.59	0.99	1.1	1.09	1.2	1.67	0.15
L1L4T	-4.9	-1.8	-0.89	-1	-0.1	3.7	1.28
L1L4Z	-4	-1.5	-0.6	-0.7	0.1	3.6	1.23
L2L4BMD	0.59	1.01	1.12	1.11	1.22	1.71	0.16
L2L4T	-5.1	-1.9	-0.93	-1	-0.1	3.9	1.32
L2L4Z	-4.4	-1.5	-0.64	-0.7	0.1	4	1.28
NeckFBMD	0.51	0.84	0.94	0.93	1.03	1.74	0.14
NeckFT	-4	-1.6	-0.85	-1	-0.2	5.1	1.08
NeckFZ	-3.1	-0.9	-0.28	-0.3	0.3	5.4	0.93
WardsBMD	0.34	0.66	0.78	0.76	0.88	1.79	0.16
WardsT	-4.5	-2.2	-1.32	-1.4	-0.5	6.4	1.26
WardsZ	-3.4	-1.2	-0.5	-0.6	0.1	6.9	1.08
TrochBMD	0.35	0.7	0.79	0.78	0.87	1.61	0.13
TrochT	-4.3	-1.8	-0.97	-1.1	-0.2	6.2	1.17
TrochZ	-3.8	-1.4	-0.65	-0.7	0	6.2	1.07
TotalFBMD	0.3	0.88	0.97	0.97	1.06	1.67	0.14
TotalFT	-4	-1.5	-0.73	-0.8	0	4.1	1.09
TotalFZ	-3.3	-1	-0.3	-0.4	0.3	5.2	0.99
TotalBMD	0.72	1.09	1.16	1.15	1.23	1.9	0.11
BMI	14.04	21.44	23.94	23.62	26.08	40.81	3.6
BMI_cat	0	1	1.37	1	2	3	0.65
FMI	0.78	3.6	5.85	5.53	7.46	19.71	2.98
FFMI	10.71	15.49	17.09	17.2	18.53	27.85	2.33
Apendicularleanmas	4.12	6.59	7.45	7.52	8.27	11.83	1.24
FMR	0.47	1.12	1.67	1.44	2.03	8.88	0.8
FTrunkgFLegsg	0.26	0.56	0.63	0.63	0.7	0.99	0.1
Indexdistributionfat	0.36	1.37	2.08	1.83	2.57	8.11	1
FtrunkpFlimbsp	0.36	0.56	0.67	0.64	0.74	1.83	0.17
FtrunkgFtotalg	0.26	0.56	0.63	0.63	0.7	0.99	0.1
FLegsgFtotalg	0.05	0.19	0.27	0.27	0.33	0.55	0.09
FlimbsgFtotalg	0.11	0.27	0.35	0.35	0.41	0.74	0.1
LLegFgBMI	14.59	50.52	91.89	84.72	125.22	344.28	50.27
LLegFpBMI	0.15	0.5	0.85	0.79	1.11	2.16	0.44
LipoSarcop	0	1	0.82	1	1	1	0.38
phenotype	1	5	8.64	6	14	16	4.54
minTscore	-4.9	-2.2	-1.47	-1.5	-0.8	2.4	1.04

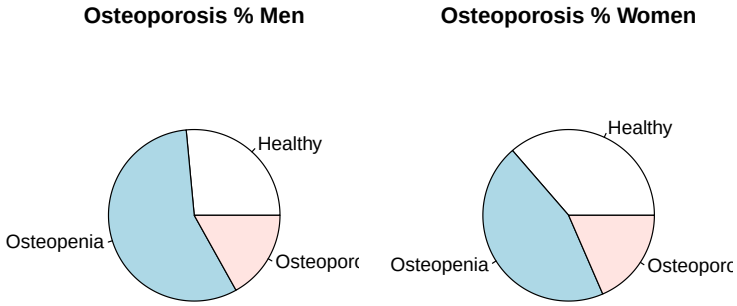
3.2 MAIN VARIABLES

The main variables are the presence or absence of the three diseases and age and weight. We add also total bone mineral density.

The presence or absence in the three diseases for men and women in the DEXA database is as follows:

Osteoporosis

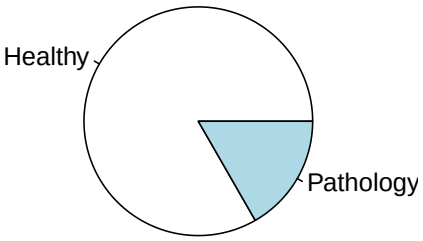
	No Bone disease	Osteopenia	Osteoporosis
Men	291.000	621.000	186.000
Women	128.000	159.000	65.000
Men%	0.265	0.566	0.169
Women%	0.364	0.452	0.185



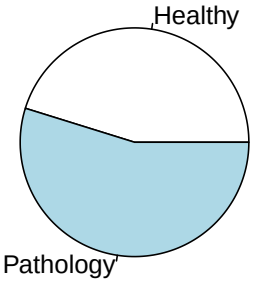
Sarcopenia

	No Sarcopenia	Sarcopenia
Men	915.000	183.000
Women	159.000	193.000
Men%	0.833	0.167
Women%	0.452	0.548

Sarcopenia % Men



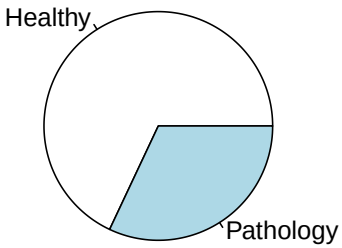
Sarcopenia % Women



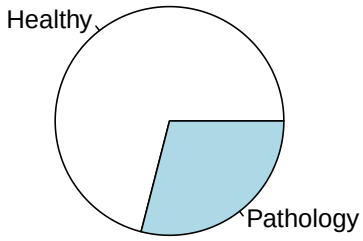
Lipodistrophy

	No Lipodistrophy	Lipodistrophy
Men	746.00	352.00
Women	251.00	101.00
Men%	0.68	0.32
Women%	0.71	0.29

Lipodistrophy % Men



Lipodistrophy % Women

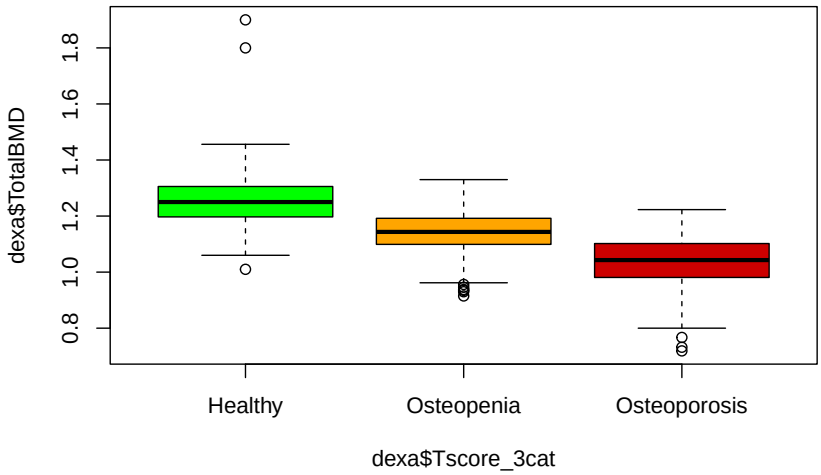


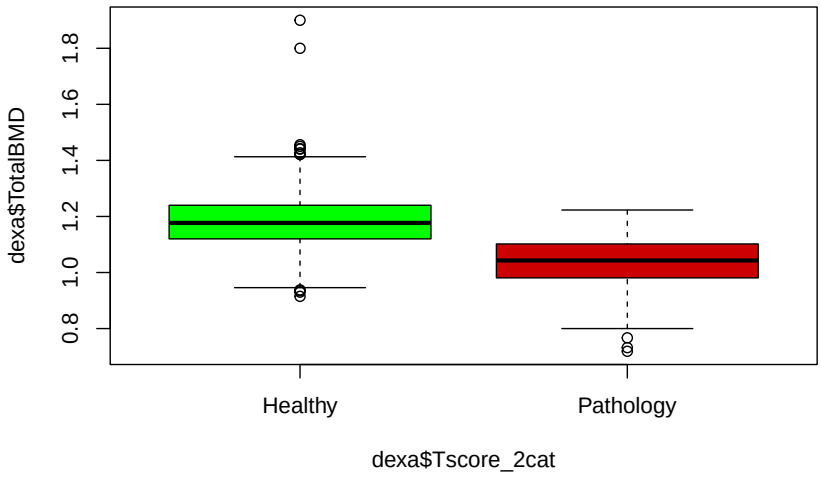
TotalBMD

A part from the the three main variables regarding to the three pathologies,we consider also that in Osteoporosis TotalBMD has to be considered because it should be a clear value of the presence/absence of this disease.

TotalBMD is a numeric data of all the body mineral density. We will take as an answer variable too.

We can see the distribution of TotalBMD regarding Osteoporosis with the three categories and two categories Osteoporosis disease.



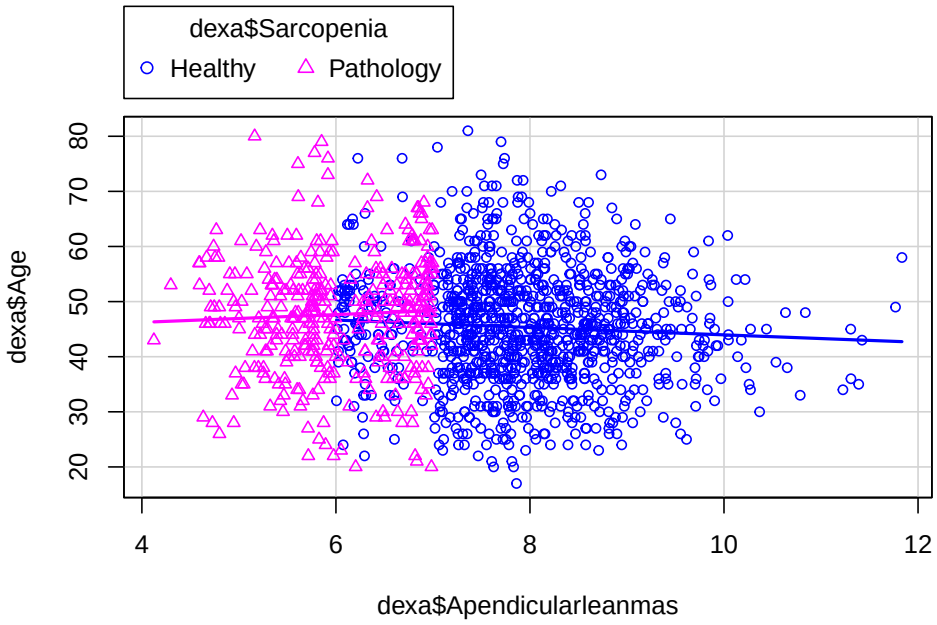


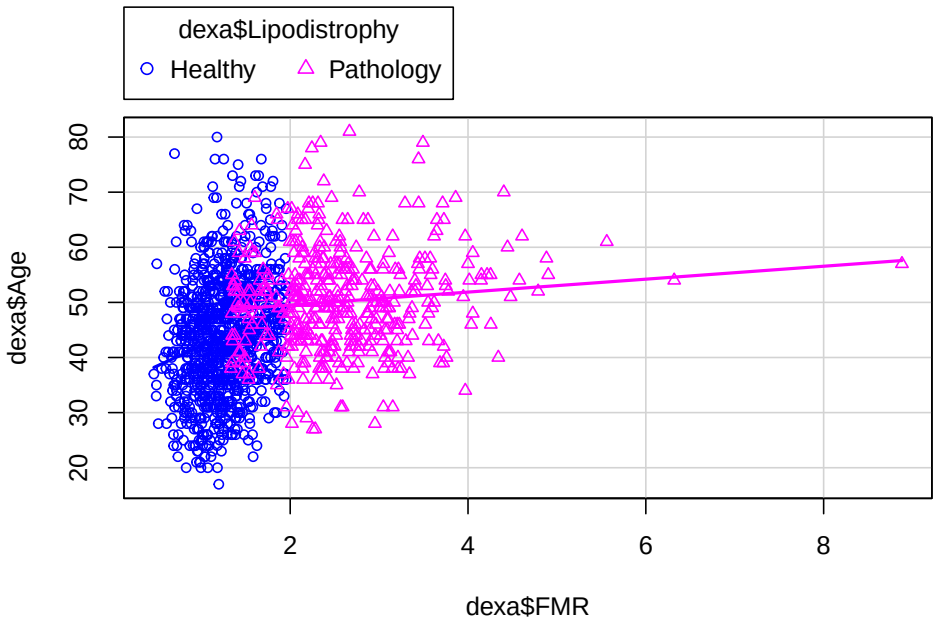
We can see that there are differences between the levels.

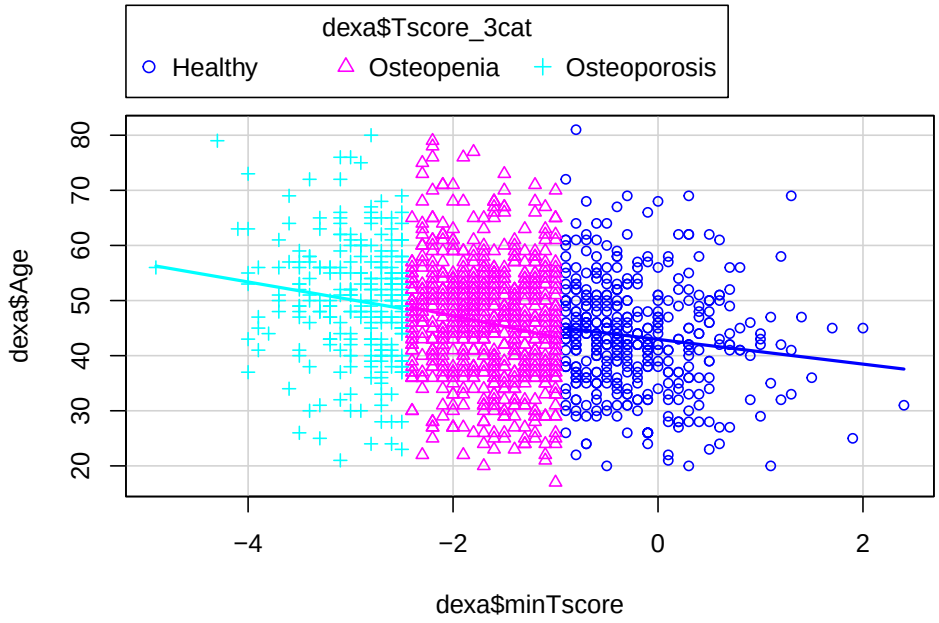
Age

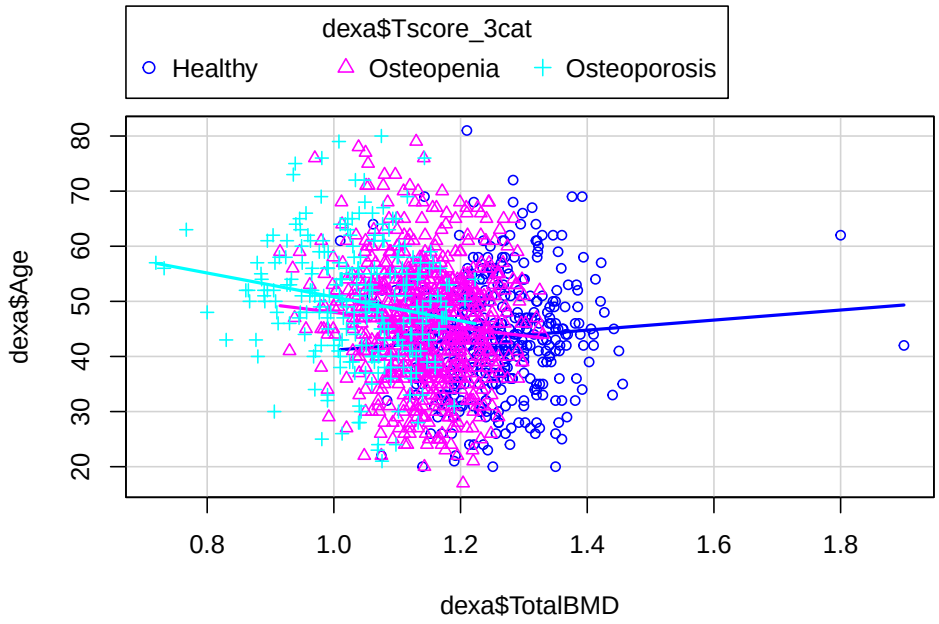
Another variable to consider is the age, because the minTscore it comes from the minimum Tscore value, and this Tcore it comes from the comparison with people of a certain age(30), so at an older age further from the reference value.

In all of the pathologies age is also important because we are talking about diseases that are age degenerative.









In the four graphics we can see that with more age there is more pathology.

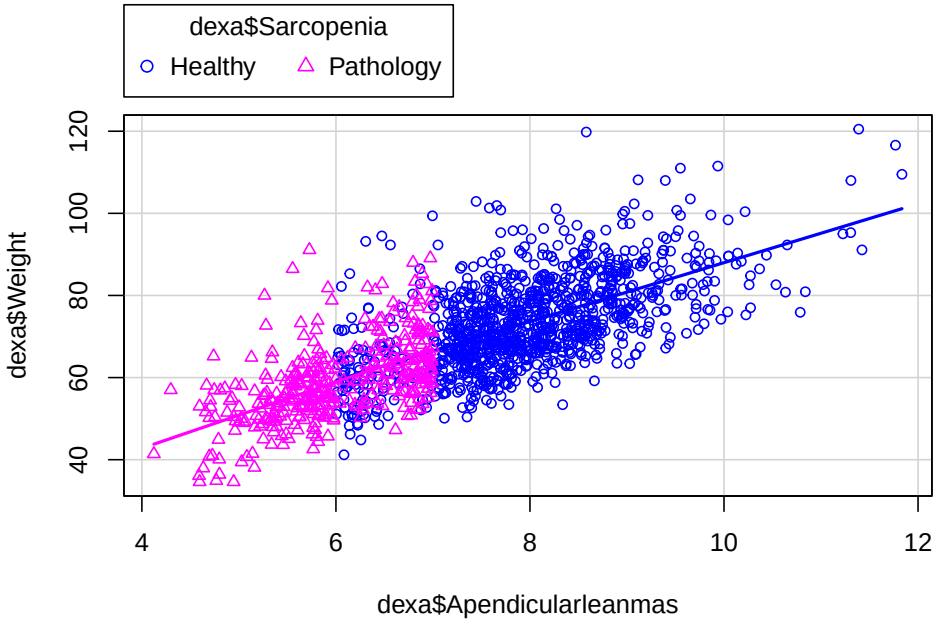
Other graphics of that variable can be seen in appendix 3.

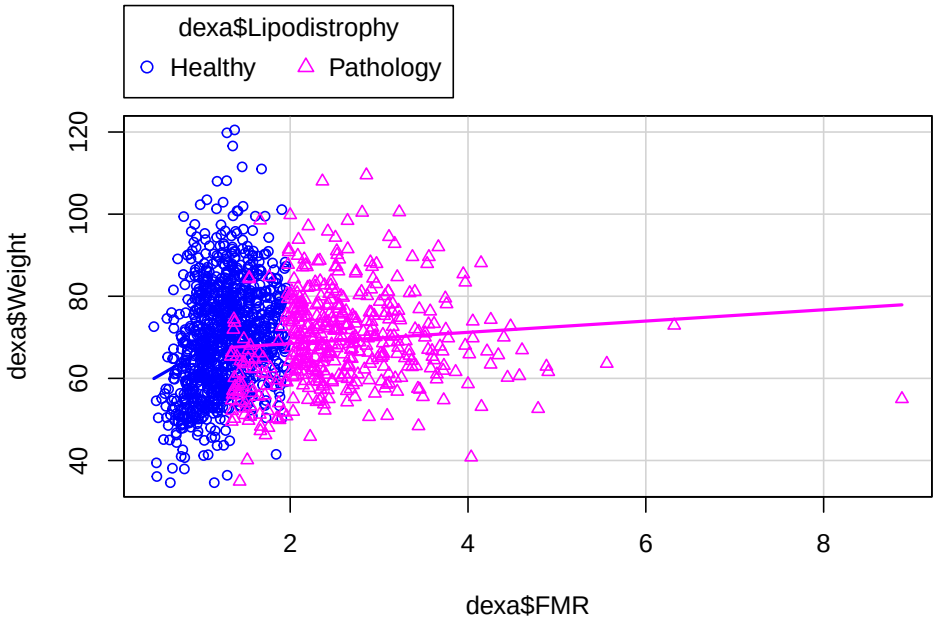
People with more age have less levels of appendiculerleanmas, minTscore and TotalBMD, and have more FMR.

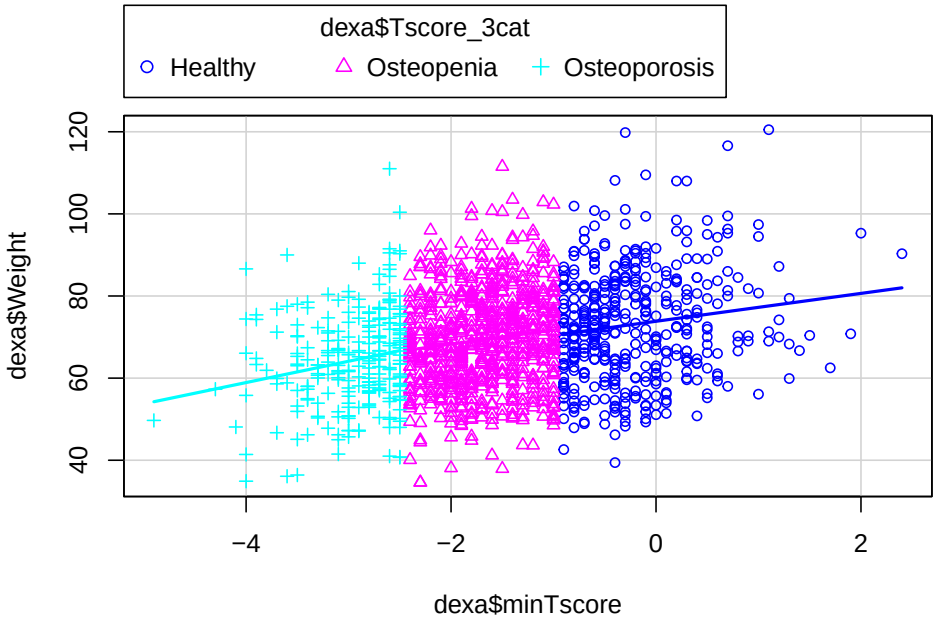
This indicates clearly that the age is a predisposing factor.

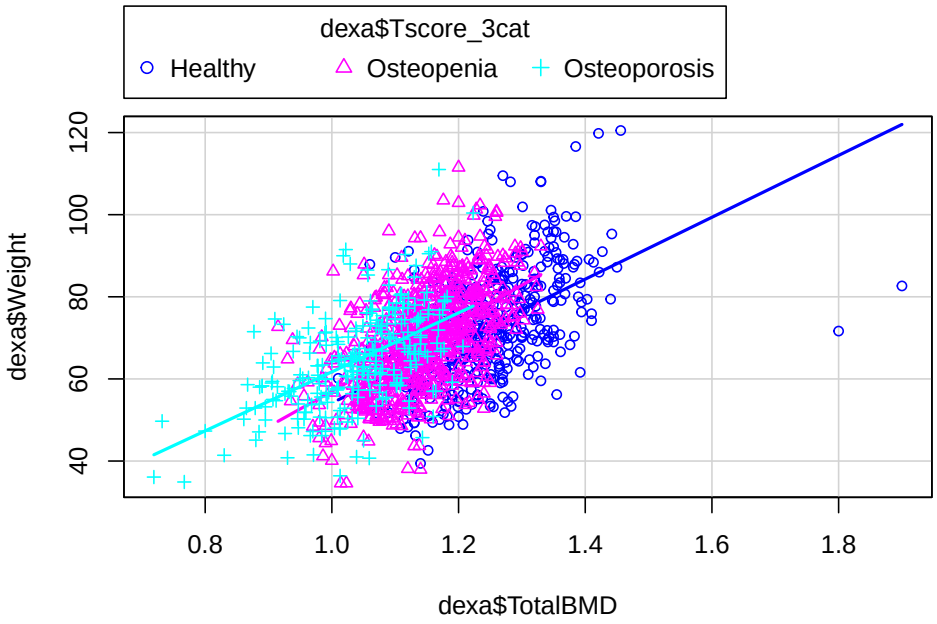
Weight

Another variable to consider is the weight as this diseases affect bone, fat and muscle.









In this graphics we can see that with less weigth there are more disease for bone disease and Sarcopenia, and that for Lipodistrophy with more weight there are more disease.

Other graphics can be seen in appendix 3.

Normality

We look for the normality of this six variables.In the three pathologies we look to the related numeric variable.

Normal qqplots can be seen in appendix 3.

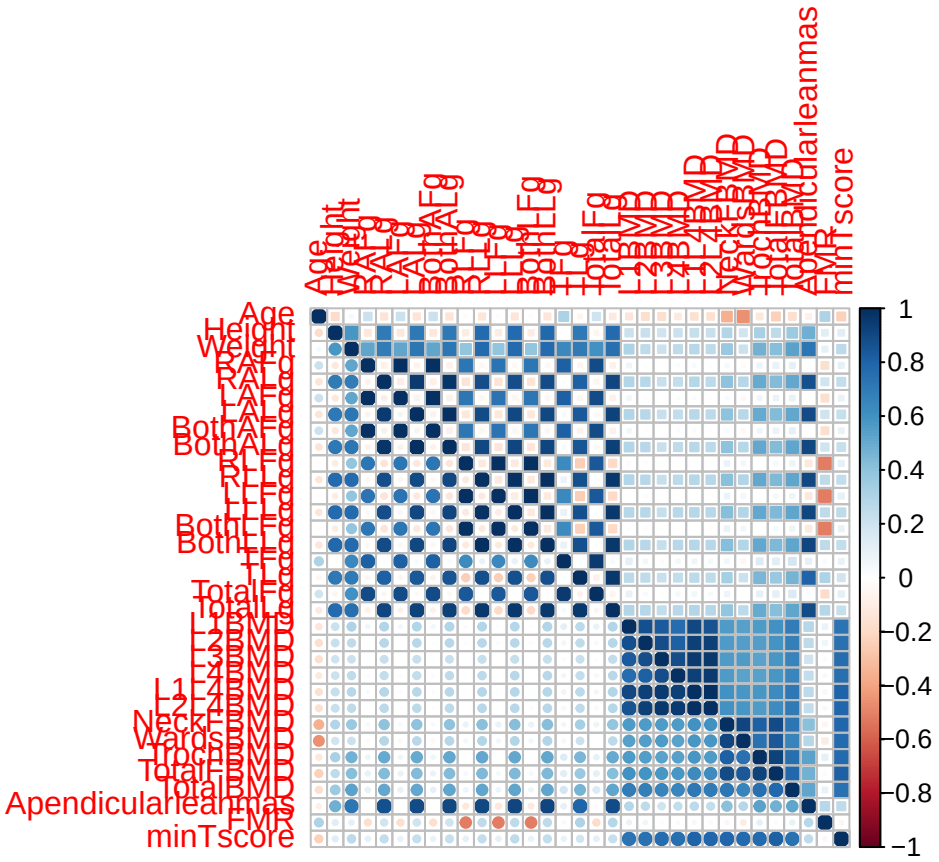
We check for a normality test.

Test	Variable	Statistic	p value	Normality
Anderson-Darling	Age	1.2107	0.0037	NO
Anderson-Darling	Height	3.7881	<0.001	NO
Anderson-Darling	TotalBMD	1.8414	1e-04	NO
Anderson-Darling	Apendicularleanmas	2.2522	<0.001	NO
Anderson-Darling	FMR	49.4211	<0.001	NO
Anderson-Darling	minTscore	1.7081	2e-04	NO

None of the variables follow a normal distribution.

3.3 CORRELATION

As there are a lot of high correlatives variables the ones selected for the correlation are the fat and lean variables with grams and the bone mineral density in each part of the body as we want to see the relation of that parts with the TotalBMD, Apl,FMR and minTscore, the outcomes.



We can see that the variables of the same diseases are highly correlated.

Lean variables are higher correlated with TotalBMD than the fat variables.Has also some correlation with minTscore.

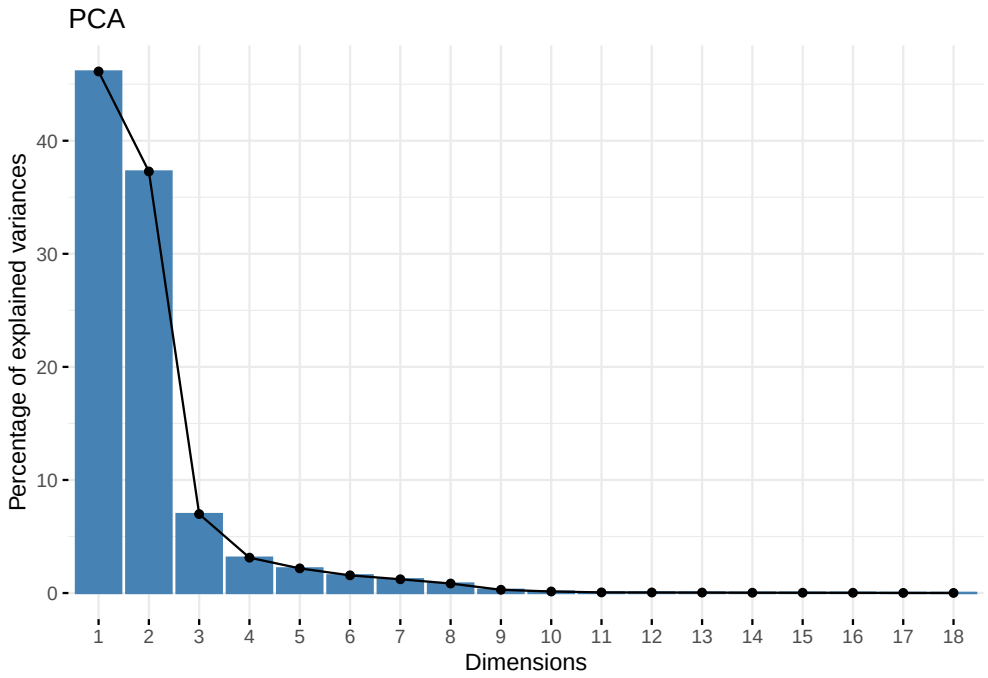
Age has a negative correlation with TotalBMD, Alm and minTscore. It has also a positive correlation with FMR, as the values of this disease are more higher more disease,unlike the other three.

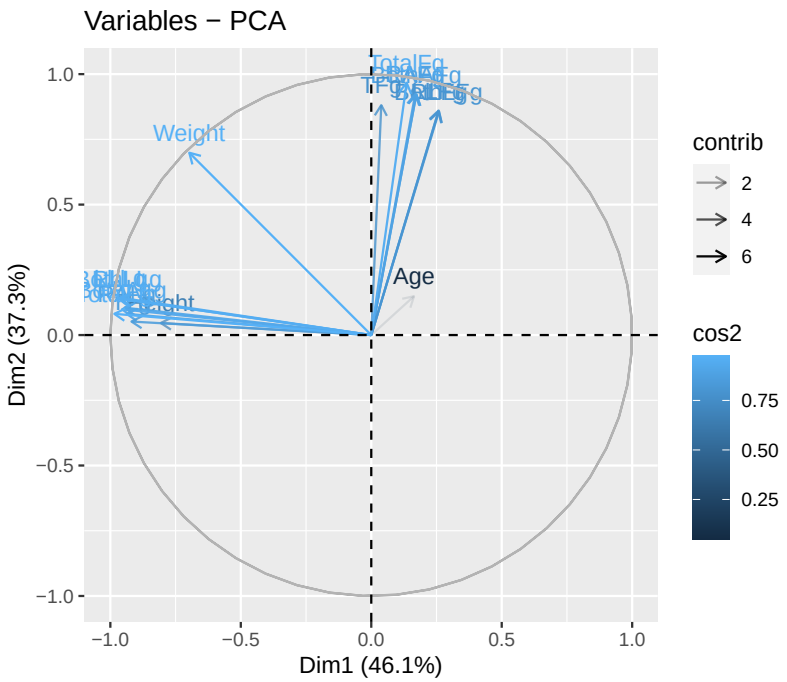
3.4 PCA

We want to predict Osteoporosis from the variable TotalBMD. This variable have a numeric value but not a categorical factor. So as we want to predict TotalBMD from fat and lean measures we will look the PCA only with fat and lean variables, and the ones with grs.

```
##             PC1    PC2     PC3     PC4     PC5
## Age      0.0553 0.0559 -0.7208 -0.6379 -0.0964
## Height  -0.2725 0.0173  0.1253 -0.1770  0.7303
## Weight  -0.2356 0.2625 -0.0615 -0.0115  0.0627
## RAFg     0.0569 0.3468 -0.1563  0.3161 -0.0171
## RALg    -0.3166 0.0305  0.0335 -0.0174 -0.3553
## LAFg     0.0584 0.3465 -0.1515  0.3195 -0.0237
## LALg    -0.3195 0.0395  0.0333 -0.0268 -0.3480
## BothAFg  0.0582 0.3466 -0.1563  0.3196 -0.0214
## BothALg -0.3227 0.0371  0.0336 -0.0216 -0.3407
## RLFg     0.0877 0.3220  0.3225 -0.2832 -0.0785
## RLLg    -0.3249 0.0520  0.0150 -0.0294  0.0824
## LLFg     0.0872 0.3234  0.3172 -0.2819 -0.0789
## LLLg    -0.3253 0.0538  0.0142 -0.0248  0.0768
## BothLFg  0.0869 0.3221  0.3219 -0.2849 -0.0836
## BothLLg -0.3273 0.0526  0.0141 -0.0271  0.0828
## TFg      0.0130 0.3311 -0.2384  0.0126  0.1958
## TLg     -0.3105 0.0194 -0.1245  0.0985  0.0514
## TotalFg  0.0458 0.3633 -0.0498 -0.0432  0.0934
## TotalLg -0.3326 0.0305 -0.0461  0.0354  0.0056
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.9604 2.6615 1.15195 0.77097 0.64371 0.54540 0.48113
## Proportion of Variance 0.4613 0.3728 0.06984 0.03128 0.02181 0.01566 0.01218
## Cumulative Proportion  0.4613 0.8341 0.90391 0.93520 0.95701 0.97266 0.98485
##                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.39987 0.23418 0.16086 0.10251 0.09677 0.08849 0.07354
## Proportion of Variance 0.00842 0.00289 0.00136 0.00055 0.00049 0.00041 0.00028
## Cumulative Proportion  0.99326 0.99615 0.99751 0.99806 0.99856 0.99897 0.99925
##                           PC15    PC16    PC17    PC18    PC19
## Standard deviation     0.06963 0.06537 0.04563 0.04194 0.03527
## Proportion of Variance 0.00026 0.00022 0.00011 0.00009 0.00007
## Cumulative Proportion  0.99951 0.99973 0.99984 0.99993 1.00000
```





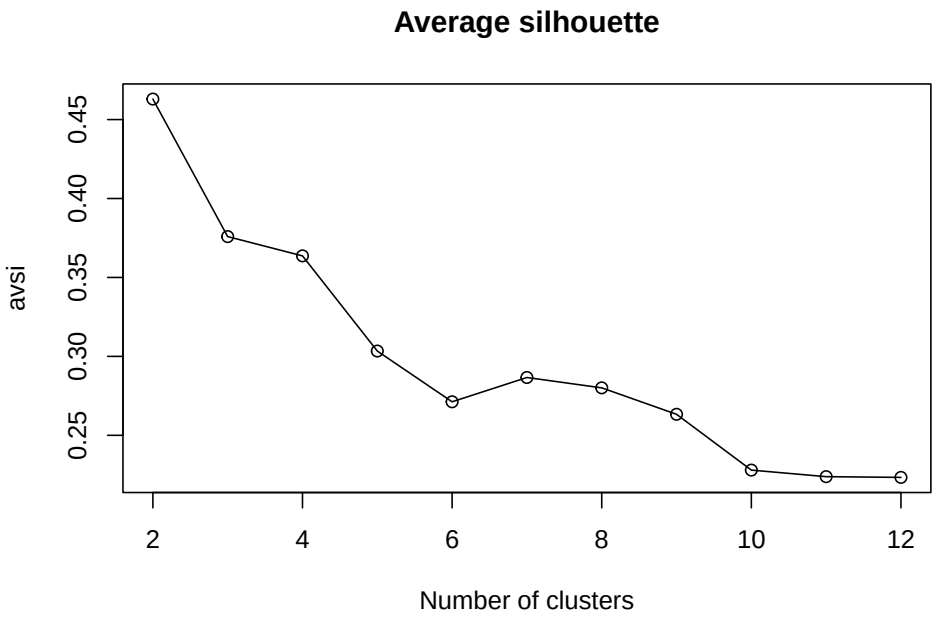
We can see that with the first two PCA the 83% of the variance is explained.

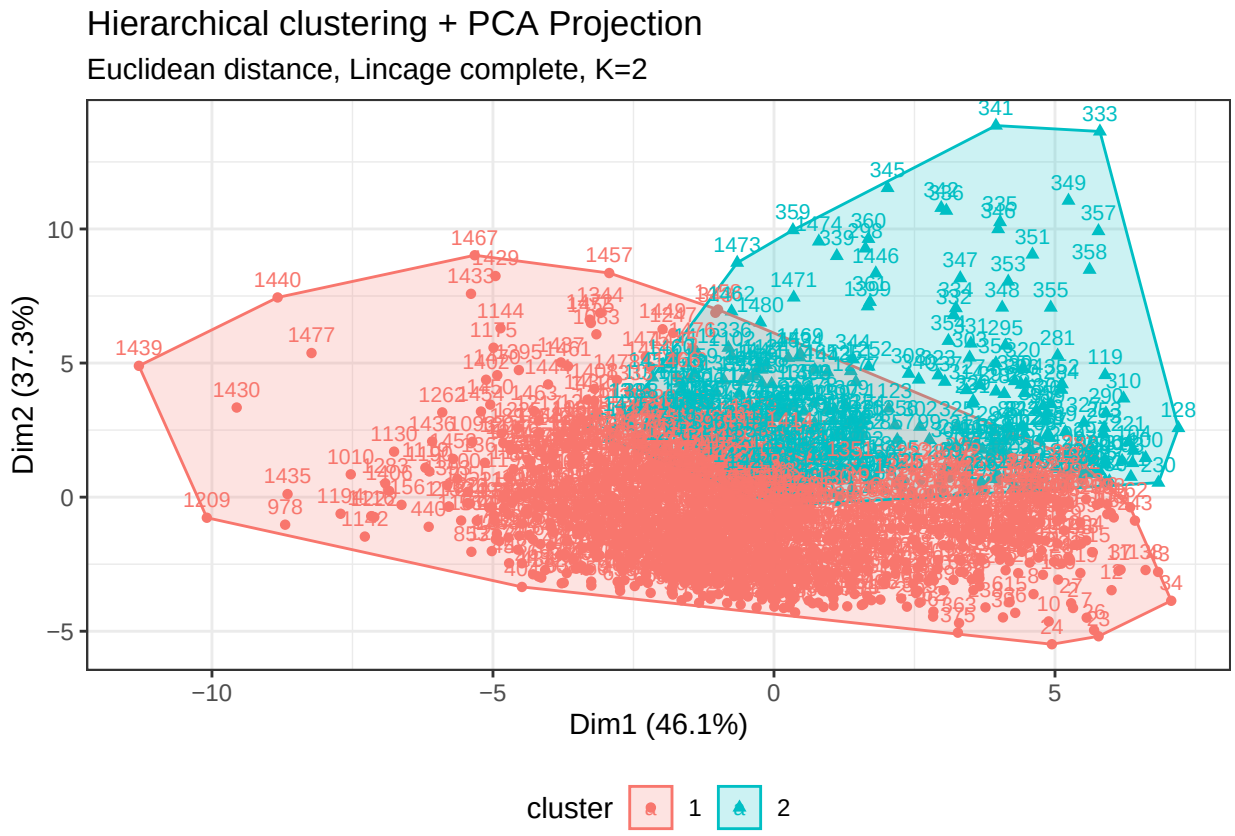
A scatterplot 3D can be seen in appendix 3.

3.5 CLUSTERING

We make with the same variables a hierarchical cluster.

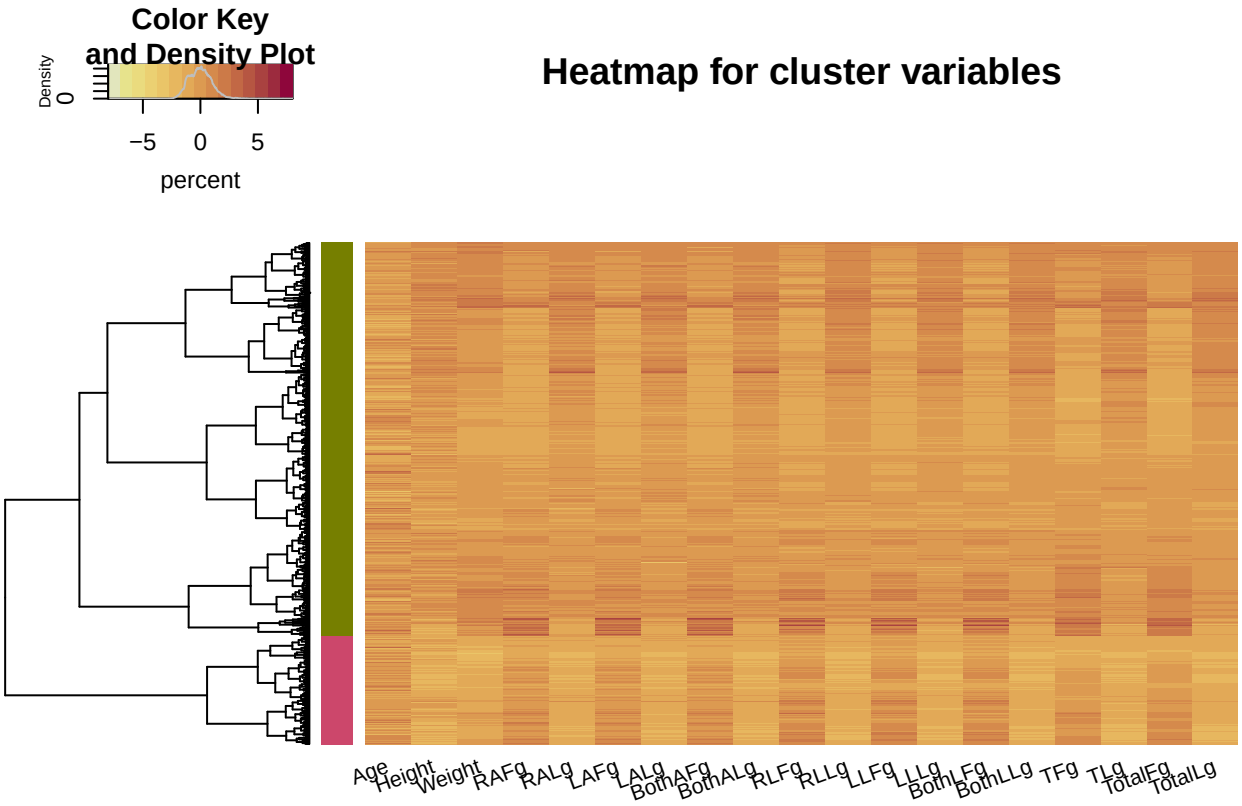
We look for the best number of clusters with the silhouette method.





A cluster dendogram amb a PAM cluster can be seen in Appendix 3.

We can looked in the heatmap the different intensity in each group.



We can see that some variables have a little differences in one group, RAFg,LAFg,BothAFg,RLFg,LLFg,BothLFg,TFg,TotalFg.

With that information we create a new variable, BMD_2cat, as we want to predict the BMD from fat/lean variables. It's made with the hierarchichal cluster.

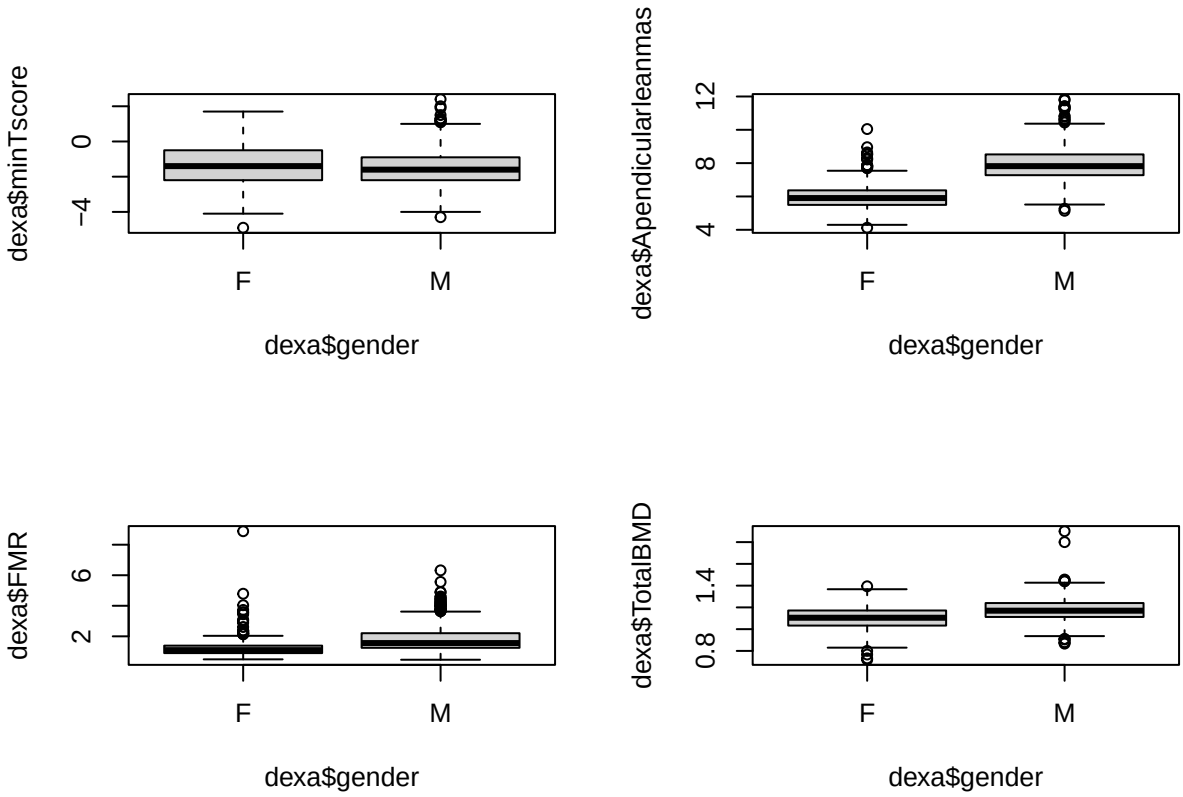
We can see the coincidence between Tscore_2cat and BMD_2cat

	Healthy	Pathology
Healthy	943	192
Pathology	256	59

3.6 FACTOR SEX

It's clear that the variable sex have some influence on the data.

We make a boxplot of the four answers divided by sex.

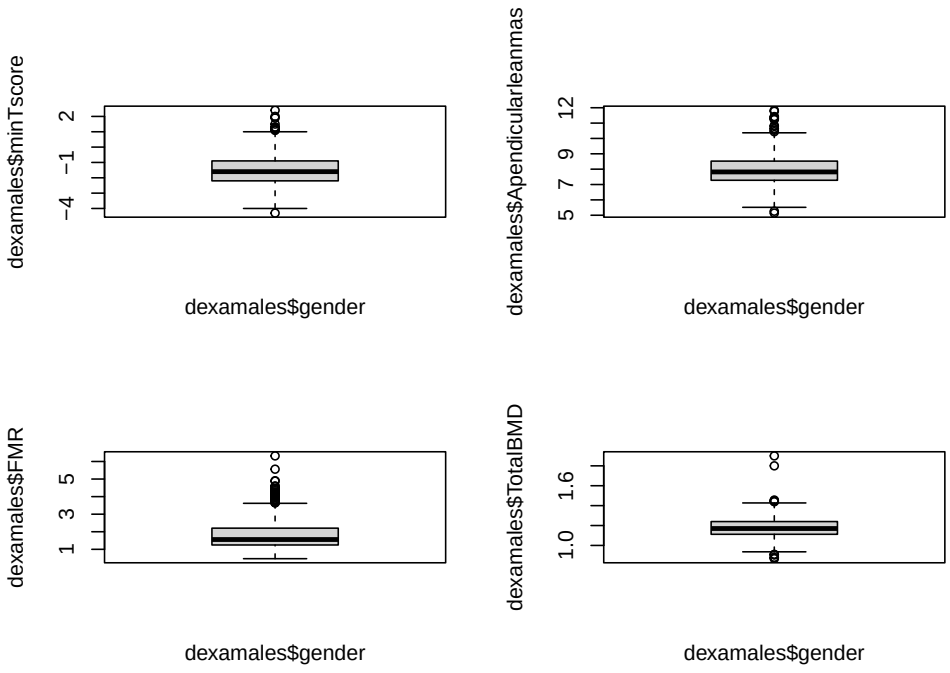


With that difference we will divide the database in two, by gender.

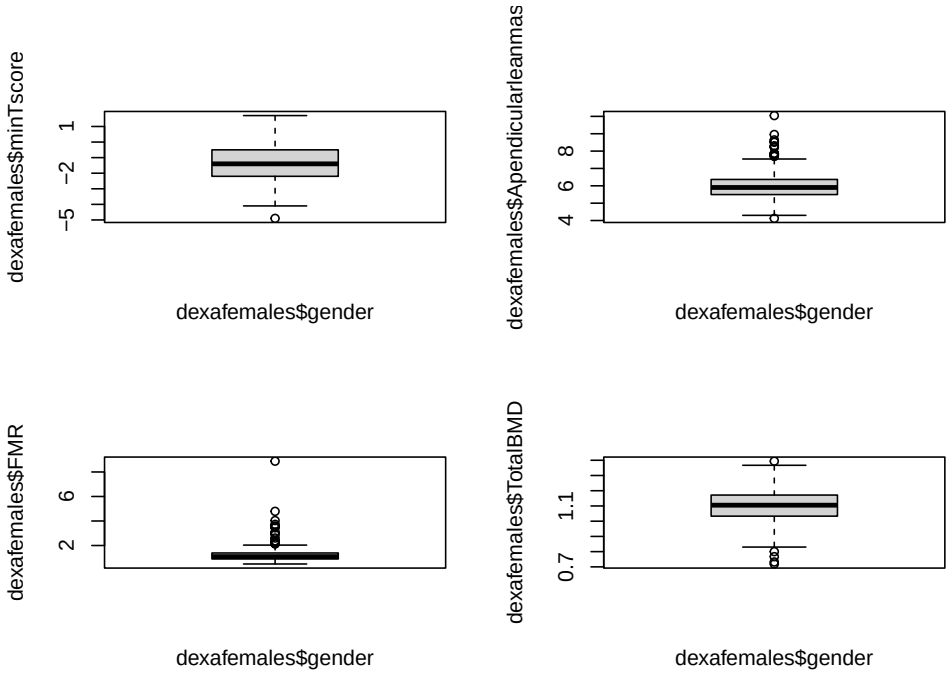
	rows	columns
Males	1098	84
Females	352	84

3.7 OUTLIERS

3.7.1 Males



3.7.2 Females



We put out the outliers.

Final databases have these dimensions.

	Rows	columns
Males	1030	84
Females	307	84

4 FINAL DATABASES

With that we create four databases,BMD,Osteoporosis,Sarcopenia and Lipodistrophy for the predictions. It's made for both sex.

The first variable is the outcome.

We have then eight databases:

4.1 CATEGORICAL DATABASES

BMDmales
Osteoporosismales
Sarcopeniamales
Lipodistrophymales
BMDfemales
Osteoporosisfemales
Sarcopeniafemales
Lipodistrophyfemales

All are saved.

We make the same with the numeric outcome..

4.2 NUMERICAL DATABASES

BMDmalesn
Osteoporosismalesn
Sarcopeniamalesn
Lipodistrophymalesn
BMDfemalesn
Osteoporosisfemalesn
Sarcopeniafemalesn
Lipodistrophyfemalesn

5 RESULTS

5.1 CATEGORICAL PREDICTIONS

With that created tables, we can pass the data in a R Markdown classificator with the main algorithms an obtain a result for the classification.

The chosen algorithms are “k-NN”,“NaiveBayes”,“Neuralnet”,“SVM”,“Decision Tree”,“RandomForest”,“k-NN Caret”,“NaiveBayes Caret”,“Neuralnet Caret”,“SVM Caret”,“Decision Tree Caret”,“RandomForest Caret”.

Some algorithms has been made with more than one option. k-NN have been made with k = 1,10,20 and the best one has been chosen. SVM has been made with lineal or rbf kernel and the best model has been chosen In caret model SVM has been made with lineal and radial model and best model has been chosen. In random forest there are two models with a different number of trees. The best model is chosen regarding the accuracy value.

The results are measured with accuracy, kappa and AUC.

The data is normalized.

We put two databases as example, Sarcopenia females and BMD males.

The models has been tested also making principal components analisis(PCA) with the first four components but the results are worse than the original data, so this results are not shown.

Males have 1030 samples and females 307. The samples are divided in train (67%) and test (33%).

Sarcopenia females has this distribution in the outcome

	No Sarcopenia	Sarcopenia
Number	128.00	179.00
%	0.42	0.58

The number of values are quite similar in both groups.

If we make the same for BMDmales.

	Healthy	Pathology
Number	857.00	173.00
%	0.83	0.17

We can see that the pathology class is much less represented than the healthy one. When this happens there are ways to oversampling and increase the samples of the underrepresented class. For that reason with TotalBMD males we make the algorithms with the original data and we make another model resampling the class pathology with 800 values and then the two classes are left with a similar percentage.

	Healthy	Pathology
Number	857.00	800.00
%	0.52	0.48

5.1.1 SARCOPENIA FEMALES

Algorithm	Accuracy	Kappa	AUC
k-NN	0.71	0.38	0.61
NaiveBayes	0.65	0.26	0.76
Neuralnet	0.77	0.53	0.84
SVM	0.86	0.71	0.34
Decision Tree	0.62	0.20	0.75
RandomForest	0.67	0.30	0.74
k-NN Caret	0.57	0.01	0.69
NaiveBayes Caret	0.67	0.31	0.72
Neuralnet Caret	0.67	0.31	0.91
SVM Caret	0.85	0.69	0.91
Decision Tree Caret	0.62	0.20	0.76
RandomForest Caret	0.69	0.36	0.79

There are only 307 samples.

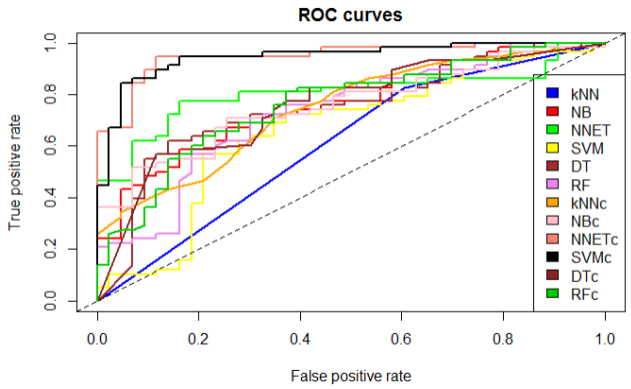
In this case SVM from caret is the best model to predict the disease.

We can see the confusion matrix and the results for this model.

```
Confusion Matrix and Statistics

          Reference
Prediction  Healthy Pathology
  Healthy        31         3
  Pathology      12        55

               Accuracy : 0.8515
                 95% CI : (0.7669, 0.9144)
    No Information Rate : 0.5743
    P-Value [Acc > NIR] : 2.023e-09

                  Kappa : 0.6878

 Mcnemar's Test P-Value : 0.03887

            Sensitivity : 0.7209
            Specificity : 0.9483
         Pos Pred Value : 0.9118
         Neg Pred Value : 0.8209
             Prevalence : 0.4257
         Detection Rate : 0.3069
   Detection Prevalence : 0.3366
      Balanced Accuracy : 0.8346

       'Positive' Class : Healthy
```

The kappa value, 0.69, can be improved but with the few data that there is, it is in the range of good agreement.

It has also and AUC of 0.9.If we see graphically all the AUC for all the models



We can see that best AUC is SVM from caret with black color.

With a few data the predictions results are quite good.

5.1.2 TOTAL BMD MALES

Algorithm	Accuracy	Kappa	AUC
k-NN	0.94	0.81	0.90
NaiveBayes	0.91	0.73	0.96
Neuralnet	0.91	0.72	0.96
SVM	0.92	0.71	0.97
Decision Tree	0.91	0.70	0.96
RandomForest	0.92	0.72	0.97
k-NN Caret	0.90	0.59	0.97
NaiveBayes Caret	0.87	0.66	0.96
Neuralnet Caret	0.91	0.69	0.97
SVM Caret	0.91	0.69	0.97
Decision Tree Caret	0.92	0.73	0.97
RandomForest Caret	0.92	0.72	0.97

5.1.3 RESAMPLED TOTAL BMD MALES

Algorithm	Accuracy	Kappa	AUC
k-NN	0.99	0.97	0.99
NaiveBayes	0.91	0.83	0.96
Neuralnet	0.92	0.84	0.96
SVM	0.95	0.90	1.00
Decision Tree	0.97	0.94	0.99
RandomForest	0.97	0.94	1.00
k-NN Caret	0.92	0.84	0.98
NaiveBayes Caret	0.90	0.80	0.97
Neuralnet Caret	0.91	0.81	0.97
SVM Caret	0.97	0.95	1.00
Decision Tree Caret	0.98	0.96	1.00
RandomForest Caret	0.97	0.94	1.00

As this factor, healthy or payhology has been created making a cluster from the data, the results are very good.

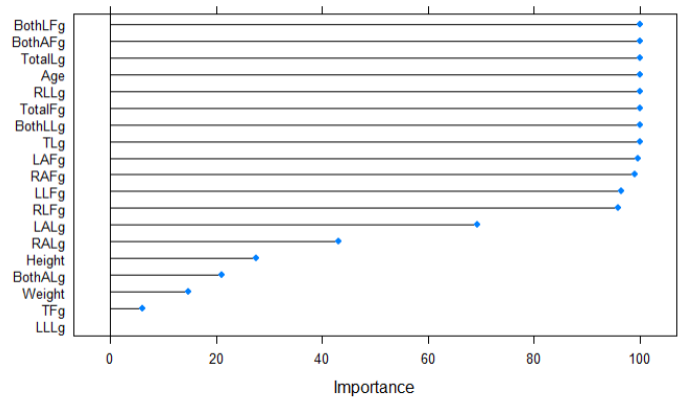
The results are quite similar, the best are obtained with k-NN and decision tree in caret.

It results good how the results improves when a resampling has been made because the kappa statistic improves a lot in this case. Accuracy and AUC are good in both models because they are measures that measure the results of the positive class and, as in the first case, the positive class is highly represented, it gives very high values in these two measures.

We can see the good results of the k-NN model in the resampled model.

```
Confusion Matrix and Statistics

          Reference
Prediction  Healthy Pathology
  Healthy       284         0
  Pathology       7       256

               Accuracy : 0.9872
                 95% CI : (0.9738, 0.9948)
    No Information Rate : 0.532
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.9743

 Mcnemar's Test P-Value : 0.02334

            Sensitivity : 0.9759
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.9734
             Prevalence : 0.5320
         Detection Rate : 0.5192
   Detection Prevalence : 0.5192
      Balanced Accuracy : 0.9880

       'Positive' Class : Healthy
```

If we look at the second best model, decision tree in caret, and their importance variables.



Age is in the firsts positions with the maximum score, also both legs and both arms fat.

With the resampling data, where each group has a similar representation the results are very good.

5.2 NUMERICAL PREDICTIONS

In the numerical predictions we choose the other two pathologies that are Lipodistrophy and Osteoporosis. We make one with males and the other with females. Males have 1030 samples and females 307. The samples are divided in train (67%) and test (33%).

There are the following algorithms "Linear Model","Regression Trees","SVM","Random Forest","Boosting","Caret Ridge","Caret Lasso","Caret RF","BayesGLM","Ensemble","PLS","NNet","Ctree".

This algoritms has been chosen because the two main ways to make numerical predictions come from linear regression and decision trees.

The algorithms are made with the normal data except the last three where the variables of the data has been selected keeping only the ones with a correlation under 0.8. So, instead of having 20 predictor variables we only keep 8.

The metric for this predictions are the MAE and RMSE.

5.2.1 LIPODISTROPHY MALES

Algorithm	MAE	RMSE
Linear Model	0.435	0.573
Regression Trees	0.497	0.649
SVM	0.426	0.580
Random Forest	0.505	0.634
Boosting	0.465	0.601
Caret Ridge	0.435	0.571
Caret Lasso	0.435	0.570
Caret RF	0.471	0.601
BayesGLM	0.434	0.572
Ensemble	0.466	0.602
PLS	0.440	0.575
NNet	0.439	0.577
Ctree	0.481	0.612

The SVM and regression models have the best predictions, linear model,ridge and lasso regression,Bayes GLM. Also NNET with only not high correlated variables has similar results.

If we look at the linear regression model

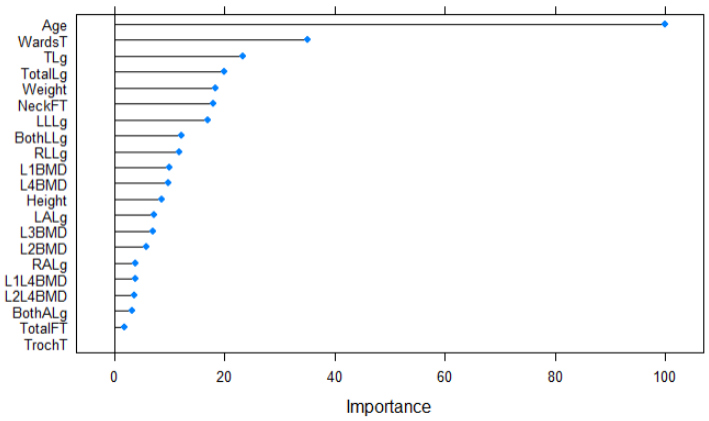
```
Call:
lm(formula = y_train ~ ., data = X_train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.48237 -0.37241 -0.07751  0.32979  1.72802

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.905e+00  6.728e-01   2.832 0.004767 **
Age          1.680e-02  2.801e-03   5.997 3.29e-09 ***
Height      -1.589e+00  4.392e-01  -3.618 0.000319 ***
Weight      -2.291e-02  3.344e-03  -6.851 1.66e-11 ***
RALg        -8.437e-05  8.616e-05  -0.979 0.327803
LALg        -8.686e-05  1.718e-04  -0.505 0.613382
BothALg     -1.321e-04  1.146e-04  -1.153 0.249169
RLLg        -2.631e-05  9.197e-05  -0.286 0.774871
LLLg         2.222e-04  1.510e-04   1.472 0.141507
BothLLg     -1.445e-04  1.174e-04  -1.231 0.218893
TLg         -1.003e-04  3.749e-05  -2.674 0.007677 **
TotalLg      1.430e-04  3.614e-05   3.956 8.45e-05 ***
L1BMD        6.196e-01  6.219e-01   0.996 0.319434
L2BMD        4.163e-01  4.991e-01   0.834 0.404547
L3BMD       -1.098e+00  8.741e-01  -1.256 0.209460
L4BMD        5.442e-03  8.269e-01   0.007 0.994751
L1L4BMD      9.907e-01  2.756e+00   0.360 0.719324
L2L4BMD     -4.474e-01  2.779e+00  -0.161 0.872129
NeckFT       5.148e-03  6.943e-02   0.074 0.940924
WardsT      -1.361e-01  5.489e-02  -2.479 0.013413 *
TrochT       4.406e-02  5.357e-02   0.823 0.411083
TotalFT      3.474e-02  8.482e-02   0.410 0.682246
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5776 on 668 degrees of freedom
Multiple R-squared:  0.2548,    Adjusted R-squared:  0.2314
F-statistic: 10.88 on 21 and 668 DF,  p-value: < 2.2e-16
```

We can see that age, hight and weight are significant.There is a positive relation between FMR and age, for every 0.016 years more there is one point more of FMR.

There is a negative relation between weight and height with FMR.

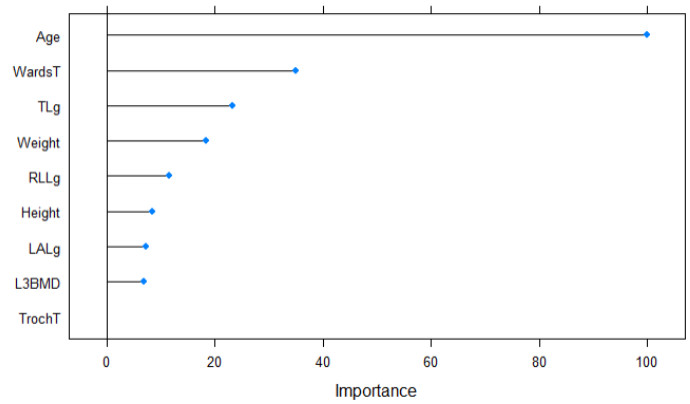
TLg,TotalLg and WardsT are also significant.

If we see Bayes GLM importance



We can see that the first five variables are the significance in linear regression.

If wee see the importance in neuralnet model with no high correlated variables



We can see the variables are mostly the same. In this case RLLg is in the five position.

The cutoff of FMR in men is 1.961. The best prediction has a MAE of 0.426 so the predictions in this case are not very good.

5.2.2 OSTEOPOROSIS FEMALES

Algorithm	MAE	RMSE
Linear Model	0.756	1.363
Regression Trees	0.735	0.932
SVM	0.637	0.837
Random Forest	0.699	0.886
Boosting	0.645	0.846

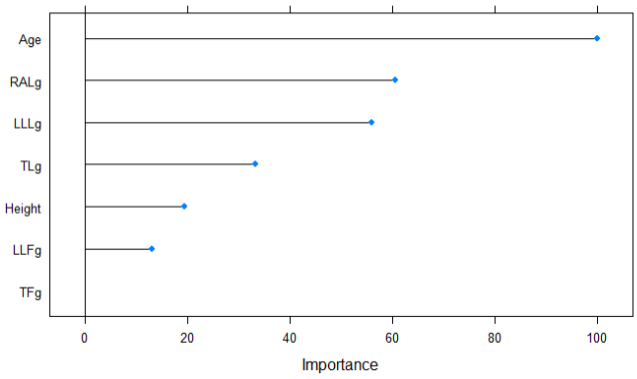
Algorithm	MAE	RMSE
Caret Ridge	0.637	0.837
Caret Lasso	0.637	0.838
Caret RF	0.656	0.866
BayesGLM	0.759	1.366
Ensemble	0.683	0.898
PLS	0.633	0.829
NNet	0.601	0.790
Ctree	0.693	0.876

The best prediction results are neuralnet of caret, followed by PLS and robust regressions. Models based on trees, random forest, ensemble have worst results.

Neuralnet and PLS models are made with few variables because high correlated variables are out, if we look which variables are more important in the model:

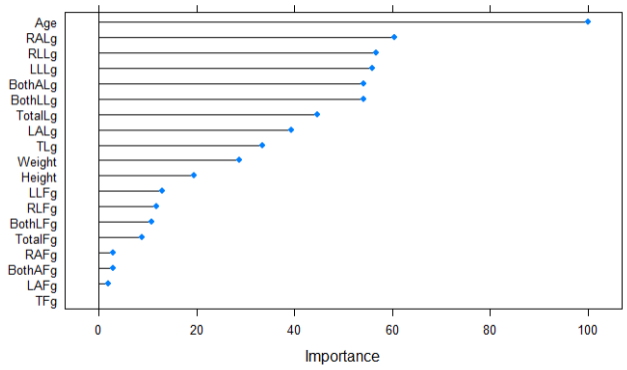
We look only Neuralnet because in PLS importance variables follow the same order. PLS importance can be seen in appendix 3.

Neuralnet



If we looked at the best model with all the variables included, that is model Ridge regression with caret and we look at the importance variables.

Ridge regression



How the data is in the model, the best possible score for a variable is 100, and we can see that the age has 100 in both models. The importance order of variables is very similar.

With the number of data that there is, these are the best predictions that are obtained.

As the prediction of categorical Total BMD in males, has quite good results we try to make a numerical prediction with that outcome.

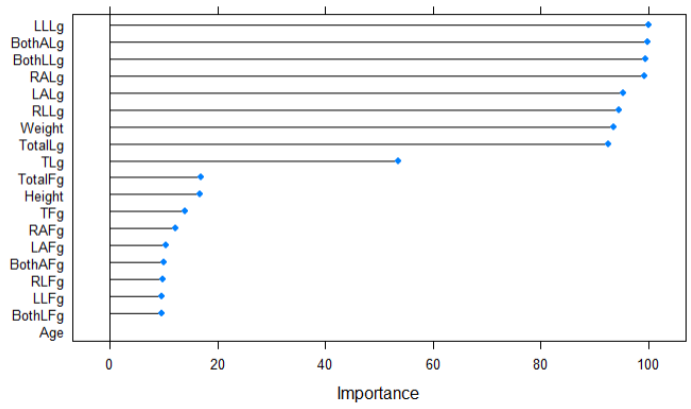
5.2.3 NUMERICAL PREDICTION TOTAL BMD MALES

Algorithm	MAE	RMSE
Linear Model	0.058	0.073
Regression Trees	0.066	0.083
SVM	0.060	0.076
Random Forest	0.067	0.084
Boosting	0.066	0.083
Caret Ridge	0.060	0.074
Caret Lasso	0.060	0.075
Caret RF	0.064	0.079
BayesGLM	0.058	0.073
Ensemble	0.065	0.081
PLS	0.061	0.076
NNet	0.060	0.075
Ctree	0.065	0.082

The results are good. Best models are linear regression and Bayes GLM.

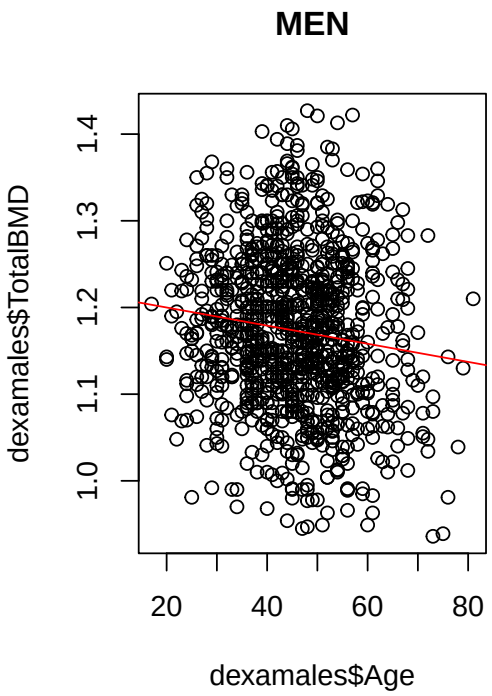
Differences between the predicted values and real values are similar in all models.We can see a density plot of the differences in appendix 3.

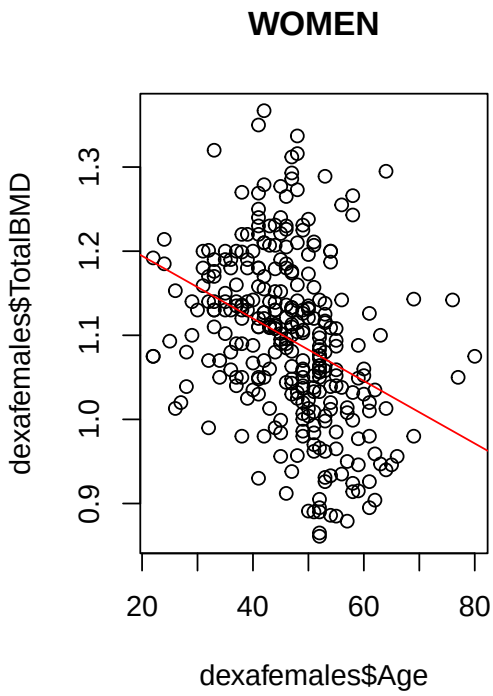
If we look at the importance of Bayes GLM.



The first variable is left leg lean. The first group are the variables related to the lean, the second group the fat. Age is in the last position.

The relation between age and Total BMD remains fairly constant and this helps in the good prediction of the results. In women this does not happen.


MEN
dexamales$TotalBMD
1.0
1.1
1.2
1.3
1.4
20
40
60
80
dexamales$Age


WOMEN
dexafemales$TotalBMD
0.9
1.0
1.1
1.2
1.3
20
40
60
80
dexafemales$Age

6 CONCLUSIONS

The categorical predictions has only good results when a resampling of the class pathology has been made.

In Sarcopenia females the number of samples is not very high and that makes not a good results, specially the kappa values.

In Total BMD the values of kappa are only in good agreement but improve with the sample resampled. The Accuracy and AUC are better also with the resampled sample.

It's clear that when one of the class has few values the results are not so good.

The most important variables in that predictions are Both legs,Both arms,TotalLg and age. The first individual body part in importance is RLLg.

In numerical predictions MAE and RMSE are high for Lipodistrophy males and Osteoporosis females. In both predictions age is the most important variable.In osteoporosis females RALg and LLLg are the following importance variables.

In the prediction of Total BMD males the results are better.In this case age is the last important variable and the first is LLLg. This probably happens because Total BMD is not a value compared to a specific age, such as the minTscore value, and this allows better predictions.

Total BMD is a general value, and we has seen that with more age there is less Total BMD but in men the descent is lower than in women.

The importance variables order is saying that lean is more important than fat for predicting this variable.

In all predictions except the last one age is a very important variable. It's normal as this pathologies are sex and age dependents. In this case factor sex has been blocked.

As the numerical predictions of Total BMD are good, if there were some reference values regarding age and sex of this variable, it could be helpful information. What is clear is that lower Total BMD values increase the risk of osteoporosis and bone fracture. If Total BMD (all body) could be a good predictor in Osteoporosis is out of this work.

With that results we can say that the numerical prediction of one pathology with the variables of the other two diseases has not good results.

My growth in learning ML has been great. I have delved into the subject and I have known books such as Applied Predictive Modeling, Max Khun[15], which has helped me see how the subject of predictions works and develop solutions to problems with data.

I have also learned to plan a job, set dates, milestones and follow the order set, which in this case has been successful.

One of the objectives of the work was to create a report in Rmarkdown with the choosen algorithms in a dynamic report so that it could be used in other databases. The objective has been met and have been created two dynamic reports, one for categorical predictions "MLCAT" and the other for numerical predictions "MLNUM". Only changing the database the dinamic R Markdown makes the solutions.

6.1 FUTURE WORKS

In the predictions of Osteoporosis and Total BMD, RALg and LLLg seems to have importance in the prediction. Other works can go in the direction to delve into the relationship of these variables with the outcomes.

Exploring the relationship between total BMD and Osteoporosis may also be another work.[10]

In this work we wanted to learn how the most important machine learning algorithms behaved. Algorithms have been mixed with others from the caret package. The work can be done only with the caret package in all its extension and results. It was not one of the objectives of this work but seeing the possibilities offered by this package has made me think that it is a very good option.

It could also be tried to make the predictions with a transformation of some of the main variables, for example with the Box-Cox method.The Cox model has the advantage of preserving the variable in its original quantitative form, and of using a maximum of information. However, very restrictive conditions of application of this model make its use rather limited. Some works goes on that direction.[21]

7 GLOSSARY

AIDS: Acquired immunodeficiency syndrome

HIV: Human immunodeficiency virus

BMD: Bone mineral density

TotalBMD: Total Body bone mineral density.

DEXA: dual-energy x-ray absorptiometry

MAE: Mean absolute error

RMSE: Root mean squared error

ROC curve: Receiver Operator Characteristic curve

AUC: area under the corve.

PCA: Principal components analisis

Alm: Apendicular lean mas

FMR: Fat mass ratio

SVM: suported vector machine.

PLS: Partial least squares

RF: Random Forest

Bayes GLM: Bayes generalised linear model

NNet: neuralnet

Ctree: Decision tree

8 BIBLIOGRAPHY

[1]Abdul Aziz, Siti Azdiah, et al. "Assessment of sarcopenia in virally suppressed HIV-infected Asians receiving treatment." Aids 32.8 (2018): 1025-1034.

[2]Bonjoch, Annaa; Figueras, Martab; Estany, Carlaa; Perez-Alvarez, Núriaa,b; Rosales, Joaquimc; del Rio, Luísc; di Gregorio, Silvanac; Puig, Jordia; Gómez, Guadalupea,b; Clotet, Bonaventuraa,d; Negredo, Eugèniaa the Osteoporosis Study Group High prevalence of and progression to low bone mineral density in HIV-infected patients: a longitudinal cohort study, AIDS: November 27, 2010 - Volume 24 - Issue 18 - p 2827-2833 doi:10.1097/QAD.0b013e328340a28d

[3]Brown, Todd T., and Roula B. Qaqish. "Antiretroviral therapy and the prevalence of osteopenia and osteoporosis: a meta-analytic review." Aids 20.17 (2006): 2165-2174.

[4]Compston, Juliet. "HIV infection and osteoporosis." BoneKEy reports 4 (2015).

[5]Dos Santos, A. P., Navarro, A. M., Schwingel, A., Alves, T. C., Abdalla, P. P., Venturini, A. C. R., … & Machado, D. R. (2018). Lipodystrophy diagnosis in people living with HIV/AIDS: prediction and validation of sex-specific anthropometric models. BMC public health, 18(1), 1-14.

[6]Everitt, Brian, et al. An R and S-PLUS companion to multivariate analysis. No. 519.5 E8.. London: Springer, 2005.

[7]Echeverría, Patricia, et al. "High prevalence of sarcopenia in HIV-infected individuals." BioMed research international 2018 (2018).

[8]Fernández Olivares, G. A. Análisis exploratorio multivariante para el estudio longitudinal de la distribución anormal de grasa, masa magra y/o masa ósea en individuos infectados por VIH..

[9]Géron, Aurélien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media, 2019.

[10]Gnudi, Saverio, Emanuela Sitta, and Nicoletta Fiumi. "Relationship between body composition and bone mineral density in women with and without osteoporosis: relative contribution of lean and fat mass." Journal of bone and mineral metabolism 25.5 (2007): 326-332.

[11]Guaraldi, Giovanni, et al. "The natural history of HIV-associated lipodystrophy in the changing scenario of HIV infection." HIV medicine 15.10 (2014): 587-594.

[12]Lantz, Brett. "Machine Learning with R. Packt Publishing." Birmingham Mumbai (2015).

[13]Lantz, Brett. Machine learning with R: expert techniques for predictive modeling. Packt publishing ltd, 2019.

[14]Lorente-Ramos, Rosa, et al. "Dual-energy x-ray absorptiometry in the diagnosis of osteoporosis: a practical guide." American Journal of Roentgenology 196.4 (2011): 897-904.

[15]Kuhn, Max, and Kjell Johnson. Applied predictive modeling. Vol. 26. New York: Springer, 2013.

[16]OLIVEIRA, Danielle Lima de. Avaliação das anormalidades do metabolismo mineral ósseo em pacientes portadores de lipodistrofia e HIV positivos. 2012. 112 f. Dissertação (Mestrado) – Universidade Federal do Pará, Núcleo de Medicina Tropical, Belém, 2012. Programa de Pós-Graduação em Doenças Tropicais.

[17]Pelegrín Cuartero, C. (2019). Exploring dimensionality reduction and machine learning methods for the prediction of body composition abnormalities among an HIV+ population.

[18]Regué Alsina, A. (2021). Exploratory analysis of a biological database (DEXA) and application of Machine Learning models to detect osteoporosis in HIV-positive patients

[19]Santos, A. P., Machado, D. R. L., Schwingel, A., Chodzko-Zajko, W. J., Alves, T. C., Abdalla, P. P., … & Navarro, A. M. (2019). Anthropometric cutoff points to identify lipodystrophy characteristics in people living with HIV/AIDS: an observational study. Nutrición hospitalaria: Organo oficial de la Sociedad española de nutrición parenteral y enteral, 36(6), 1315-1323.

[20]Tien, Phyllis C., and Carl Grunfeld. “What is HIV-associated lipodystrophy? Defining fat distribution changes in HIV infection.” Current opinion in infectious diseases 17.1 (2004): 27-32.

[21]Zhang, Tonglin, and Baijian Yang. “Box–cox transformation in big data.” Technometrics 59.2 (2017): 189-201.

[22]https://www.unaids.org/es/resources/fact-sheet (Visited September 2021)

[23]https://www.who.int/es/news-room/fact-sheets/detail/hiv-aids

[24]http://topepo.github.io/caret/index.html

9 APPENDIX

9.1 APPENDIX 1 (EXCLUDED DATA)

Main variables of the excluded individuals. The variables have a normal distribution.

ID	Gender	Age	Weight	Height	ALM	FMR	minTscore
45806	F	40	57.246	NA	NA	1.33	-1.8
289085	F	45	45.878	1.6	5.44	1.7	-2.3
10451281	F	45	44.6	1.57	5.35	1.21	-2.1
447739	F	57	41	1.53	4.85	0.48	-3
145445	F	48	52.216	1.55	5.34	0.81	-2.5
35981	F	44	61.12	1.66	5.86	1.16	-0.6
10006435	F	49	46.271	1.55	5.38	1.26	-1.8
93264	M	29	71.774	NA	NA	2.59	-1.8
503754	M	36	62.739	1.73	7.63	1.15	1.5
526215	M	44	63.821	1.76	7.69	1.71	-1.9
537097	M	31	75.084	1.84	8.47	1.41	-0.5
273533	M	40	62.8	1.7	8.1	2.13	-0.4
500488	M	47	62	1.715	7.33	1.12	-0.8
562458	M	49	62.994	1.73	7.36	0.94	-1.9
143630	M	61	52.3	1.64	6.05	1.22	-3.8
14511840	M	46	82.1	1.83	7.39	1.28	-0.8
193544	M	39	73.86	1.67	7.59	2.44	-2.1
341762	M	46	61.1	1.45	9.17	2.19	-1.3
474320	M	42	82.7	1.69	9.48	1.43	0.3
90908	M	46	79.6	1.76	8.06	1.12	-1.2
10000634	M	37	88	1.83	7.53	0.78	-0.8
261221	M	52	92.4	1.8	8.8	1.78	-1.2

Test	Variable	Statistic	p value	Normality
Anderson-Darling	exc.Age	0.3497	0.4374	YES
Anderson-Darling	exc.Weight	0.4122	0.3082	YES
Anderson-Darling	exc.Height	0.2580	0.6806	YES
Anderson-Darling	excAlm	0.6361	0.0832	YES
Anderson-Darling	excfmr	0.4734	0.2161	YES
Anderson-Darling	exc.minTscore	0.2495	0.7107	YES

9.2 APPENDIX 2 (DEXA DATABASE VARIABLES)

General information and anthropometric variables

ID	Patient Number
gender	Sex
gender_num	Sex numeric
Age	Age in years
Age_cat	Age categorical
Height	Height
Weight	Weight

Fat/lean variables

RAFp	Right Arm Fat%
RAFg	Right Arm Fat grs
RALg	Right Arm Lean grs
LAFp	Left Arm Fat%
LAFg	Left Arm Fat grs
LALg	Left Arm lean grs%
BothAFp	Both Arms Fat %
BothAFg	Both Arms Fat grs
BothALg	Both Arms Lean grs
RLFp	Right Leg Fat%
RLFg	Right Leg Fat grs
RLLg	Right Leg Lean grs
LLFp	Left Leg Fat%
LLFg	Left Leg Fat grs
LLLg	Left Leg Lean grs
BothLFp	Both Leg Fat%
BothLFg	Both Leg Fat grs
BothLLg	Both Leg Lean grs
TFp	Trunk Fat%
TFg	Trunk Fat grs
TLg	Trunk Lean grs
TotalFp	Total Fat%
TotalFg	Total Fat grs
TotalLg	Total Lean grs

Bone measures with T and Z values

L1BMD	Lumbar 1 BMD
L1T	Lumbar 1 T value
L1Z	Lumbar 1 Z value
L2BMD	Lumbar 2 BMD
L2T	Lumbar 2 T value
L2Z	Lumbar 2 Z value
L3BMD	Lumbar 3 BMD
L3T	Lumbar 3 T value
L3Z	Lumbar 3 Z value
L4BMD	Lumbar 4 BMD

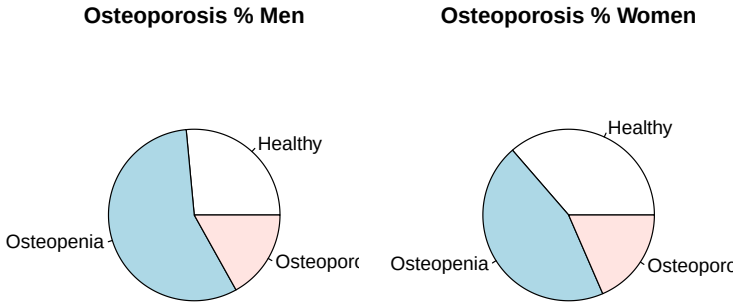
L4T	Lumbar 4 T value
L4Z	Lumbar 4 Z value
L1L4BMD	Lumbar 1-4 BMD
L1L4T	Lumbar 1-4 T value
L1L4Z	Lumbar 1-4 Z value
L2L4BMD	Lumbar 2-4 BMD
L2L4T	Lumbar 2-4 T value
L2L4Z	Lumbar 2-4 Z value
NeckFBMD	Neck Femur BMD
NeckFT	Neck femur T value
NeckFZ	Neck Femur Z value
WardsBMD	Wards BMD
WardsT	Wards T value
WardsZ	Wards Z value
TrochBMD	Trochanter BMD
TrochT	Trochanter T value
TrochZ	Trochanter Z value
TotalFBMD	Total Femur BMD
TotalFT	Total Femur T value
TotalFZ	Total Femur Z value
TotalBMD	Total Body BMD

Disease and Calculated variables

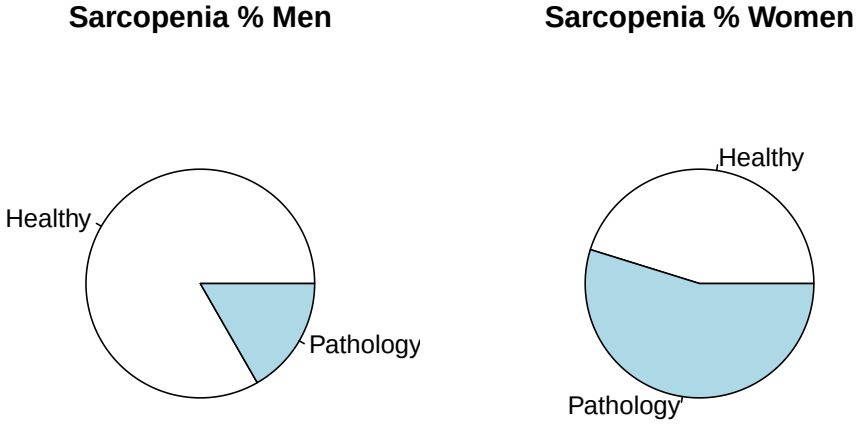
BMI	Body mass index
BMI_cat	Body mass index cat
FMI	Fat mass index
FFMI	Free fat mass index
Apendicularleanmas	Apendicularleanmas
FMR	Fat mass ratio
FTrunkgFLegsg	Fat Trunk/Fat legs grs
Indexdistributionfat	Index distribution fat
FtrunkpFlimbsp	Fat Trunk/Fat limbs%
FtrunkgFtotalg	Fat Trunk/Fat Total grs
FLegsgFtotalg	Fat Legs/Fat Total grs
FlimbsgFtotalg	Fat Limbs/Fat Total grs
LLegFgBMI	Left Leg Fat/BMI grs
LLegFpBMI	Left Leg Fat/BMI %
Lipodistrophy	Lipodistrophy
Sarcopenia	Sarcopenia
LipoSarcop	Lipodistrophy or Sarcopenia
phenotype	phenotipe
minTscore	minTscore
Tscore_3cat	Tscore 3 levels
Tscore_2cat	Tscore 2 levels

9.3 APPENDIX 3 (GRAPHICS)

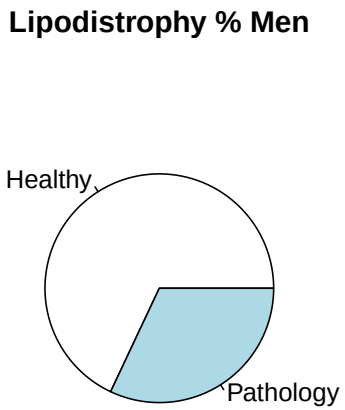
Graphic Osteoporosis % Men and Women

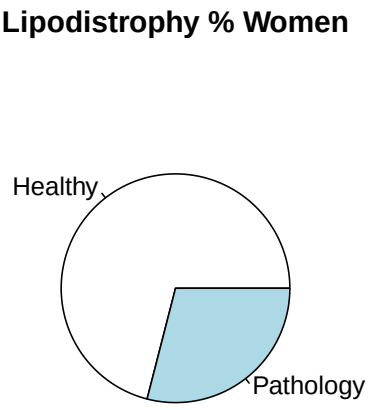


Graphic Sarcopenia % Men and Women

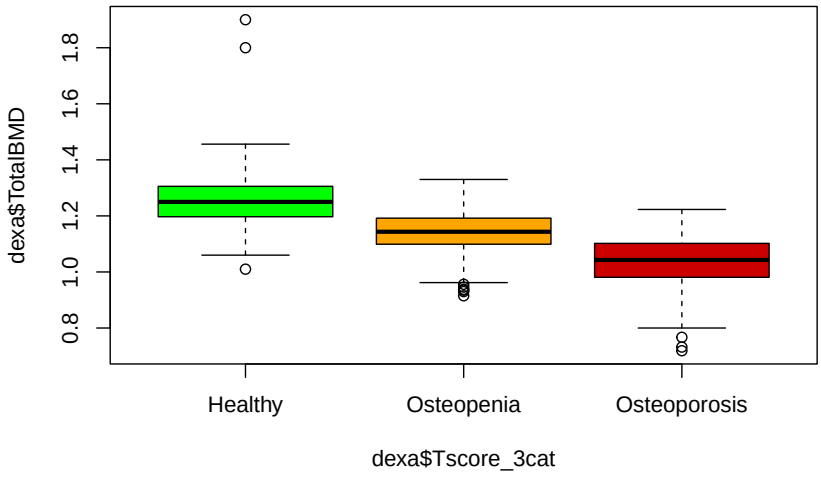


Graphic Lipodostrophy % Men and Women

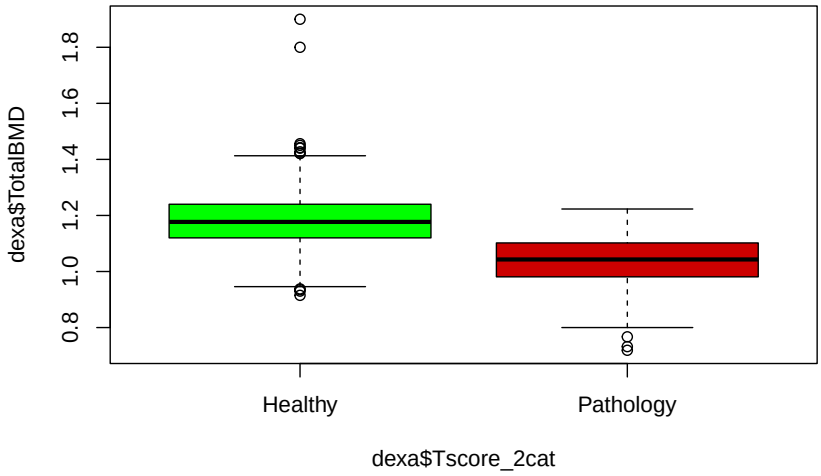




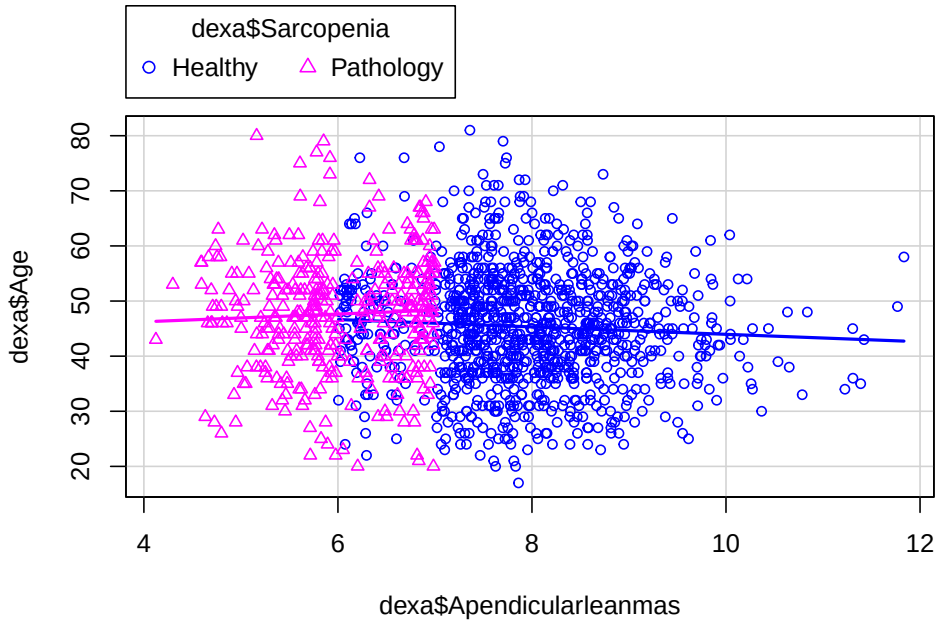
Graphic of Total BMD in Osteoporosis/Osteopenia/Healthy

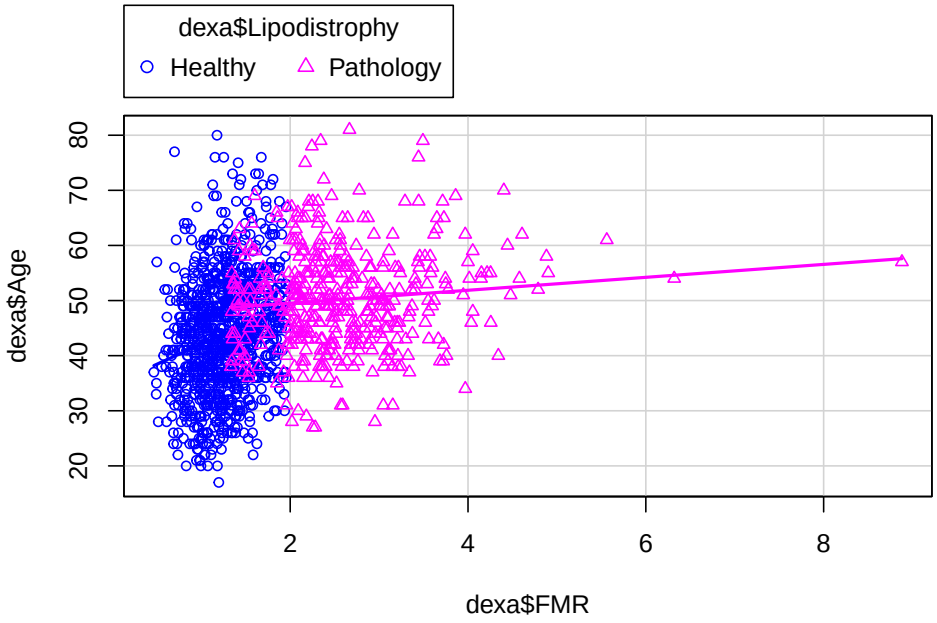


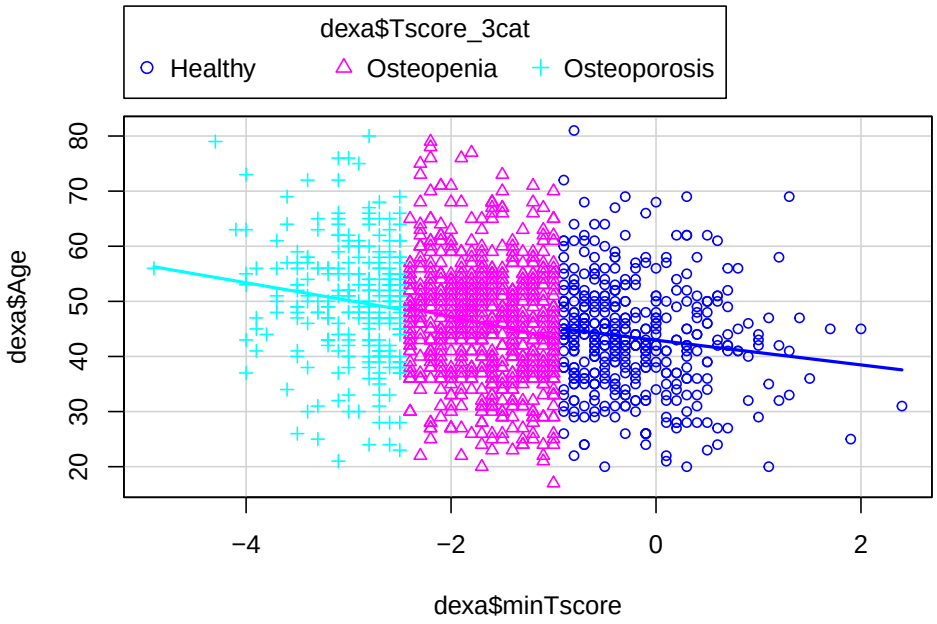
Graphic Osteoporosis/Healthy

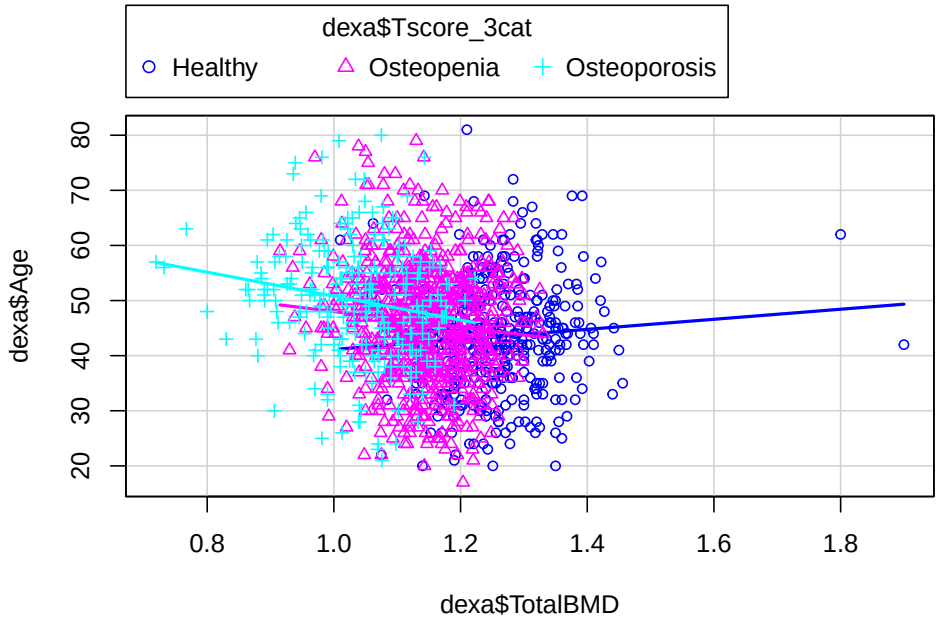


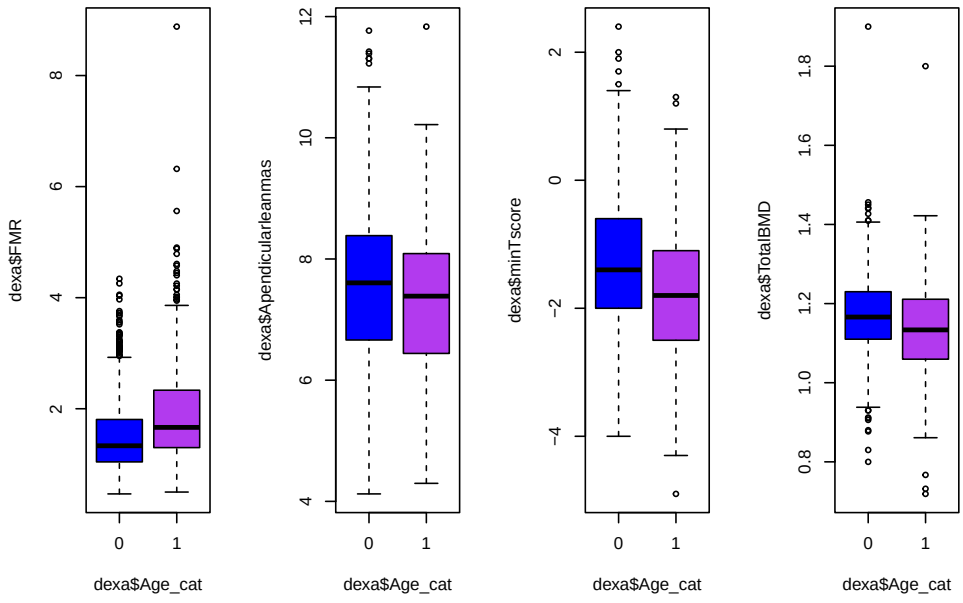
Graphics of variable age related the the three pathologies and TotalBMD.



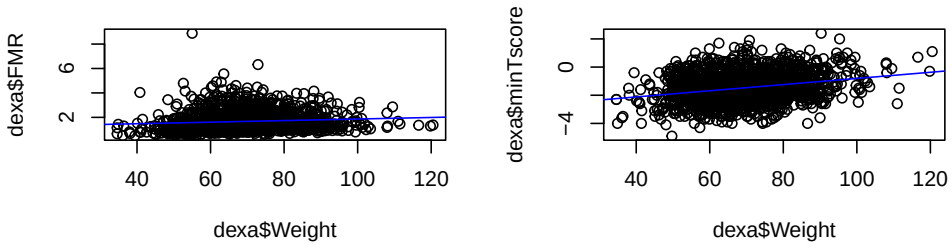

dexa$Lipodistrophy
Healthy
Pathology
dexa$Age
20
30
40
50
60
70
80
2
4
6
8
dexa$FMR

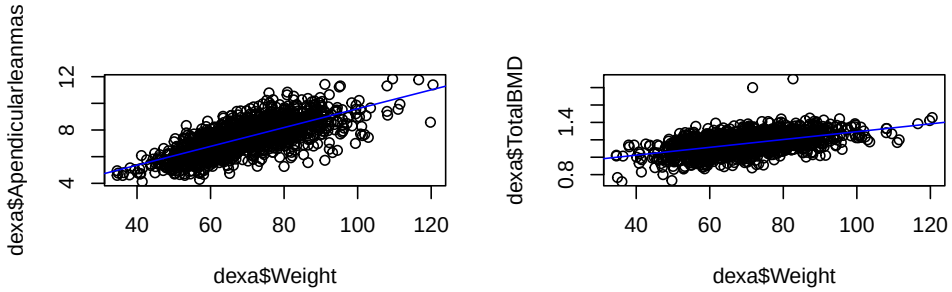

dexa$Tscore_3cat
Healthy
Osteopenia
Osteoporosis
dexa$Age
20
30
40
50
60
70
80
−4
−2
0
2
dexa$minTscore

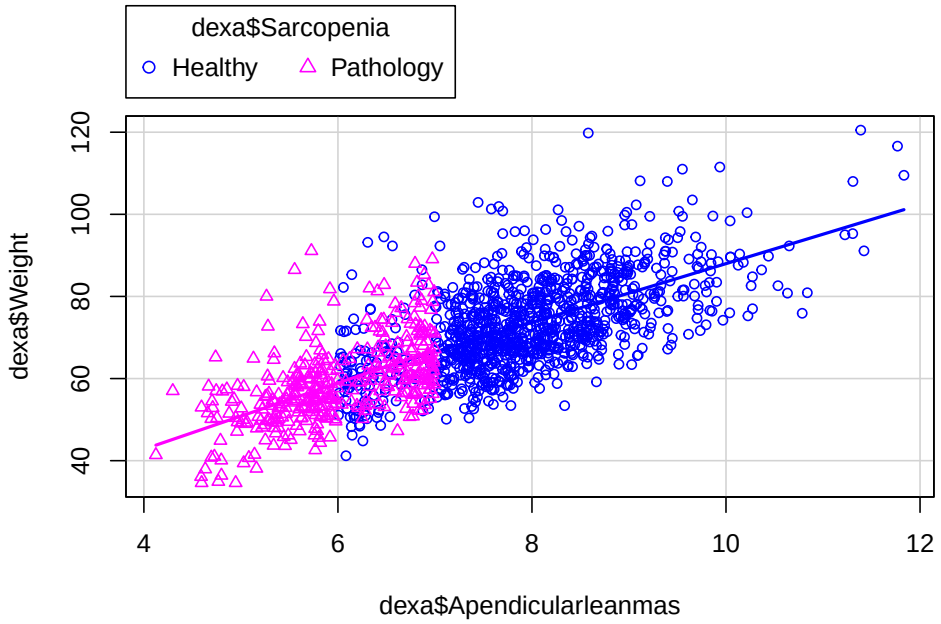

dexa$Tscore_3cat
Healthy
Osteopenia
Osteoporosis
dexa$Age
20
30
40
50
60
70
80
0.8
1.0
1.2
1.4
1.6
1.8
dexa$TotalBMD

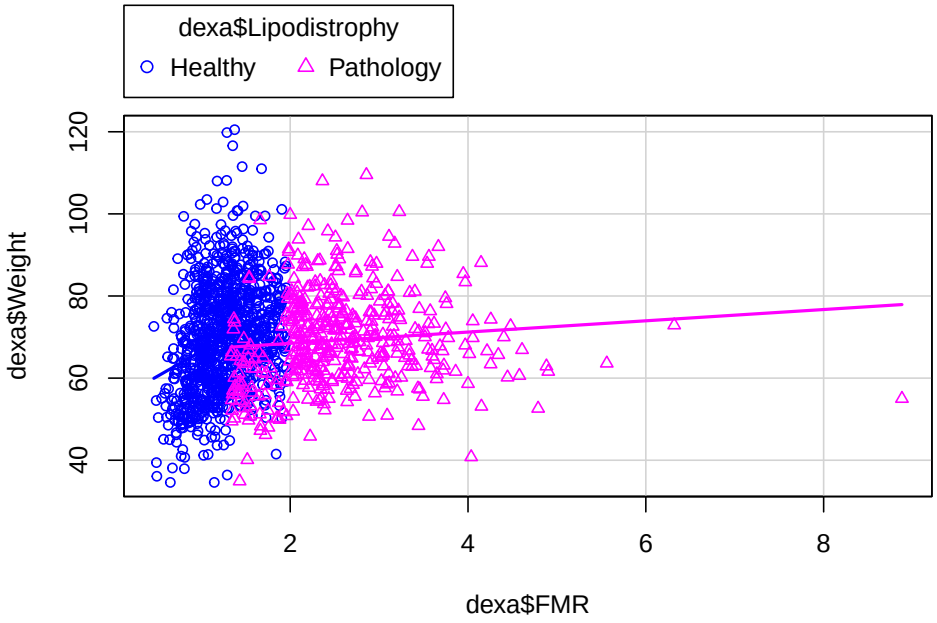

dexa$FMR
2
4
6
8
0
1
dexa$Age_cat
dexa$Apendicularleanmas
4
6
8
10
12
0
1
dexa$Age_cat
dexa$minTscore
−4
−2
0
2
0
1
dexa$Age_cat
dexa$TotalBMD
0.8
1.0
1.2
1.4
1.6
1.8
0
1
dexa$Age_cat

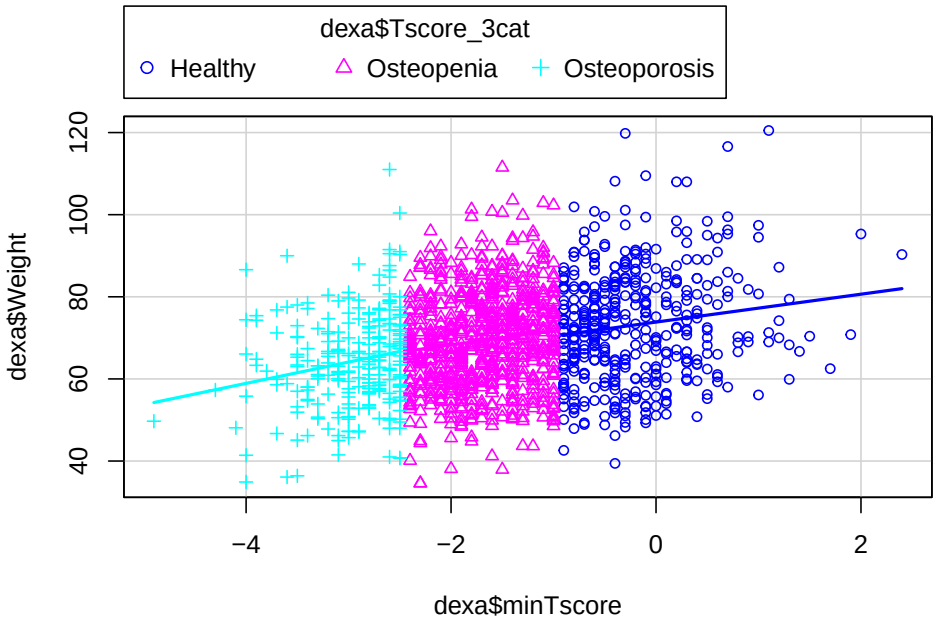
Graphics of variable weight related to the three pathologies and TotalBMD.

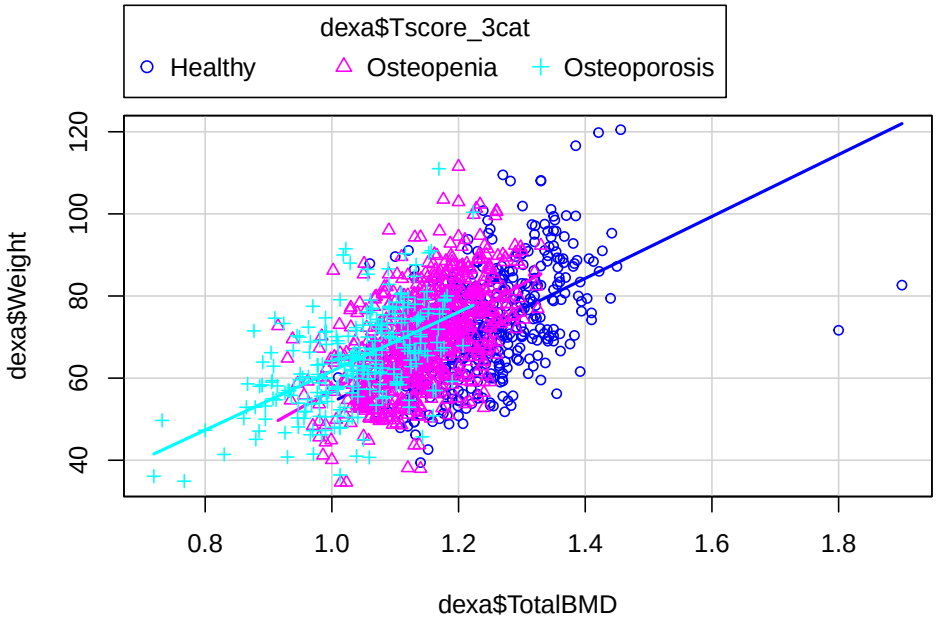




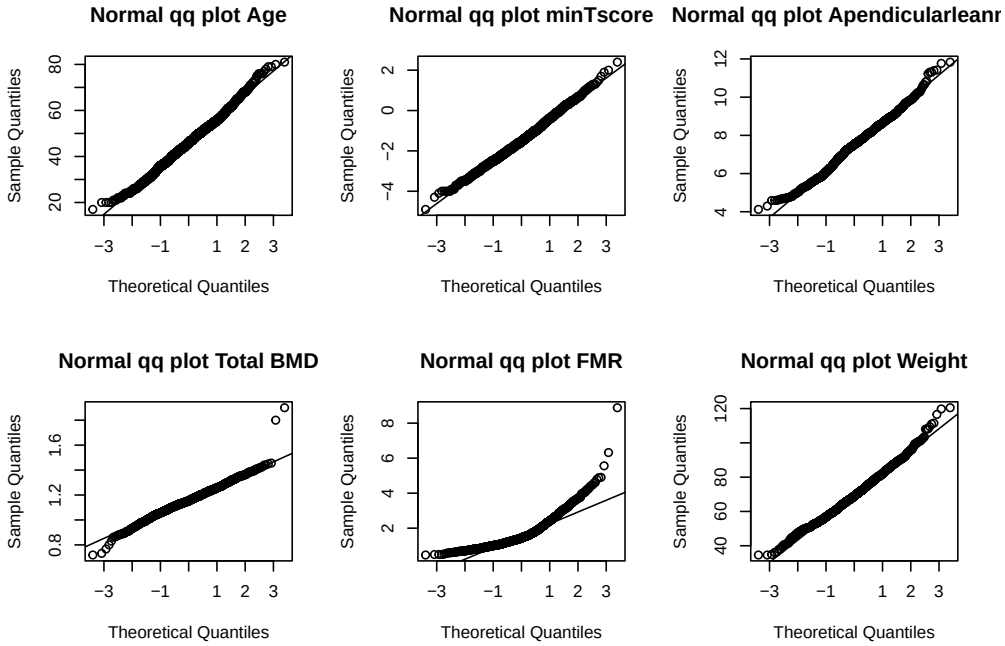



dexa$Lipodistrophy
Healthy
Pathology
dexa$Weight
40
60
80
100
120
2
4
6
8
dexa$FMR

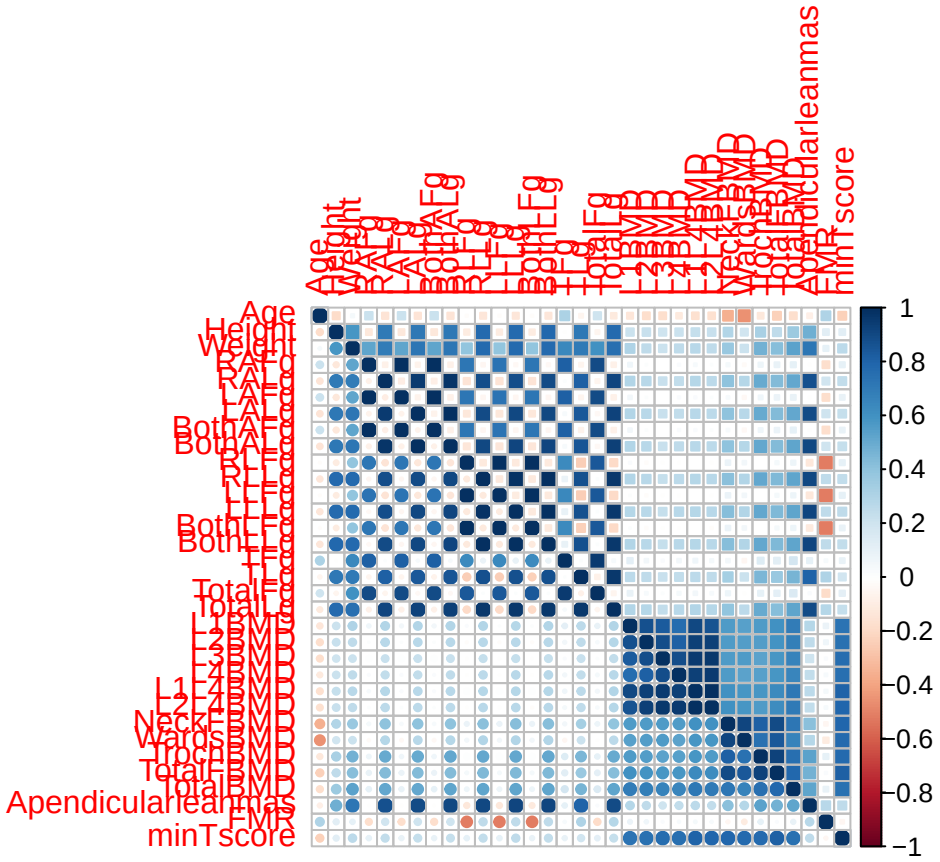

dexa$Tscore_3cat
Healthy
Osteopenia
Osteoporosis
dexa$Weight
40
60
80
100
120
−4
−2
0
2
dexa$minTscore



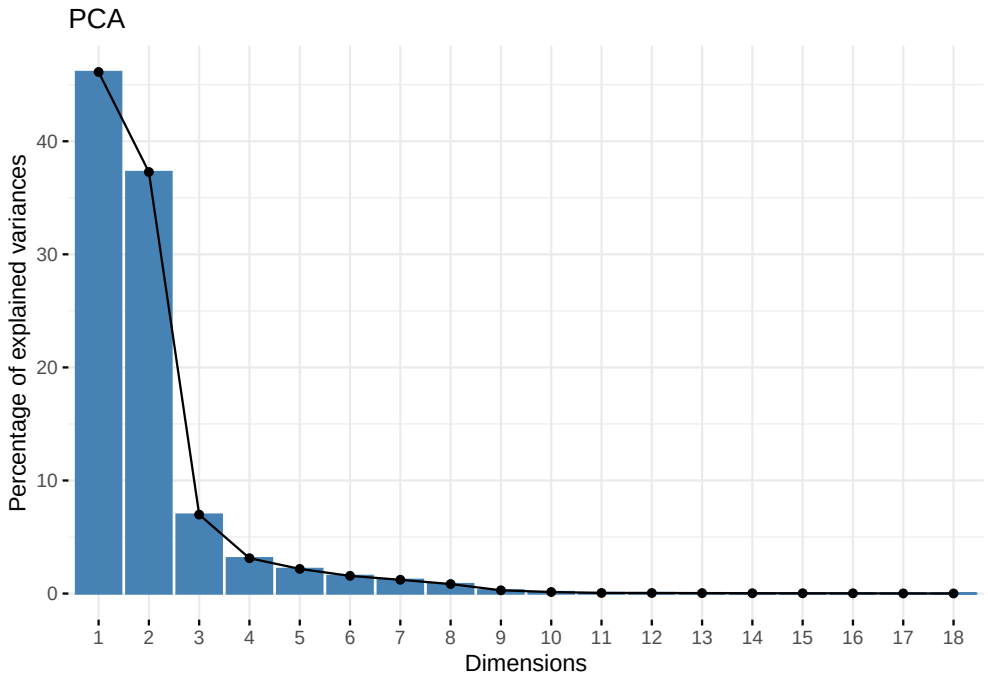
Normal qqplot of age,TotalBMD,minTscore,FMR,Alm and weight

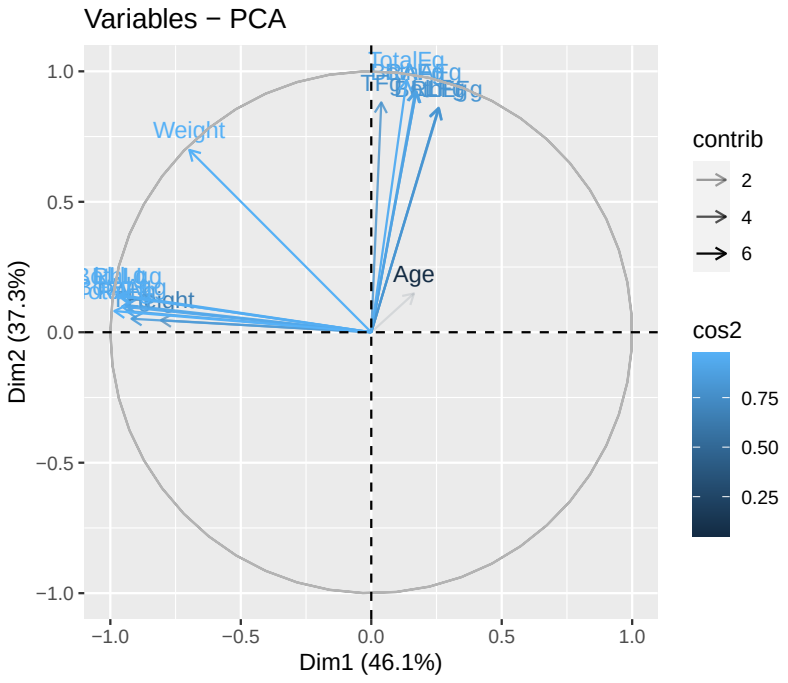


Correlation between main variables and the variables of the three pathologies and TotalBMD.

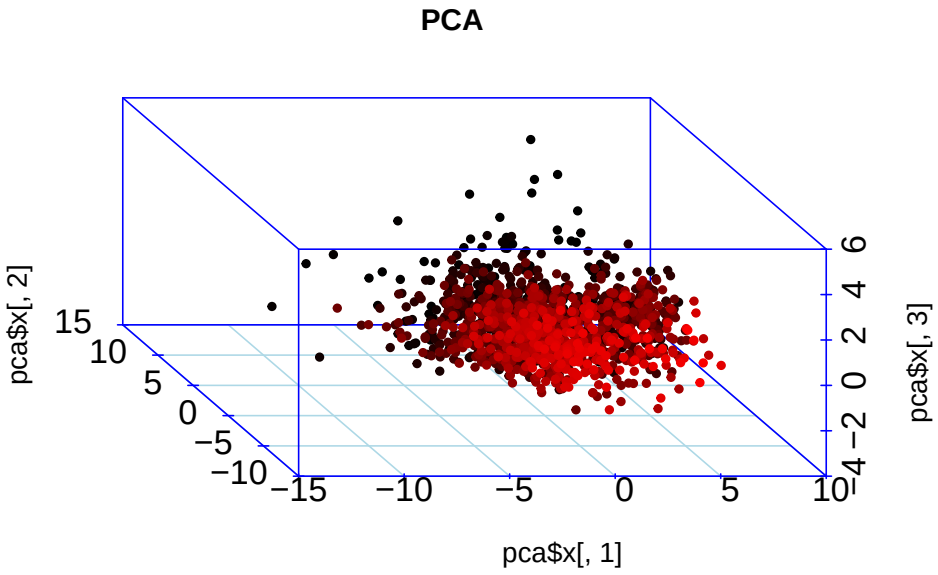


PCA graphics of the mean and fat variables with TotalBMD.

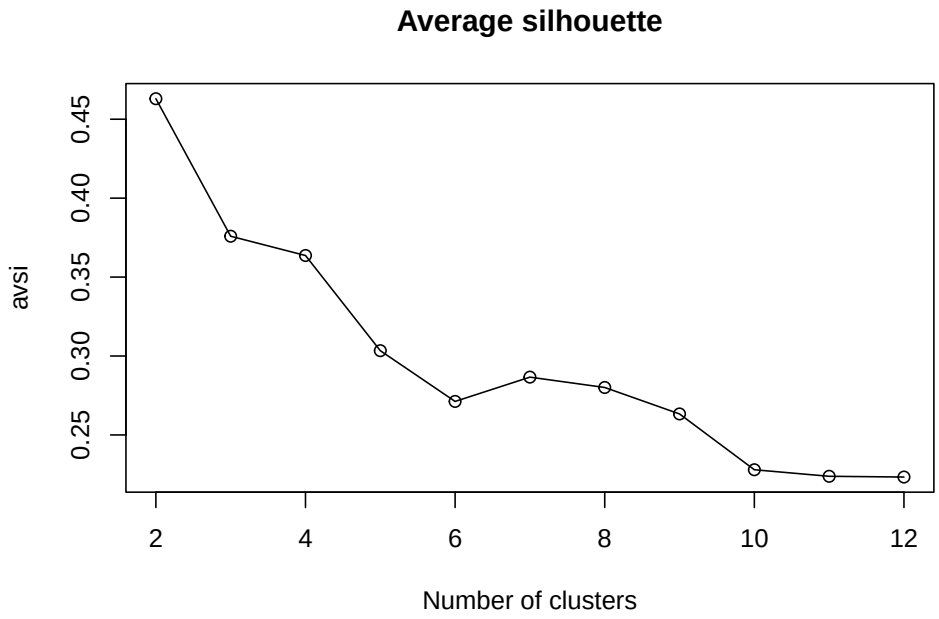




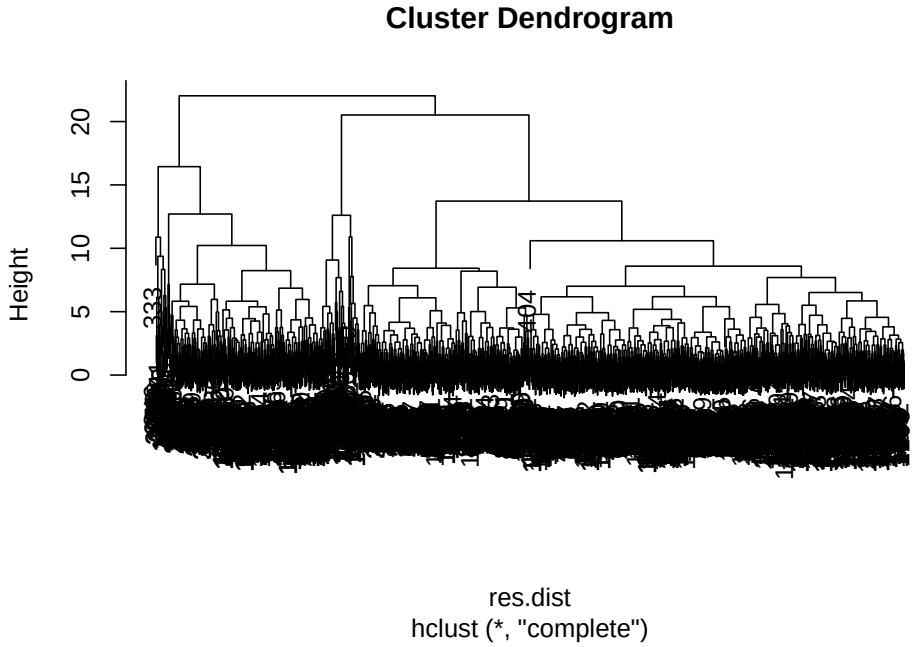
Scatterplot of the three first components.



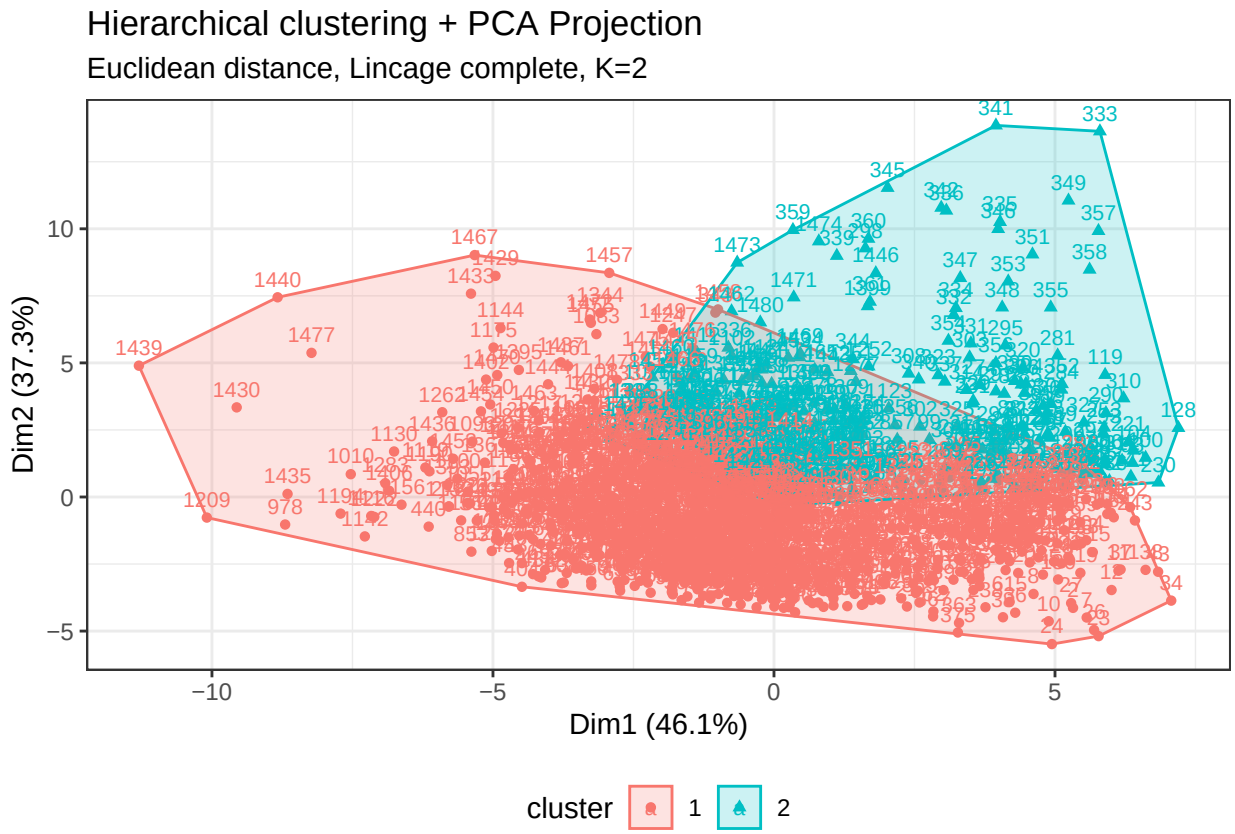
Silouette graphic



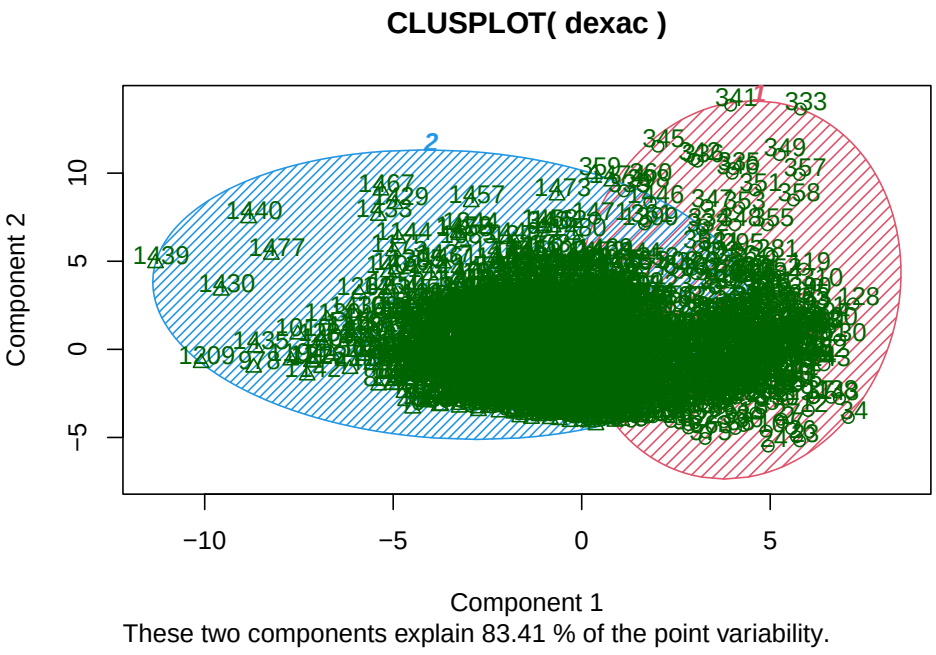
Cluster Dendogram



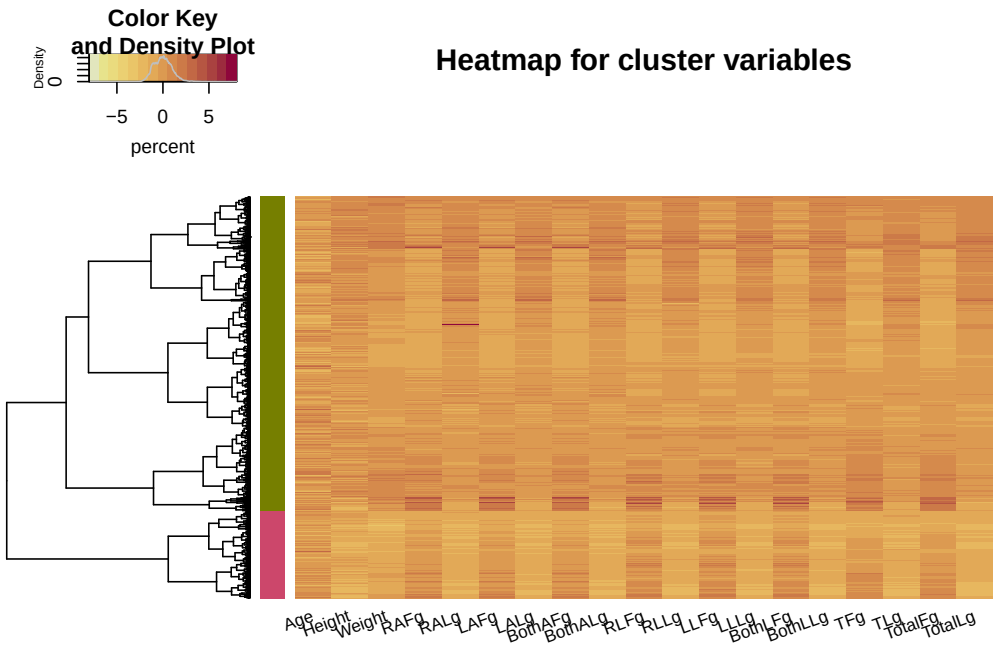
Hierarchical clustering



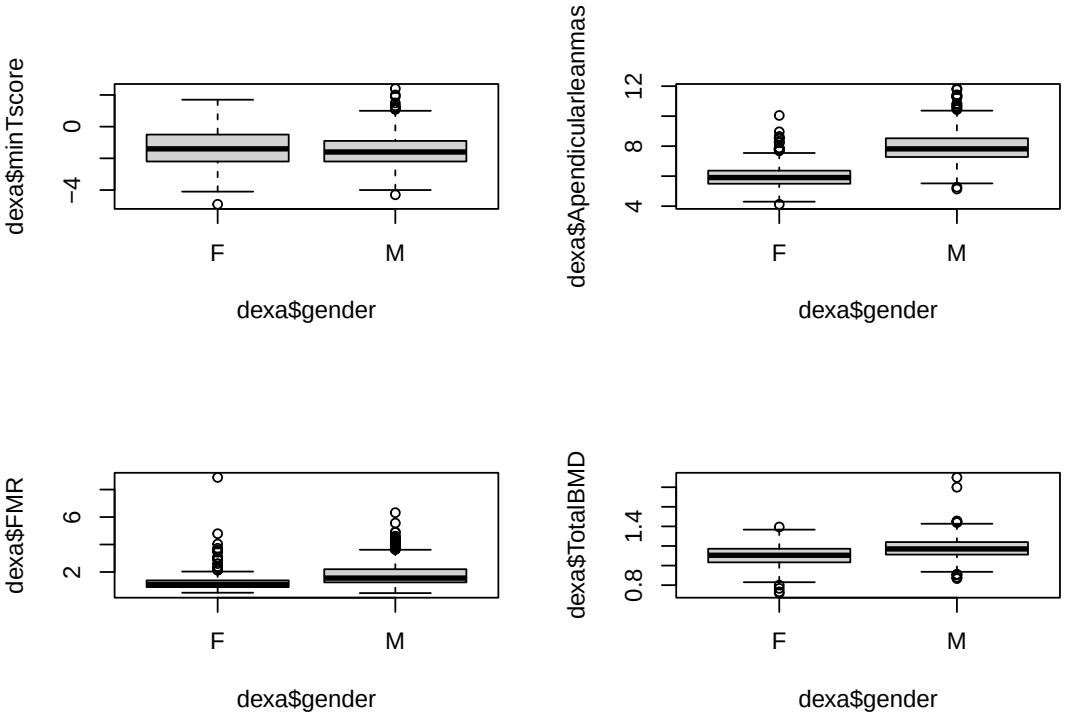
PAM Cluster



Heatmap

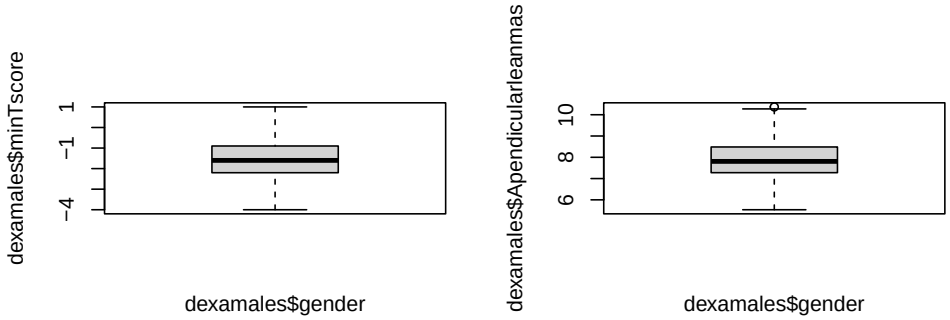


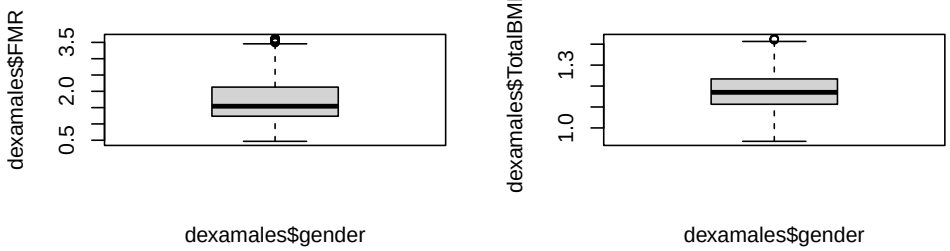
Sex variable related to the three pathologies and TotalBMD



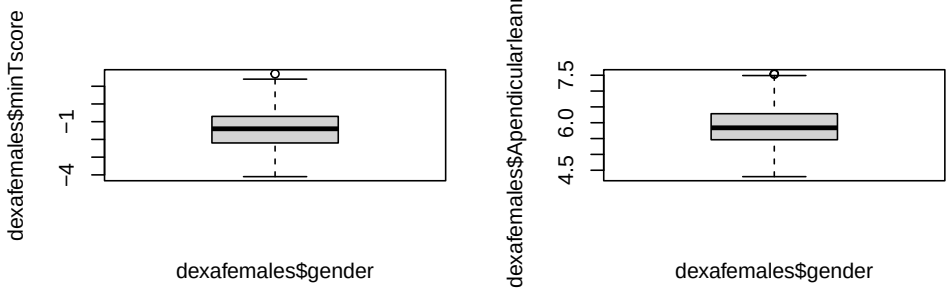
Outliers

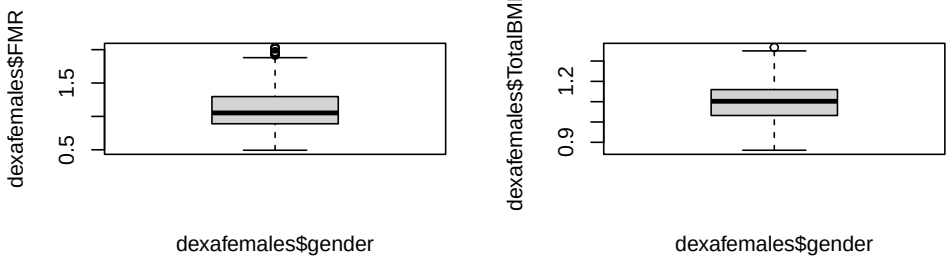
Males





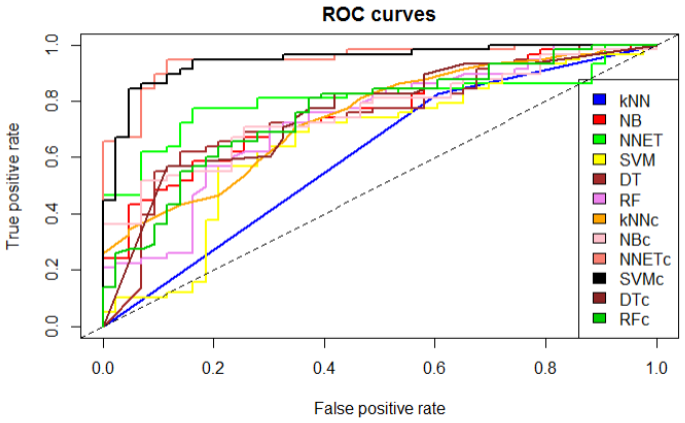
Females



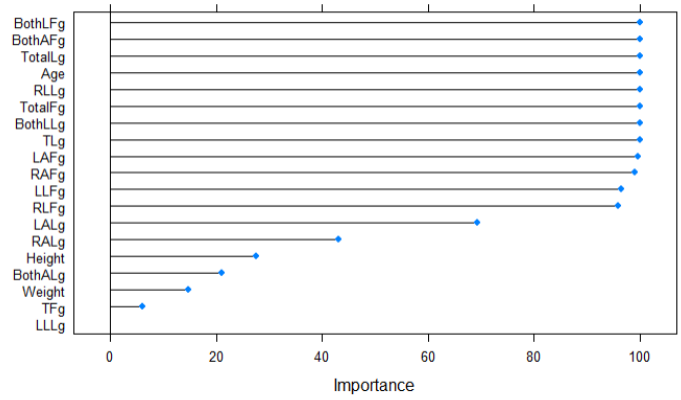


9.3.1 RESULTS'S GRAPHICS

CATEGORICAL SARCOPENIA FEMALES

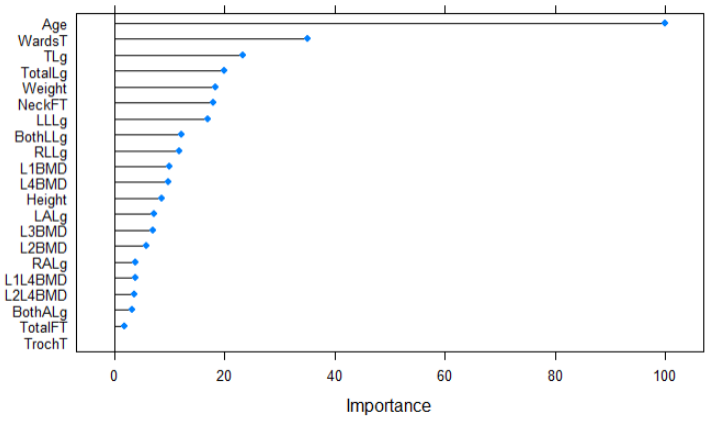


CATEGORICAL TOTAL BMD MALES

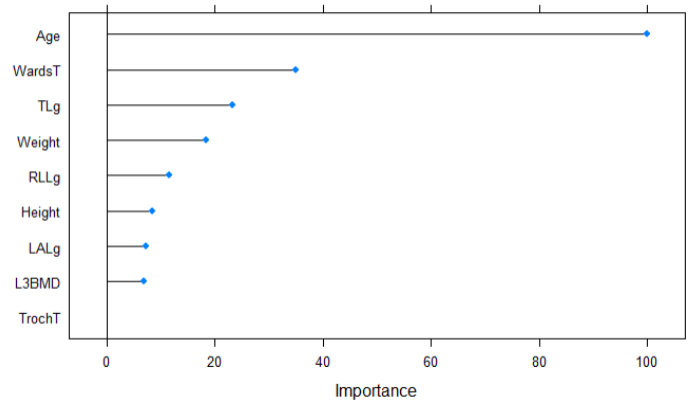


NUMERICAL LIPODISTROPHY MALES

Bayes GLM importance

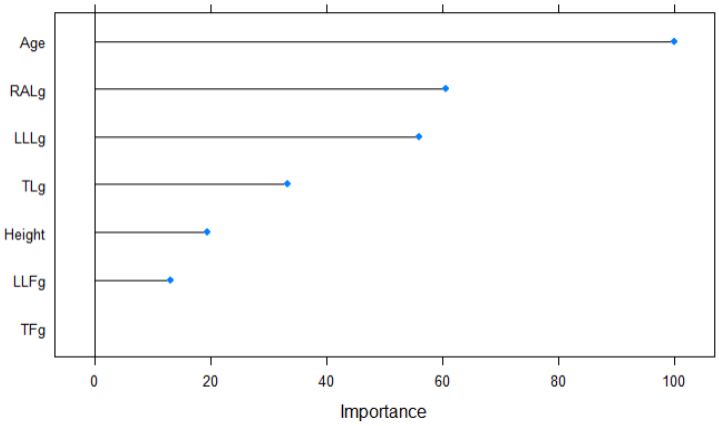


Neuralnet importance with no high correlation variables.

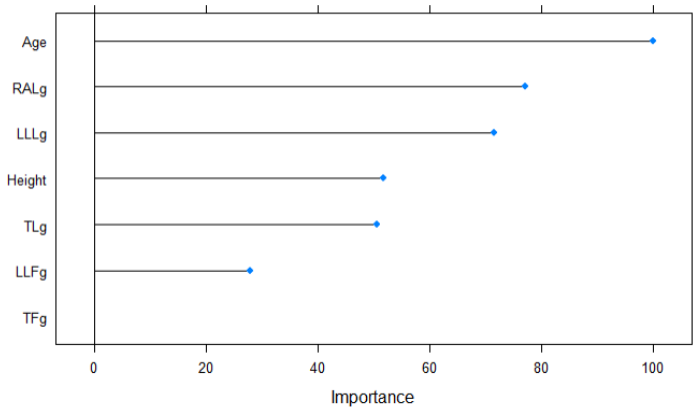


NUMERICAL OSTEOPOROSIS FEMALES

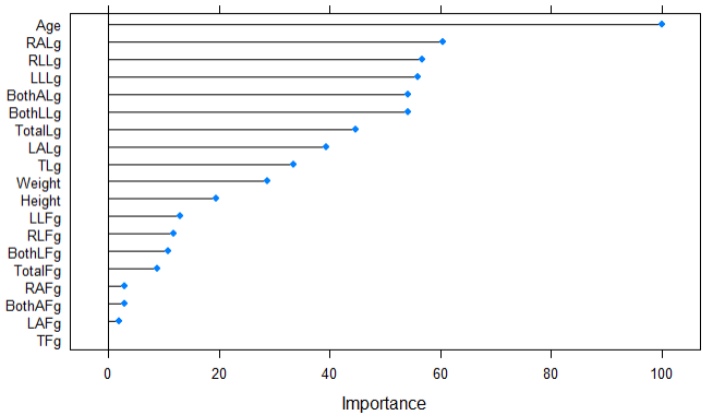
Neuralnet mportance with no high correlation variables.



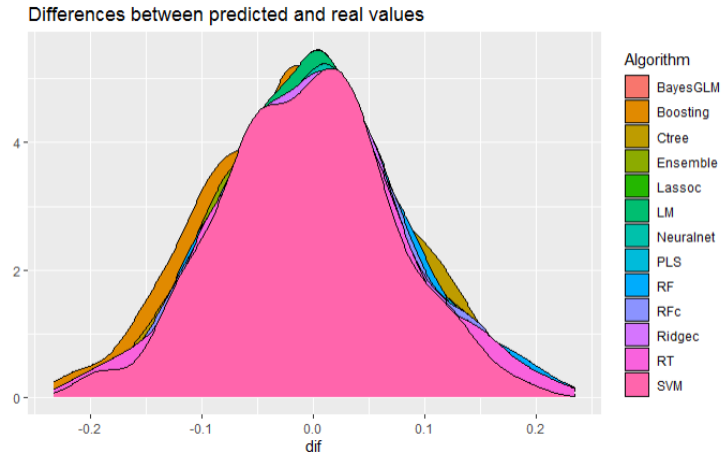
PLS importance with no high correlation variables.

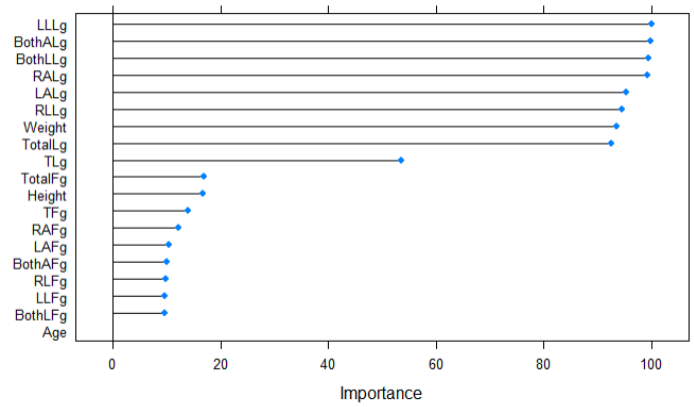


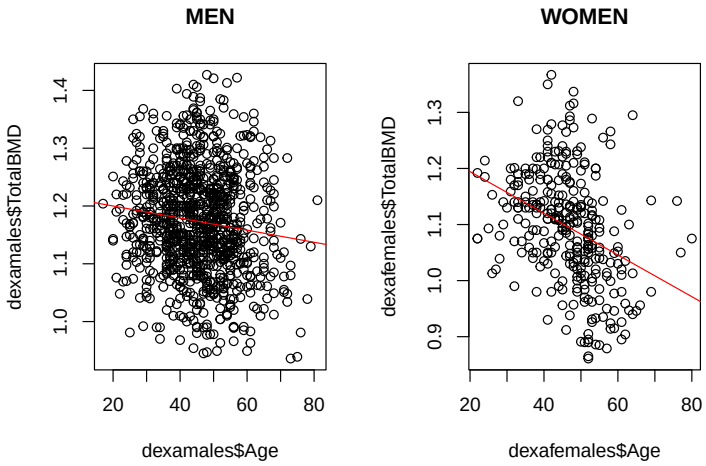
Ridge regression



NUMERICAL PREDICTION TOTAL BMD MALES






MEN
dexamales$TotalBMD
1.0
1.1
1.2
1.3
1.4
20
40
60
80
dexamales$Age
WOMEN
dexafemales$TotalBMD
0.9
1.0
1.1
1.2
1.3
20
40
60
80
dexafemales$Age