QATAR UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

IMPLEMENTATION OF MACHINE LEARNING ALGORITHMS FOR CLASSIFICATION OF BONE MINERAL DENSITY TYPES BASED ON QATAR BIOBANK DATA

BY

MOHAMMED AHMED

A Thesis Submitted to

the College of Arts and Sciences

in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Applied Statistics

January 2023

COMMITTEE PAGE

The members of the Committee approve the Thesis of
Mohammed Ahmed defended on [Defense Date].

Dr. Saddam Akber Abbasi
Thesis/Dissertation Supervisor

Dr. Atiyeh Abdallah
Co-supervisor

Dr. Adegoke Nurudeen
Committee Member

Dr. Faiz Elfaki
Committee Member

Approved:

Ahmed Elzatahry, Dean, College of Arts and Sciences

ABSTRACT

AHMED, M., Masters : January : 2023, Applied Statistics

Title: Implementation of Machine Learning Algorithms for Classification of Bone Mineral Density types based on Qatar Biobank Data

Supervisor of Thesis: Advisor's First Name, Middle Initial, Last name only.

Bone Mineral Density (BMD) test measures the amount of calcium and other minerals in specific areas of bone. Low BMD is a well-known problem and results in bone fractures in millions of people around the world. BMD can be affected by demographic factors (such as age, gender, etc.) and clinical features (such as Vitamin D level, Calcium, etc.). A large population is known to have issues related to bones due to low vitamin D levels. BMD can be generally classified into normal and low (Osteopenia) by using the BMD t-scores. It is of interest to know which factors can affect BMD and help in classification of BMD types.

We applied machine learning techniques to classify BMD levels into "Normal" or "low" using Qatar Biobank dataset. The aim is to highlight the most important variables in classifying BMD levels, and to identify which machine learning algorithm has the ability to accurately and precisely classify BMD levels.

Results showed that Random Forest (RF) was the best performing algorithm followed by Gradient Boosting. While the most important variables are "BMI", "Testosterone", "Hip-Waist ratio", "Uric Acid", "eGFR", "Ferritin", "Gender" and "Age". Research showed that we could rely on machine learning algorithms for early diagnosis of low BMD issues, which will spare time and cost

DEDICATION

To my professors.

To my family.

To Dr. Hadi Abu-Rasheed.

ACKNOWLEDGMENTS

"I would like to acknowledge the support of Qatar Biobank for providing data to achieve the requirements of this study"

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Table 1

Variables abbrevation

Abbreviation	Meaning
BMD	Bone mineral density
BMI	Body mass index
DXA	Dual X-ray Absorptiometry
SXA	Single X-ray Absorptiometry
ML	Machine learning
AI	Artificial intelligence
OSTI	Osteoporosis Self-assessment Tool Index
AUROC	Area Under the Receiver Operating Characteristics Curve
SUA	Serum Uric Acid
FM	Fat mass
LM	Lean mass
AUC	Area under the curve
DT	Decision tree
RF	Random forest
LDA	Linear discriminant analysis
QDA	Quadraric discriminant analysis
KNN	K-nearest neighbors
SVM	Support vector machine
XGBoost	extreme gradient boosting

CHAPTER 1: INTRODUCTION

1.1 Bone Mineral Density and Osteoporosis

Bone Mineral Density (BMD) test measures the amount of calcium and other minerals in specific areas of bone. Higher BMD results means denser bones that results in low risk of fracture whereas a low BMD can increase the risk of bone fractures. Low BMD leads to Osteoporosis which is a bone disorder where the bones are more fragile and more likely to break. It affects more than 75 million people around the world. In Europe and the USA, more than 2.3 million fractures occur annually. Specifically speaking about hip fractures, it is estimated to have 3 million cases by 2025, and this number could increase since age was not accounted for in the estimation (Figure 1).

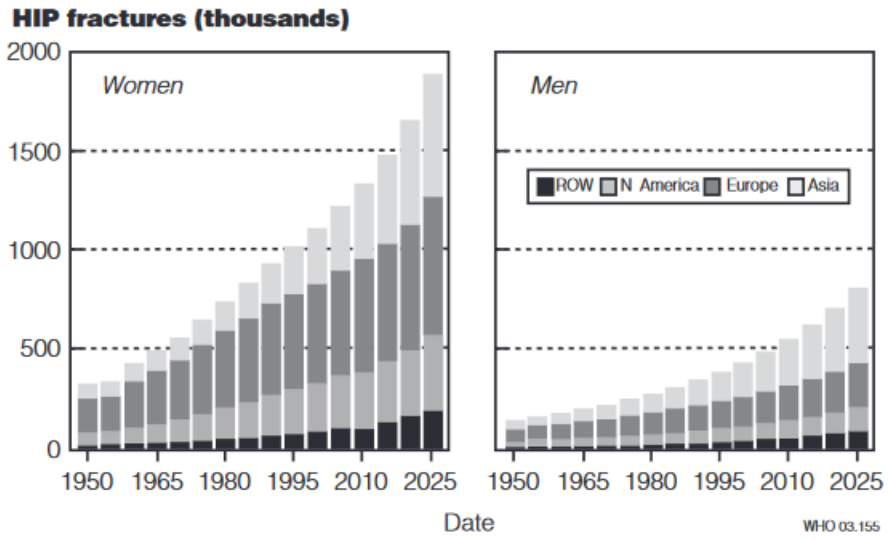

Figure 1. Hip fracture estimates from 1950 to 2025 by gender and region. (WHO, 2003).

The danger of the disease rises from the associated fractures. Hip fractures, vertebrae fractures and forearm fractures are highly associated with osteoporosis. However, fractures in other sites of the body are at least partly due to low BMD issues.

Bone mineral density (BMD), which accounts for the majority of the variations in bone tissue strength, is frequently used clinically to diagnose and assess the severity

of osteoporosis. There are several methods to measure the BMD like, Dual X-ray Absorptiometry (DXA) or Single X-ray Absorptiometry (SXA) which are methods to estimate the mineral content in the whole body or in a specific site of the body. DXA is used to measure BMD in hip and spine, while SXA is used to measure BMD in heels or wrists. However, it is also possible to use DXA to measure BMD in heels and wrists. BMD readings differs with different sites, hence, hip readings are used since it is the most correlated to osteoporosis fractures. Nowadays, DXA is considered the "Gold Standard" since it is well technically developed and biologically validated. Dual X-ray Absorptiometry (DXA) is an x-ray technology with a low dose used to measure the reduction in the x-ray beams that are passed through varying densities of bone tissue.

BMD is usually reported in T-scores, which measures how different your bone mass from a bone mass of a healthy adult (usually between 25 to 35 years). According to the WHO, BMD could be categorized into four categories based on T-scores, which are:

1- Normal: where T-score is larger the -1.
2- Low: also called Osteopenia, defined as T-score between -1 and -2.5.
3- Osteoporosis: defined as T-score less than -2.5.
4- Severe Osteoporosis: defined as T-score less than -2.5 with an existence of a fracture.

Osteoporosis is called “silent disease” because its symptoms remain hidden in the early stages of the disease, and it is not discovered until fracture occurs. Hip fracture are the most serious type of fractures. They are painful and usually require a hospital care. The mean hospital stay for a hip fracture is around 30 days, which costs the health care system a lot of expense (Figure 2). In EU, the medical cost of osteoporosis was estimated at 37 billion euros in 2010. In Japan, adding one year with a good

quality of life could cost a number between 10,000$ to 89,000$ of treatments and screening. Furthermore, osteoporosis has more serious consequences which could reach to death. In Sweden1% of all deaths is casually related to hip fractures, which is somewhere between deaths due to pancreatic cancer and deaths due to breast cancer. In 2010 in UK, 43,000 of deaths were casually related to osteoporosis. Discovering osteoporosis early is vital in bone fracture care management specially in elder people and people with diabetes.

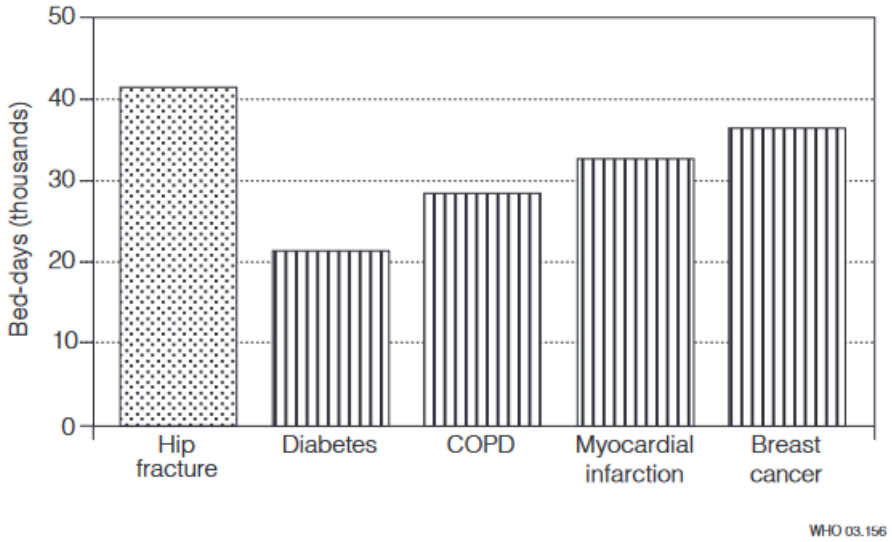

Figure 2. Hospital bed days for hip fractures compared to other diseases in women 45 years or older, in Trent region in UK. (WHO, 2003).

A lot of variables are linked to BMD. According to the (WHO 2003) “Osteoporosis is three times more common in women than in men, partly because women have a lower peak bone mass and partly because of the hormonal that occur at the menopause”. Also, greater age at menopause, estrogen, diet, height, weight and calcium intake all these are found to be positively correlated with BMD. While, age, cigarette smoking, caffeine intake, history of gastric surgery and mother fracture history are found to be negatively correlated with BMD. Furthermore, many other factors are

found to be related BMD, such as: Vitamin D, protein, phosphate, vitamin K. With the increasing number of data every day, we believe that machine learning techniques can play a role in early diagnosing osteoporosis. We aim in this research to assess the performance of machine learning algorithms in classifying BMD into “Normal” or “Low”, and to highlight the variables with the highest influence on BMD.

1.2 Machine Learning

There is an increasing demand for effective methods to process large volume of data and data with high number of dimensions. Machine learning is popular solution these days for analysis big amount of data. So, what is machine learning and what is the difference between it and the classical statistics techniques known to us?

1.2.1 ***Difference and Comparison between ML and Conventional Statistic Approaches:***

The main difference between conventional statistics approaches and machine learning (ML) techniques is that statistics methods give estimates of phenomenon based on a probabilistic model, on the other side, machine learning (ML) techniques follow algorithmic approach to mimic the data and produce similar results for any future similar case. However, there is a lot of common land between the two fields.

Statistics approaches set some hypothesis about the data while ML algorithms don't. ML can handle huge amount of data, while conventional statistics approaches were designed to handle sample sizes that are considered small to moderate nowadays. ML can handle data with huge dimension, while conventional statistics approaches performs better when the number of variables is considerably lower than the subjects.

1.2.2 Artificial Intelligence, Machine Learning and Deep Learning:

Artificial intelligence (AI) is the science of developing machines that have the ability to mimic human intelligence or behavior. Machine learning (ML) is a subset of AI, where machines or computers are trained to learn from data and recognize the trends

embedded in the data. Similarly, deep learning is a subset of machine learning where computers are trained to mimic the human brain. Neural network is considered the backbone of deep learning, where machines learn from data through an iterative layers of modeling. Figure 3 gives a good graphical summary about the three fields. In this research we are going to focus on machine learning only.

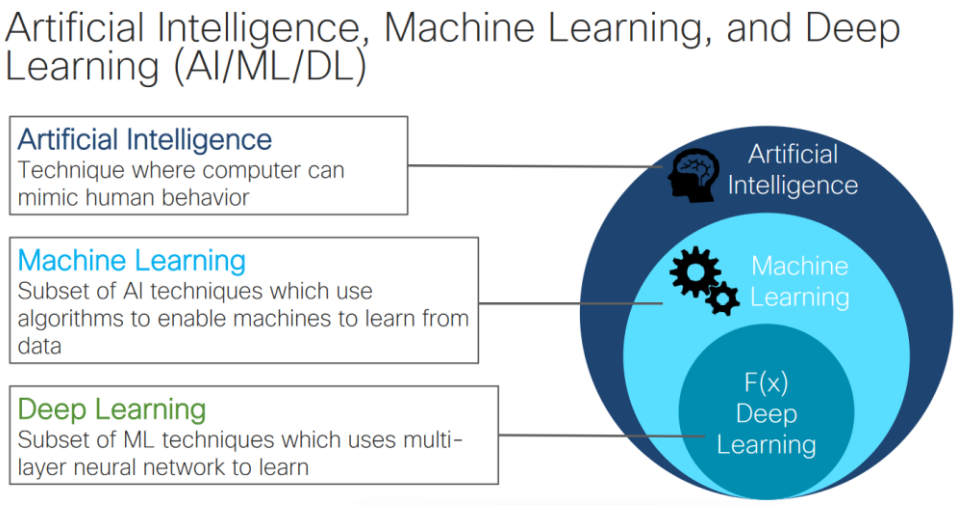

Figure 3. Relationship between Artificial Intelligence, Machine Learning and Deep Learning. (Cisco, 2019).

1.2.3 Supervised Learning vs. Un-supervised Learning:

Mainly, machine learning splits into two categories: Supervised Learning and Un-Supervised Learning. (Supervised learning) is when we use demographic and clinical variables to predict a variable of interest or an outcome. In other meaning, when we use independent variables to predict a dependent variable. The variable of interest can be either quantitative or qualitative. The presence of this outcome in the dataset, guides the learning process.

On the other hand, when we don't have a variable of interest and our goal is to distinguish and describe the clusters or the groups in the data, this is called (Un-Supervised learning). Our research scoop focuses only on supervised machine learning

1.3 Research Objective

- Calculate the reference values to provide T-scores of BMD based on age groups, gender and BMI levels for Qatar population.
- Feature selection to identify the best subset of variables that can help in classifying low BMD for Qatar population.
- Implement a variety of machine learning algorithms for classifying BMD.
- Compare the efficiency of the machine learning algorithms in classifying BMD to normal or low using a variety of performance metrics.

1.4 Research Questions

- What is the reference mean and standard deviation of BMD in Qatar population?
- What are the T-scores of BMD based on age groups, gender and BMI for Qatar population?
- Which demographic and clinical features can have a significant influence on BMD?
- Which machine learning algorithms can precisely and accurately classify the BMD type for Qatar population?

CHAPTER 2: LITERATURE REVIEW

Yoo et al. (2013) compared a number of machine learning models; Support Vector Machine (SVM), random forest, artificial neural network and logistic regression, to predict the risk of osteoporosis in Korean women and compared these methods to different conventional tools. They found that SVM outperforms all the machine learning techniques and the conventional tools as well. SVM predicted osteoporosis with 82.7% AUROC, 76.7% accuracy, 77.8% sensitivity and 76.0% specificity.

Juan et al. (2015) proposed a new model utilizes the genetic algorithm with ensemble classifier in order to predict the osteoporosis risk among Twainian women. Their approach showed a good performance with 70.43% accuracy.

Kruse et al. (2017) compared twenty-four different models in predicting hip fracture. They used DXA data collected form 4722 women and 717 men from two different Danish regions between the years 1996 to 2006. Their results showed that bootstrap aggregated flexible discriminant analysis performed the best for women with 0.92 AUC. While extreme gradient boosting performed the best for men with 0.89 AUC.

Mehta and Sebro (2019) applied SVM on dataset of 307 adults to predict bone fractures. Their model had an AUROC of 89.6%, 91.8% accuracy, 81.8% sensitivity and 97.4% specificity.

Galassi et al. (2020) applied several machine learning algorithms on a dataset of 137 post-menopausal women. Their results show that random forest has the best predictive power with an accuracy over 87%, sensitivity over 83% and specificity over 92%.

Khondaker et al. (2020) studies obesity in Qatar. Their aim is to distinguish the

healthy people from the obese ones and to highlight the risk factors associated with obesity. They implemented a case-control study and extracted the data of 500 Qatari adults (250 obese and 250 healthy) from Qatar Biobank. Their results showed a significant difference in BMD reading in obese people, where obese people have higher BMD scores.

Shim et al. (2020) applied seven different techniques to predict the risk of osteoporosis. Those methods are: k-nearest neighbors, decision tree, random forest, gradient boosting ma-chine, support vector machine, artificial neural networks and logistic regression. This study was conducted on postmenopausal women in Korea. They extracted the medical data from the Korea National Health and Nutrition Examination Surveys. Using AUROC, results showed that artificial neural network performed the best followed by random forest.

Erjiang et al. (2021) compared seven machine learning techniques to Osteoporosis Self-assessment Tool Index (OSTi). Those techniques are: catboost, extreme gradient boosting, neural network, discriminant analysis, random forest, logistic regression and support vector machine. Study was done on adult patients in West of Ireland between Jan 2000 and Nov 2018. Results shows that OSTi scored 72.3% AUROC for men and 81.0% for women. While the best performing machine learning technique was extreme gradient boosting with 76.8% AUROC for men and 83.3% for women.

Ibrahim et al. (2021) investigated the relationship between Serum Uric Acid (SUA) and the Bone Mineral Density (BMD). They extracted the information of 2981 Qatari adults from Qatar Biobank. They implemented multiple regression to test the association between SUA and BMD while accounting for gender and age, and they found a significant relationship. Further-more, they tested the relationship while also

accounting for BMI, smoking, vitamin D, alkaline phosphate (AK), and estimated glomerular filtration rate (eGFR), they relationship between SUA and BMD remained significant.

Park et al. (2021) applied three techniques which are extreme gradient boosting (XGBoost), logistic regression and neural network, on a dataset of 3309 adults aged 50 or above. They used XGBoost to select the most 20 important features, those are the features used in the analysis. Results showed that the best predictive model was the extreme gradient boosting (XGBoost) with a 73% AUROC for men and 79% for women.

Chen et al. (2022) developed a novel model utilizing extreme gradient boosting and neural network to predict fracture risk in osteoporosis patients with diabetes. They implement-ed their study on a dataset of 1603 adult patients diagnosed with diabetes and osteoporosis. They compared the performance of their model with most of the machine learning techniques, such as: extreme gradient boosting, deep neural network, SVM, logistic regression, random forest and other techniques. Their model outperforms all the other model 90.4% accuracy followed by extreme gradient boosting with 86.1% accuracy.

Kerkadi et al. (2022) studied the relationship between Bone Mineral Density (BMD) and body composition which are fat mass (FM) and lean mass (LM) in the Qatari women. They extracted the data of 2000 Qatari women from Qatar Biobank. Their results showed that the relationship between BMD and FM is not linear. Also, LM is a strong predictor of BMD.

Abdulla et al. (2022) studied the hip fracture rates in Qatar. They estimated the lifetime risk of hip fracture for people over 50 years old in Qatar and compare it with the simi-lar estimates in Kuwait, Abu-Dhabi and Saudi Arabia. Estimates of hip fracture

in Qatar was lower than Kuwait, but higher than Abu-Dhabi and Saudi Arabia estimates.

CHAPTER 3: MACHINE LEARNING ALGORITHMS

3.1 Logistic Regression

In simple linear regression, we assume that there is a linear relationship between X and Y, and this linear relationship can be expressed in the following mathematical form:

$$Y = \beta_0 + \beta_1 X \qquad (1)$$

In logistic regression, we model the probability that Y belongs to a particular category.

$$\Pr(Y = 1|X) \text{ or } \Pr(X) \qquad (2)$$

But, in order to not predict output outside the range (0,1), we use a function that gives an output between 0 and 1 for all values of X. The most famous model for this task is the logistic function:

$$\Pr(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \qquad (3)$$

Logistic regression has a well-founded theory and it is used widely by statisticians, but also very popular in machine learning field and among data scientist.

3.2 Decision Tree

In decision tree, predictor space is divided into a number of regions. To build a decision tree, the predictor space, which is the set of all possible predictor values, X_1, X_2, …, X_p, is divide into L unique and non-overlapping regions, R_1, R_2, …, R_L. For all observations in region R_i we make the same prediction, which in the classification case will be the most occurring category.

At each step, the predictor space is divided into boxes not onto high dimensional shape for simplicity and ease of interpretation. The criteria that motivates the splitting is called the Entropy, which is given by the formula:

$$Entropy = -\sum_{j=1}^{J} \hat{P}_{ij} \, log \hat{P}_{ij} \qquad (4)$$

Where $\hat{P}_{ij}$ represents the proportion of observation in region i that are from the category

j. Entropy can take values equal to zero or higher. It is of desire to make the entropy as small as possible, since it represents the variability in the data.

Decision trees are simple and easy to interpret. On the other hand, they are less accurate compared to the other machine learning techniques.

3.3 Random Forest

Random forest is an extension of decision trees. Random forest is built by first taking a sample of the observation *with replacement,* then randomly choose a subset of the variables (preferably $\sqrt{P}$), lastly build a decision tree. Repeat this process N times, usually N is chosen to be large (preferably 500).

At each step, random forest builds a decision tree using a sample of available variables in the dataset, which means, randomly selecting important variables yields in an accurate prediction, and randomly selecting variables that is weakly related to the dependent variable will yield in a poor prediction. Hence, random forest has the ability to recognize the most important variables in the dataset.

Random forest reduces decision trees variability and gives more accurate results, but on the other hand it less interpretable.

3.4 Gradient Boosting

Like random forest, gradient boosting is an extension of the decision trees. Unlike the random forest, each decision tree built by gradient boosting *depend* on the previous built tree. Gradient boosting does not use the bootstrapping techniques like random forest, instead, a normal decision tree is fitted and the residuals are calculated and stored. Second step is to build a new decision tree for the saved residuals using the dataset, and adding the predictions of the second tree to the predictions of the first one and calculate the residuals. Those steps are repeated multiple times.

3.5 Linear Discriminant Analysis (LDA) & Quadratic Discriminant Analysis (QDA)

Similar to logistic regression, those techniques are common between the conventional statistics techniques and the machine learning techniques. They can be used for dimension reduction or for classification. LDA and QDA use Bayes theorem to estimate $\Pr(Y = J|X)$.

3.6 Support Vector Machine (SVM)

SVM is generalization of the *Maximal Margin Classifier* approach. Maximal margin classifier is a line drawn in way that the distance between each group and the line is the maximum. Support vector machine is an extension for that approach to handle higher number of dimensions.

CHPTER 4: METHODOLOGY AND RESULTS FOR UNBALANCED DATA

Data was collected from the Qatar Biobank. Data included information about BMD and 46 different features or variables for 5000 patients. Most of the variables are quantitative whereas few such as gender, nationality and smoking are categorical. Some of the variables available in the BMD dataset are listed in table 1 below:

Table 2

Qatar Biobank BMD Dataset Variables

Variable	**Variable**
Gender	Calcium Corrected
Nationality	Phosphorus
Age	Uric Acid
BMI	Iron
Smoking	Total Iron Binding Capacity
Sodium	Prothrombin Time (PT)
Potassium	International Normalization Ratio
Bicarbonate	Fibrinogen
Urea	Dihydroxyvitamin D Total
eGFR	Free Thyroxine T4
Glucose	Free Triiodothyronine T3
Bilirubin Total	Thyroid Stimulating Hormone
Total Protein	Ferritin
Alkaline Phosphatase	Folate
ALT (GPT) Liver enzyme	Vitamin B12
AST (GOT)	C-Peptide

Cholesterol Total	Insulin
HDL-Cholesterol	Testosterone Total
LDL-Cholesterol Calc	Estradiol
Triglyceride	HBA 1C %
Calcium	Homocysteine

Gender is categorized as male/female, nationality as Qatari/Non-Qatari and smoking as Non-smoker/Smoker/Ex-smoker. The other 44 variables are clinical variables such as: Bone Mineral Density (BMD), Body Mass Index (BMI), Sodium, Potassium, Urea and so. All those variables are quantitative variables.

4.1 T-score Calculation:

Bone Mineral Density (BMD) readings are converted into T-scores using the following formula:

$$T - score = \frac{BMD - BMD_{ref}}{SD} \qquad (5)$$

Where $BMD_{ref}\ and\ SD$ are the BMD mean and the standard deviation of the healthy young adults (25-35 years of age) in Qatar. In our case $BMD_{ref} = 1.2218$ and $SD = 0.1241$. Furthermore, a new binary variable was created based on the T-scores. This new variable classifies the patients into Normal (if $T - score > -1$) or Low BMD ($T - score < -1$). This variable is going to be considered our variable of interest in this study, where we are going to train the machine learning algorithms to classify the BMD level into Normal or Low based on a number of demographic and clinical features.

Table 2 lists the T-score values based on three variables, which are: age, gender and BMI. We notice an increase in the T-score values when age increase. Similarly, the BMI increases, the T-scores tend to increase. All mean values of T-scores are close to

each other and standard deviations are relatively small and also close to each other. On the other hand, figure 4 and 5 clearly shows that males' BMD is higher than females.

Table 3

T-scores Reference Values based on Gender, Age Groups and BMI Groups

		Male		Female	
Age Group	BMI Group	Mean BMD	SD. BMD	Mean BMD	SD. BMD
(18,30]	(18,25]	1.225	0.108	1.102	0.088
(30,40]	(18,25]	1.220	0.106	1.121	0.096
(40,50]	(18,25]	1.227	0.121	1.141	0.099
(50,82]	(18,25]	1.192	0.140	1.120	0.103
(18,30]	(25,30]	1.284	0.113	1.172	0.089
(30,40]	(25,30]	1.269	0.111	1.160	0.092
(40,50]	(25,30]	1.280	0.115	1.174	0.100
(50,82]	(25,30]	1.241	0.117	1.100	0.096
(18,30]	(30,40]	1.317	0.107	1.209	0.088
(30,40]	(30,40]	1.317	0.108	1.214	0.095
(40,50]	(30,40]	1.299	0.117	1.217	0.091
(50,82]	(30,40]	1.291	0.095	1.137	0.105
(18,30]	(40,59.9]	1.370	0.094	1.268	0.097
(30,40]	(40,59.9]	1.381	0.101	1.244	0.077
(40,50]	(40,59.9]	1.319	0.131	1.254	0.097
(50,82]	(40,59.9]	1.393	0.045	1.190	0.129

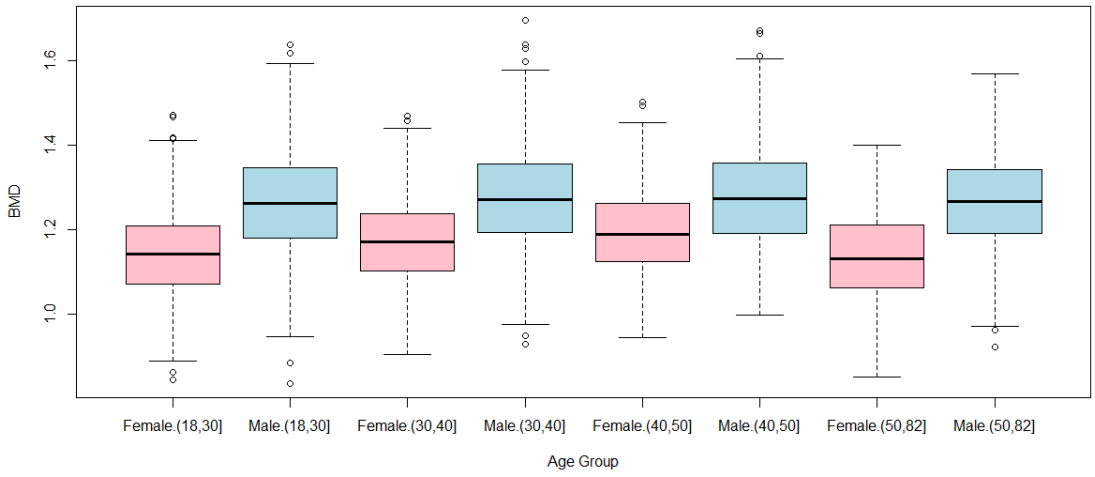

Figure 4. T-scores boxplot based on age groups and gender.

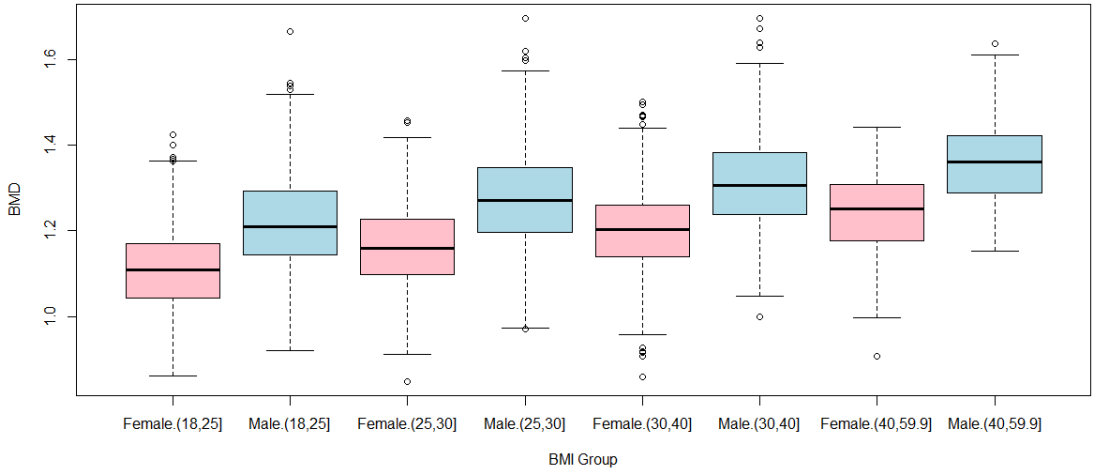

Figure 5. T-scores boxplot based on BMI groups and gender.

4.2 Data Cleaning

For several number of patients, information of many of the variables are missing, hence those observations were removed from the study. Some variables also suffered from a similar scenario; they were removed due to a high number of missing values. Moreover, some quantitative variables have in-consistent values. For example, for age, few values were <1%. These were also removed from the data. On the other hand, there were some variables suffers from few numbers of missing values which

have been kept part of the study and not removed, these missing values were substituted using the imputation techniques.

4.3 Feature selection

Since the BMD data consists of a large number of variables, random will be used to determine the most important subset of features. These features can be helpful in building useful models for classifying BMD. These important set of features are fed into different models for training.

4.4 Implementation of Machine Learning Algorithms

For the implementation of ML algorithms, the BMD dataset is divided into training and test sets. Two-thirds of the data was dedicated for training the models and one-third for testing. A number of algorithms are available in literature for classification of binary variables. Different algorithms can result in a different subset of important features. In this study, we will cover a wide range of ML models for classifying the BMD variable. Specifically, we will use the following:

- Decision Trees (DT)
- Random Forest (RF)
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- K-Nearest Neighbors (KNN)
- Logistic Regression
- Support Vector Machine (SVM)
- Gradient Boosting

For the implementation of these algorithms, R programming language will be used. Table 2 lists the R functions that will be used and their R libraries:

Table 4

ML Algorithms Functions in R and their Libraries

Method	Library	Function
Decision Tree	tree	tree
Random Forest	randomForest	randomForest
Linear Discriminant Analysis	MASS	Lda
Quadratic Discriminant Analysis	MASS	qda
K-Nearest Neighbors	caret	train
Logistic Regression	nnet	Multinom
Support Vector Machine	e1071	svm
Gradient Boosting	caret	train

4.5 Performance Measure

To compare the performance of different ML algorithms, we will make use of four criteria: Accuracy, Sensitivity, Specificity and Area Under the Receiver Operating Characteristics Curve (AUC or AUROC).

4.5.1 Accuracy

Is the most common performance measure used. It is the probability correctly predicted cases out of all predictions made.

$$Accuracy = \frac{True\ positive + True\ Negative}{Total\ Sample\ number} \qquad (6)$$

4.5.2 Sensitivity

Is the probability of correctly predicted positive cases out of all positive cases. In other words,

$$Sensitivity = \frac{True\ positive}{True\ positive + False\ Negative} \qquad (7)$$

4.5.3 Specificity

Is the probability of correctly predicted negative cases out of all negative cases.

$$Specificity = \frac{True\ negative}{True\ negative + False\ positive} \qquad (8)$$

4.5.4 Area under the Curve (AUC)

Is another performance measure that calculates the area under the Receiver Operating Characteristics (ROC) Curve. This curve is defined as the plot of sensitivity against the complement of the specificity ($1 - Specificity$). The higher the AUC, the higher is the model ability to classify the variable categories.

Table 3 include a list with all the performance measures R functions used and their R libraries.

Table 5

Performance Measures Functions in R and their Libraries

Performance Measure	Library	Function
Accuracy	MLmetric	Accuracy
Sensitivity	MLmetric	Sensitivity
Specificity	MLmetric	Specificity
AUC	MLmetric	AUC

4.6 Descriptive Statistics

Before starting to analyze the data, we take a general look on the shape of the variables. We start with the gender. Figure 4 below shows that 46.2% of our observations are females. The ratio between the two genders is very close and both genders are well represented.

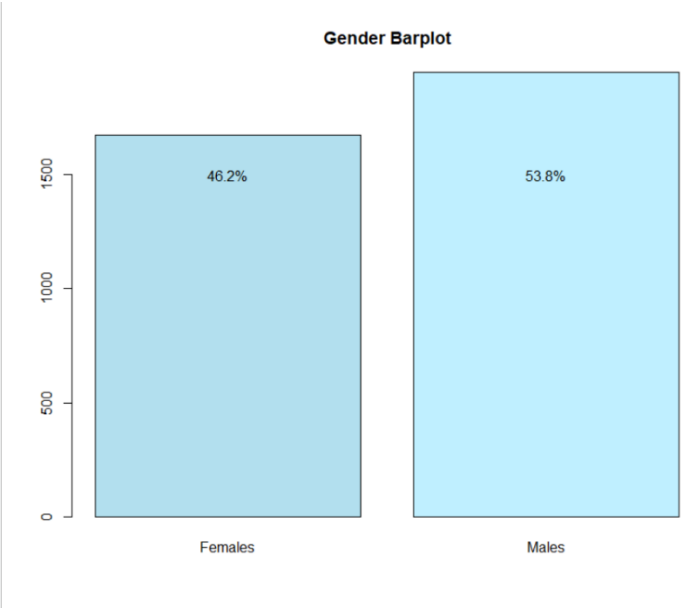

Figure 6 Gender distribution.

As you can see in Figure 5, nationality variable in divided only into Qataris and non-Qataris. The dominance of the Qatari category is obvious, more than 80% are Qataris.

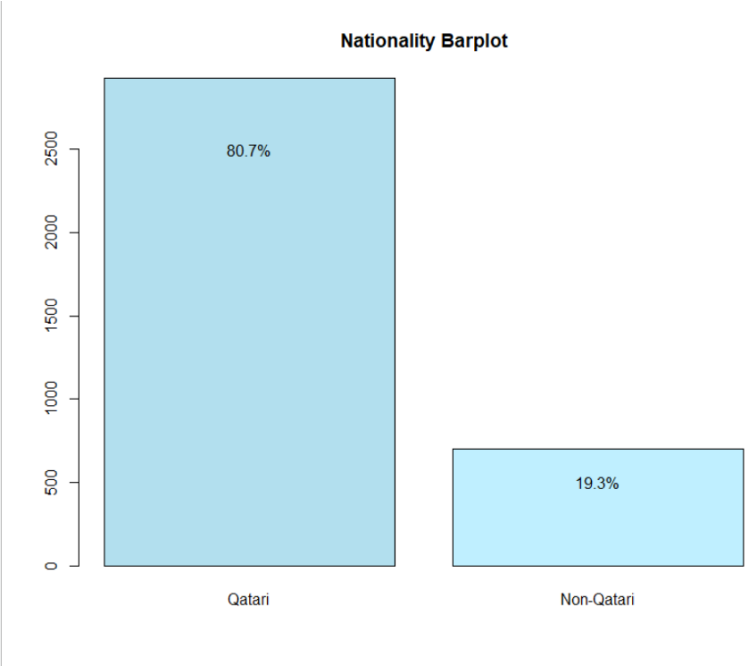

Figure 7 : Nationality distribution.

Figure 6 shows the histogram of the BMI, it looks approximately normally distributed, while the age distribution as shown in figure 7 is right skewed.

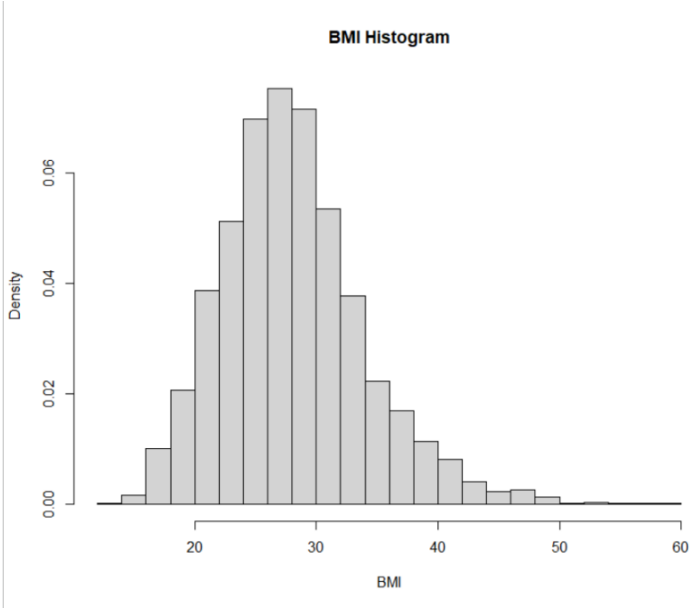

Figure 8 Body mass index (BMI) histogram.

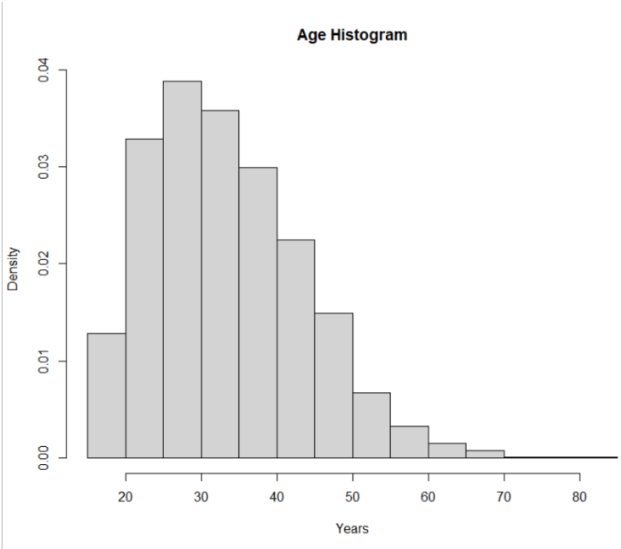

Figure 9 Age histogram.

A correlation plot is plotted to examine the relationship between the quantitative variables in the dataset. As shown in figure 8, there is no strong relationship between

the numerical independent variables.

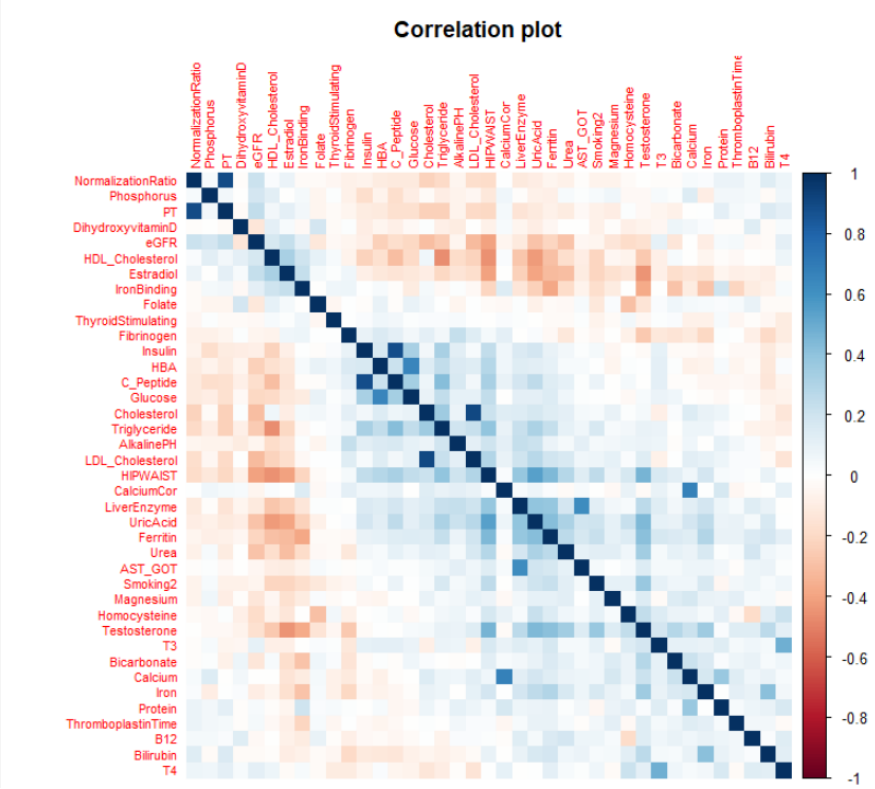

Figure 10 Correlation Plot between the Numerical Independent Variables.

Figure 9 shows that only 19% of the observations have low BMD. Apparently, the data is not balanced and there is a dominance of the Normal category over the low category.

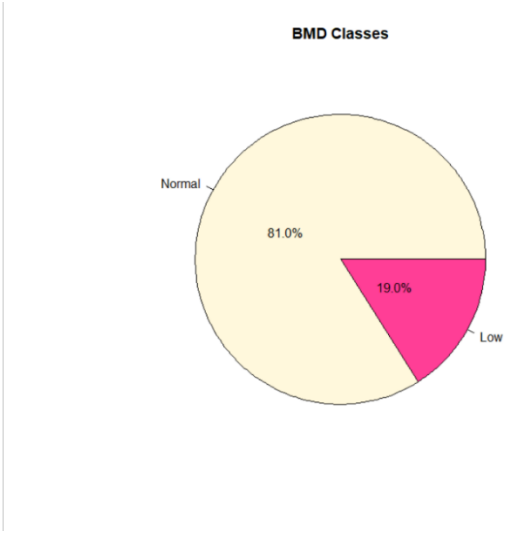

Figure 11 BMD Classes Pie Chart.

Lastly, table 4 below shows the relationship between gender and BMD. The cells represent the frequencies, and the column percentage is within the brackets. Among female, 27.2% have low BMD which is much higher than 19%. On the other hand, only 6% of males have low BMD.

Table 6

The Relationship between Gender and BMD

		Gender	
		Female	Male
BMD Levels	Normal	1220 (0.728)	1826 (0.937)
	Low	456 (0.272)	123 (0.063)

Next step is to fit and train the machine learning algorithms and to evaluate their performance.

4.7 Machine Learning Implementation

4.7.1 Decision Tree

First algorithm is decision trees. As mentioned earlier, decision tree is famous for its simplicity and ease of interpretation, plotting the decision tree explains a lot. The decision tree plot for classifying BMD levels is plotted below (figure 10):

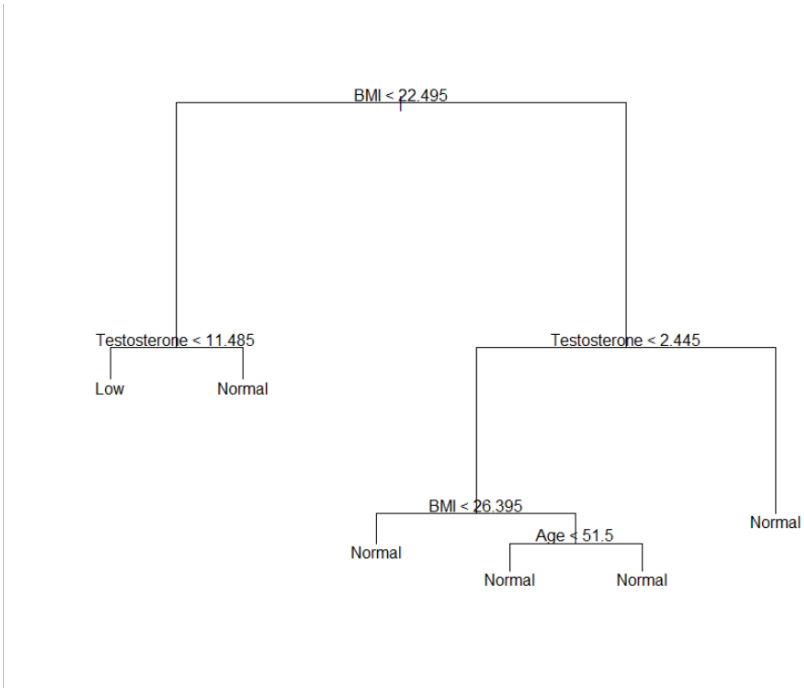

Figure 12 Decision Tree for Classifying BMD Values Based on Qatar Biobank Data.

Unfortunately, this graph isn't quite helpful. Some leaves (the nodes at the end labelled "normal" or "low" are called leaf node) can be merged since they are redundant. Hence, we pruned the tree and plotted again (figure 11):

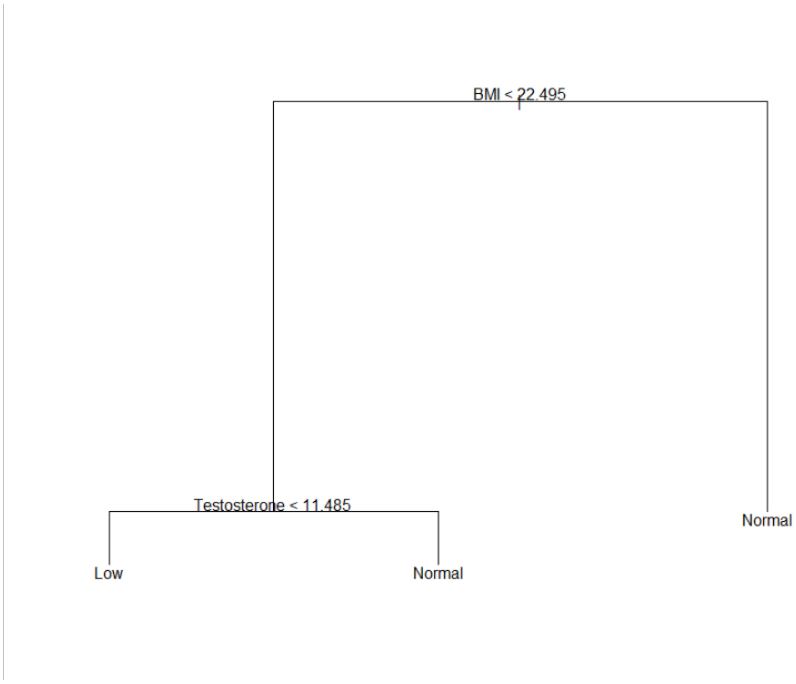

Figure 13 Pruned Decision Tree for Classifying BMD Values Based on Qatar Biobank Data.

According to the decision tree results, we need only two variables to predict the person's BMD level. Those variables are: BMI and Testosterone. This graph says that if BMI is greater than 22.5, the observation is classified as normal, otherwise we check the testosterone, if it more than 11.49 then the observation is classified as normal, otherwise it is classified as low

Decision tree performance is tested on the test set of the dataset. The confusion matrix can be found in the appendix. The confusion matrix is popular in machine learning field. It is used to take a detailed view about the algorithm performance. Moreover, the performance measures are included in tables 5 below:

Table 7

Decision Tree Performance Measures

Accuracy	Sensitivity	Specificity	AUC
0.862	0.892	0.541	0.635

Decision tree performs very good in detecting normal cases with sensitivity higher than 89%, which means decision trees can detect 89% of the normal cases correctly. On the other hand, performs very bad in detecting the low cases with 54% specificity, only 54% of the low BMD cases are detected. For better details check the decision tree confusion matrix in the appendix.

4.7.2 Random Forest

Random forest is one of the most famous and powerful machine learning algorithms. Unfortunately, its output cannot be plotted like decision tree, but it gives more accurate predictions. We trained a random forest algorithm to classify BMD levels and results are presented in table 6 below:

Table 8

RF Performance Measures

Accuracy	Sensitivity	Specificity	AUC
0.874	0.879	0.750	0.594

Random Forest results are close to the decision tree in term of accuracy and sensitivity, while we can notice an improvement in term of specificity. Accuracy was 87.4%, which means 87.4% of the prediction were correct. Sensitivity is 87.9%, which means 87.9% of the actual "Normal" BMD were correctly predicted to be "Normal" by the random forest algorithm. The model has a specificity of 75.0%, which means 75% of the actual "Low" BMD cases were correctly predicted to be "Low". AUC as an overall performance measure shows that decision tree model - unexpectedly - outperformed the random forest in classifying BMD levels.

4.7.3 Linear Discriminant Analysis (LDA)

LDA is a well-known topic among statisticians. We fitted a LDA model to classify BMD levels and results are below (table 7):

Table 9

LDA Performance Measures

Accuracy	Sensitivity	Specificity	AUC
0.880	0.894	0.688	0.645

Similarly, LDA results are close to the decision tree and random forest. Accuracy was 88%, which means 88% of the prediction were correct. Sensitivity is 89.4%, which means 89.4% of the actual "Normal" BMD were correctly predicted to be "Normal" by the LDA algorithm. The model has a specificity of 68.8%, which means 68.8% of the actual "Low" BMD cases were correctly predicted to be "Low". AUC as an overall performance measure was 0.645 which is close to decision tree and RF results.

4.7.4 Quadratic Discriminant Analysis (QDA)

QDA is an extension of the LDA technique. We fitted a QDA model to classify BMD levels and results are below:

Table 10

QDA Performance Measures

Accuracy	Sensitivity	Specificity	AUC
0.710	0.924	0.282	0.684

QDA has significantly higher sensitivity, but accuracy and specificity dropped, while AUC is close to the pervious seen methods. Accuracy was 71%, which means 71% of the prediction were correct. Sensitivity is 92.4%, which means 92.4% of the actual "Normal" BMD were correctly predicted to be "Normal" by the QDA algorithm.

The model has an extremely low specificity of 28.2%, which means 28.2% of the actual "Low" BMD cases were correctly predicted to be "Low". QDA has an AUC of 0.684.

4.7.5 K-Nearest Neighbors (KNN)

We trained a KNN algorithm to classify BMD levels and results are below (table 9):

Table 11

KNN Performance Measures

Accuracy	Sensitivity	Specificity	AUC
0.843	0.856	0.200	0.504

KNN has the worst performance so far in classifying the BMD levels into normal and high. Accuracy was 84.3%, which means 85.6% of the prediction were correct. Sensitivity is 85.6%, which means 85.6% of the actual "Normal" BMD were correctly predicted to be "Normal" by the KNN algorithm. The model has an extremely low specificity of 20.0%, which means 20.0% of the actual "Low" BMD cases were correctly predicted to be "Low". KNN has an AUC of 0.504.

4.7.6 Logistic Regression

Logistic regression is one of the most famous techniques in statistics. We fitted a logistic regression model to classify BMD level into normal and low and below are the results (table 10):

Table 12

Logistic Regression Performance Measures

Accuracy	Sensitivity	Specificity	AUC
0.868	0.891	0.582	0.634

Logistic regression results were not different than the results we saw earlier. Accuracy was 86.8%, which means 86.8% of the prediction were correct. Sensitivity is 89.1%, which means 89.1% of the actual "Normal" BMD were correctly predicted to be "Normal" by the logistic regression model. The model has a specificity of 58.2%, which means 58.2% of the actual "Low" BMD cases were correctly predicted to be "Low". Logistic regression has an AUC of 0.634.

4.7.7 Support Vector Machine (SVM)

Support vector machine is a very famous technique among data scientists. We train a SVM algorithm to classify the BMD level and below are the results (table 11):

Table 13

SVM Performance Measures

Accuracy	Sensitivity	Specificity	AUC
0.874	0.875	0.850	0.579

Accuracy was 87.4%, which means 87.4% of the prediction were correct. Sensitivity is 87.5%, which means 87.5% of the actual "Normal" BMD were correctly predicted to be "Normal" by the SVM model. The model has a specificity of 85%, which means 85% of the actual "Low" BMD cases were correctly predicted to be "Low". SVM has an AUC of 0.579.

4.7.8 Gradient Boosting

Gradient boosting is a extension to the decision tree approach like the random forest. We train a gradient boosting algorithm to classify the BMD levels and below are the results (table 12):

Table 14

Gradient Boosting Performance Measures

Accuracy	Sensitivity	Specificity	AUC
0.863	0.885	0.563	0.612

Accuracy was 86.3%, which means 86.3% of the prediction were correct. Sensitivity is 88.5%, which means 88.5% of the actual "Normal" BMD were correctly predicted to be "Normal" by the gradient boosting model. The model has a specificity of 56.3%, which means 56.3% of the actual "Low" BMD cases were correctly predicted to be "Low". Gradient boosting has an AUC of 0.612.

4.8 Feature Selection

Feature selection is an important step in the process of building a machine learning model. Decreasing the number of features or the number of variables decrease the training time and decrease the cost. In this research, we will use random forest to highlight the highest contributor variables in classifying BMD levels. We plugged all the variables in the model and used the complete dataset and ran a RF algorithm. The model will automatically highlight the most important variables. Only those variables are going to be used to fit the eight machine learning algorithms again. Seven variables were highlighted by RF as illustrated in figure 12 below.

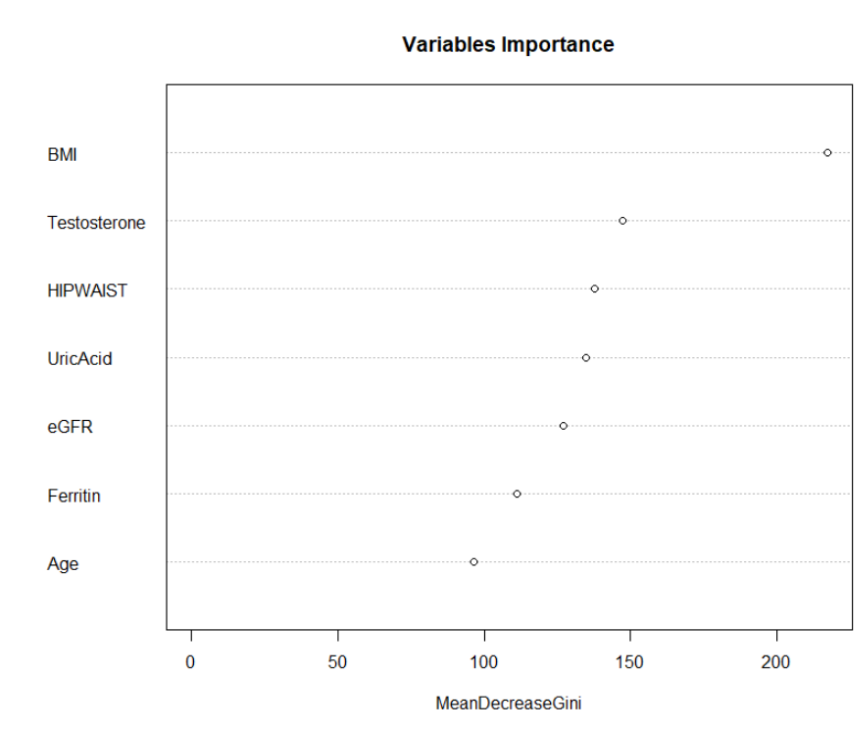

Figure 14 Top seven important variable retained by RF.

4.8.1 Decision Tree

Decision tree results didn't change at all after feature selection. Accuracy was 86.2%, which means 86.2% of the prediction were correct. Sensitivity is 89.2%, which means 89.2% of the actual "Normal" BMD were correctly predicted to be "Normal" by the decision tree model. The model has a specificity of 54.1%, which means 54.1% of the actual "Low" BMD cases were correctly predicted to be "Low". Gradient boosting has an AUC of 0.635 (Table 13).

Table 15

Decision Tree Performance Measures (With Feature Selection)

Accuracy	Sensitivity	Specificity	AUC
0.862	0.892	0.541	0.635

4.8.2 Random Forest

After feature selection the performance decreased slightly. Accuracy was 84.7%, which means 84.7% of the prediction were correct. Sensitivity is 88.2%, which means 88.2% of the actual “Normal” BMD were correctly predicted to be “Normal”. The model has a specificity of 44.2%, which means 44.2% of the actual “Low” BMD cases were correctly predicted to be “Low”. RF has an AUC of 0.635 (Table 14).

Table 16

RF Performance Measures (With Feature Selection)

Accuracy	Sensitivity	Specificity	AUC
0.847	0.882	0.442	0.598

4.8.3 Linear Discriminant Analysis (LDA)

After feature selection the performance decreased slightly. Accuracy was 86.5%, which means 86.5% of the prediction were correct. Sensitivity is 88.0%, which means 88.0% of the actual “Normal” BMD were correctly predicted to be “Normal”. The model has a specificity of 59.0%, which means 59.0% of the actual “Low” BMD cases were correctly predicted to be “Low”. LDA has an AUC of 0.597 (Table 15).

Table 17

LDA Performance Measures (With Feature Selection)

Accuracy	Sensitivity	Specificity	AUC
0.865	0.880	0.590	0.597

4.8.4 Quadratic Discriminant Analysis (QDA)

After feature selection, the accuracy and the specificity increased clearly.

Accuracy is 85.8%, which means 85.8% of the prediction were correct. Sensitivity is 89.2%, which means 89.2% of the actual “Normal” BMD were correctly predicted to be “Normal”. The model has a specificity of 51.5%, which means 51.5% of the actual “Low” BMD cases were correctly predicted to be “Low”. QDA has an AUC of 0.636 (Table 16).

Table 18

QDA Performance Measures (With Feature Selection)

Accuracy	Sensitivity	Specificity	AUC
0.858	0.892	0.515	0.636

4.8.5 K-Nearest Neighbors (KNN)

After feature selection, the specificity and AUC have increased. Accuracy is 85.0%, which means 85.0% of the prediction were correct. Sensitivity is 88.0%, which means 88.0% of the actual “Normal” BMD were correctly predicted to be “Normal”. The model has a specificity of 46.2%, which means 46.2% of the actual “Low” BMD cases were correctly predicted to be “Low”. KNN has an AUC of 0.592 (Table 17).

Table 19

KNN Performance Measures (With Feature Selection)

Accuracy	Sensitivity	Specificity	AUC
0.850	0.880	0.462	0.592

4.8.6 Logistic Regression

After feature selection, the results are very close to the pervious obtained results by logistic regression. Accuracy is 86.6%, which means 86.6% of the prediction were

correct. Sensitivity is 88.1%, which means 88.1% of the actual “Normal” BMD were correctly predicted to be “Normal”. The model has a specificity of 60.5%, which means 60.5% of the actual “Low” BMD cases were correctly predicted to be “Low”. Logistic regression has an AUC of 0.597 (Table 18).

Table 20

Logistic Regression Performance Measures (With Feature Selection)

Accuracy	Sensitivity	Specificity	AUC
0.866	0.881	0.605	0.597

4.8.7 Support Vector Machine (SVM)

Similar to the Logistic regression, the results after reducing the number of variables are very close to the pervious obtained results by SVM. Accuracy is 87.6%, which means 87.6% of the prediction were correct. Sensitivity is 87.7%, which means 87.7% of the actual “Normal” BMD were correctly predicted to be “Normal”. The model has a specificity of 82.6%, which means 82.6% of the actual “Low” BMD cases were correctly predicted to be “Low”. SVM has an AUC of 0.587 (Table 19).

Table 21

SVM Performance Measures (With Feature Selection)

Accuracy	Sensitivity	Specificity	AUC
0.876	0.877	0.826	0.587

4.8.8 Gradient Boosting

Again, the results after reducing the number of variables are very close to the pervious obtained results by gradient boosting. Accuracy is 87.0%, which means 87.0%

of the prediction were correct. Sensitivity is 88.7%, which means 88.7% of the actual "Normal" BMD were correctly predicted to be "Normal". The model has a specificity of 62.2%, which means 62.2% of the actual "Low" BMD cases were correctly predicted to be "Low". Gradient boosting has an AUC of 0.620 (Table 20).

Table 22

Gradient Boosting Performance Measures (With Feature Selection)

Accuracy	Sensitivity	Specificity	AUC
0.870	0.887	0.622	0.620

CAPTER 5: BALANCING AND RESULTS

One drawback of machine learning algorithms is that it performs poorly when there is a dominance of one category. As we saw earlier, less than 20% of our sample has low BMD. This fact reflected on the results, all algorithms had the ability to well detect the Normal category, which is not the case for the Low category. This is known as the “Accuracy paradox”.

Hence, we followed a new approach and “Balanced” the dataset. There are several techniques to handle this obstacle, “Oversampling” is just one of them. In this technique we increase the amount of the under-represented category, until it matches the amount of the second category. We randomly selected and repeated some of the cases that labelled as “Low”. We increased the sample size until the number of low BMD cases is equal to the number of normal BMD that already exist in the data.

In the next section we will plot the explanatory graphs to examine the changes in the structure of the data.

5.1 Descriptive Statistics after balancing

As shown in figure 13, the ratio between the two BMD levels is balanced.

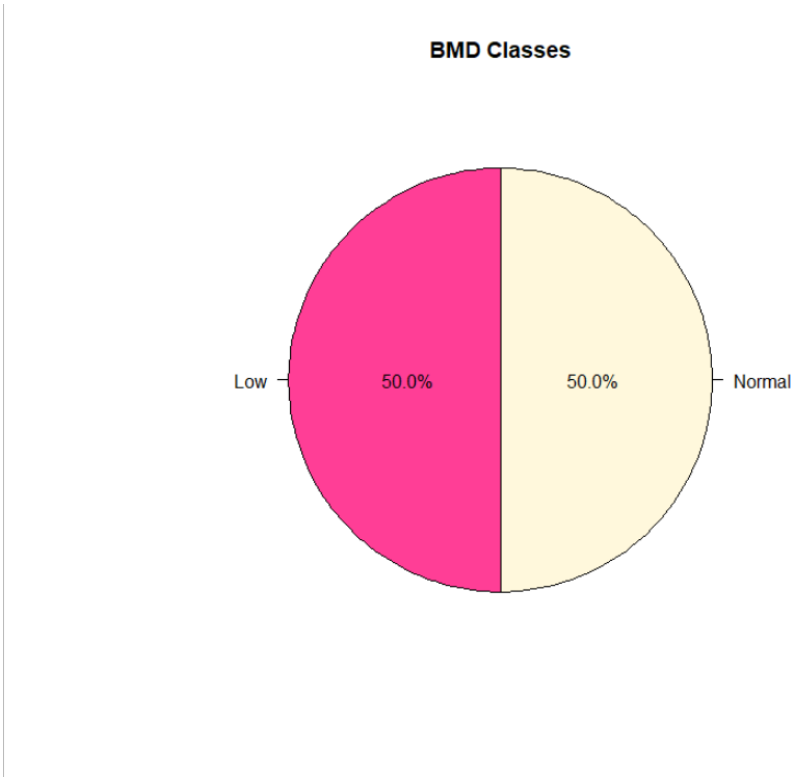

Figure 15 BMD levels after balancing the dataset.

Similarly, the gender distribution has been affected too. 59% of sample are females now (Figure 14). Based in this we expect gender to be an important factor in classifying the BMD levels.

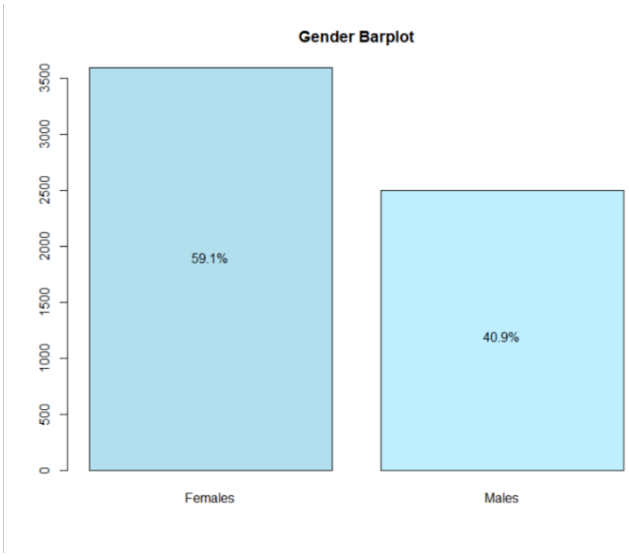

Figure 16 Gender bar chart after balancing the data.

On the other hand, nationality distribution seems to be similar. Now 17% of the population are not Qataris (Figure 15). Similarly, we don't expect nationality to be one since we could not notice a significant change in its distribution.

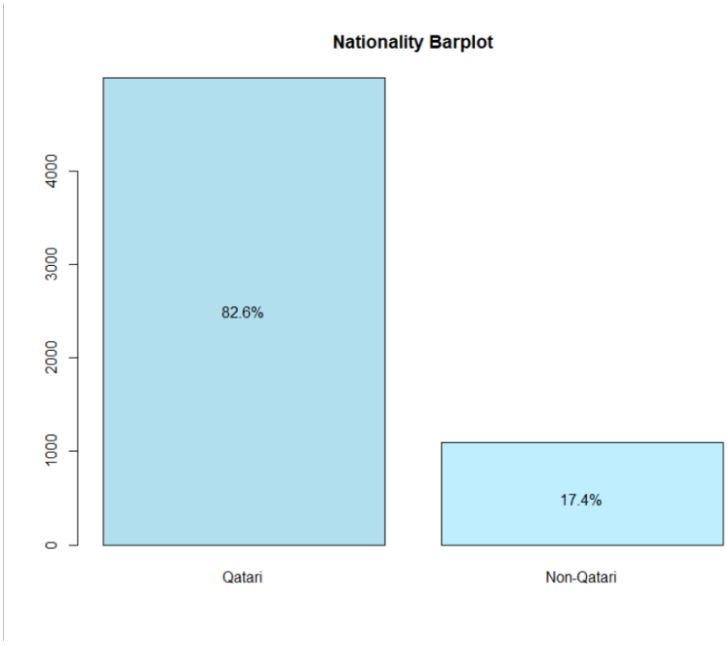

Figure 17 Nationality bar chart after balancing the data.

Age and BMI histogram looks quite similar to the results before balancing as well as the correlation plot, as can be seen in figures 16 and 17 below:

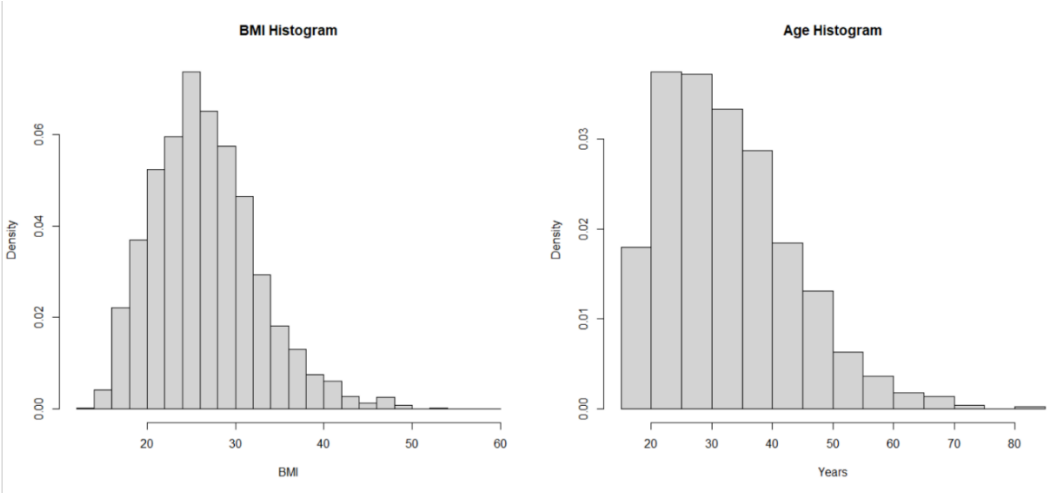

Figure 18 BMI and age histogram after balancing the data.

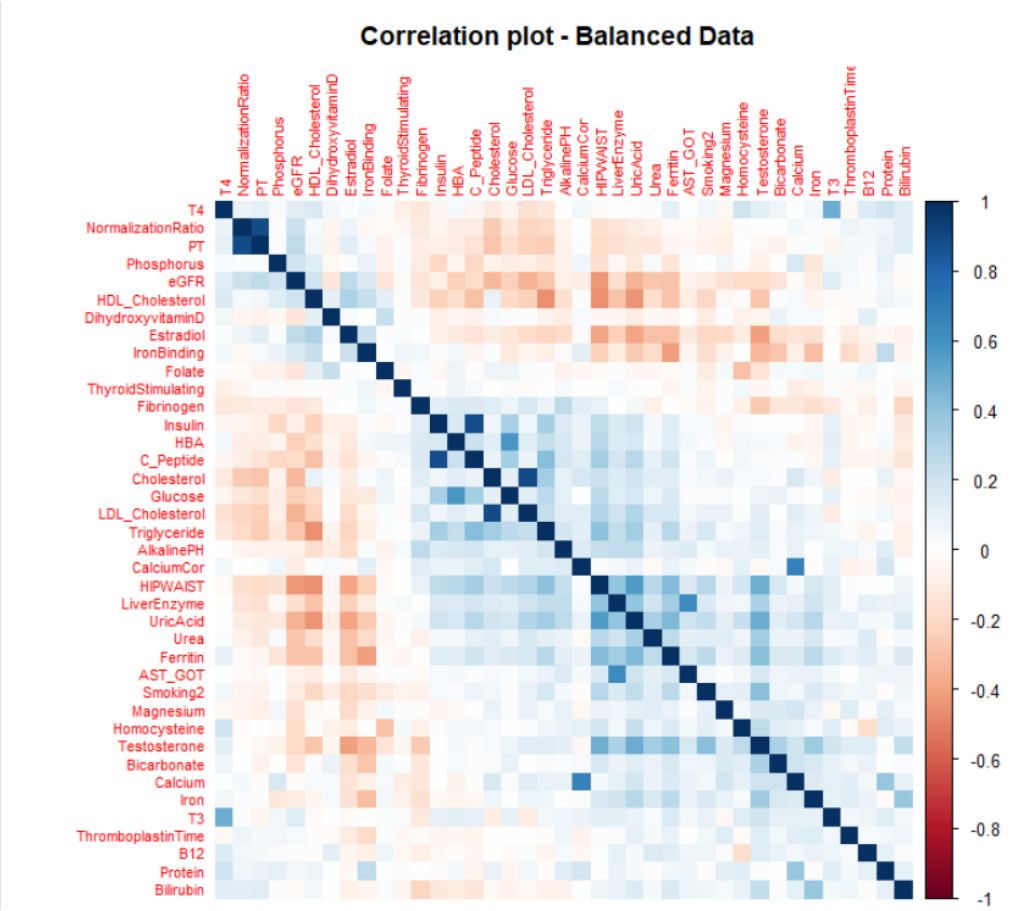

Figure 19 Correlation plot after balancing the data.

Table 21 below gives a detailed idea about the relationship between the gender and BMD levels, after balancing the data. Percentages between the brackets demonstrate the column percentages. We notice that the percentage of female with low BMD out of all females has risen remarkably.

Table 23

Relationship between Gender and BMD after the Balancing

		Gender	
		Female	Male
BMD Levels	Normal	1220 (34.0%)	1826 (73.0%)
	Low	2372 (66.0%)	674 (27.0%)

5.2 Machine Learning Implementation after Balancing

We implemented the machine learning algorithms on the modified dataset to train the model to classify the BMD values using all the variables. Below are the results.

5.2.1 Decision Tree

According to figure 18 below, Testosterone, BMI and eGFR are the most important variables in classifying the BMD variables, according to the decision tree. Decision tree model performance was tested on the test section of the data, and results are presented in table 22.

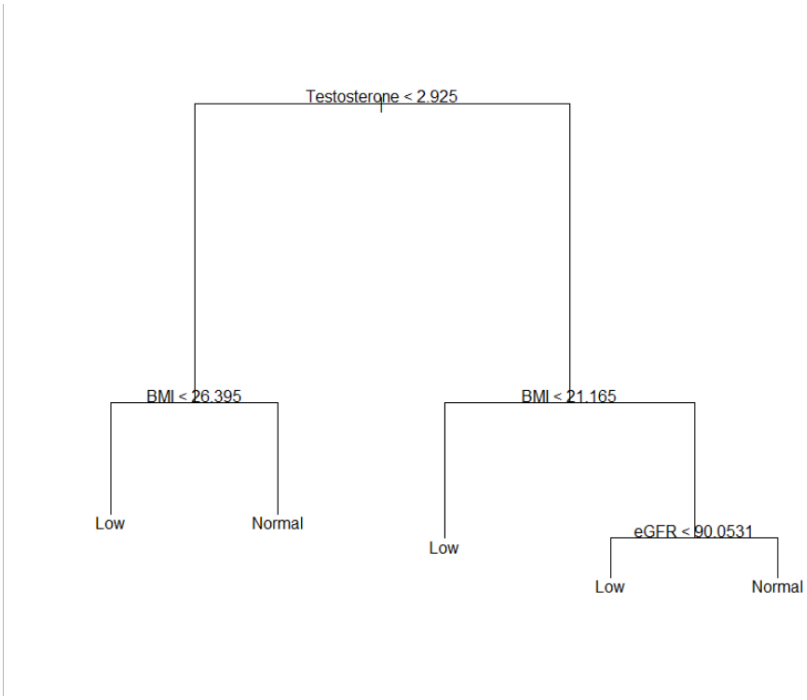

Figure 20 Pruned decision tree for classifying BMD levels after balancing.

Table 24

Decision Tree Performance Measures After Balancing

Accuracy	Sensitivity	Specificity	AUC
0.716	0.678	0.772	0.717

The performance differs totally after balancing the dataset. Accuracy is 71.6%, which means 71.6% of the predictions were correct. Sensitivity is 67.8%, which means 67.8% of the actual “Normal” BMD were correctly predicted to be “Normal”. The model has a specificity of 77.2%, which means 77.2% of the actual “Low” BMD cases were correctly predicted to be “Low”. Decision tree has an AUC of 0.717.

5.2.2 Random Forest

Random forest performance is outstanding. Accuracy is 96.6%, which means 96.6% of the predictions were correct. Sensitivity is 99.0%, which means 99.0% of the actual “Normal” BMD were correctly predicted to be “Normal”. The model has a specificity of 94.4%, which means 94.4% of the actual “Low” BMD cases were correctly predicted to be “Low”. RF has an AUC of 0.965 (Table 23).

Table 25

RF Performance Measures After Balancing

Accuracy	Sensitivity	Specificity	AUC
0.966	0.990	0.944	0.965

5.2.3 Linear Discriminant Analysis (LDA)

LDA performance wasn’t bad. Accuracy is 74.0%, which means 74.0% of the predictions were correct. Sensitivity is 74.7%, which means 74.7% of the actual “Normal” BMD were correctly predicted to be “Normal”. The model has a specificity of 73.4%, which means 73.4% of the actual “Low” BMD cases were correctly predicted to be “Low”. LDA has an AUC of 0.740 (Table 24).

Table 26

LDA Performance Measures After Balancing

Accuracy	Sensitivity	Specificity	AUC
0.740	0.747	0.734	0.740

5.2.4 Quadratic Discriminant Analysis (QDA)

Accuracy is 73.7%, which means 73.7% of the predictions were correct. Sensitivity is 81.9%, which means 81.9% of the actual “Normal” BMD were correctly predicted to be “Normal”. The model has a specificity of 68.9%, which means 68.9% of the actual “Low” BMD cases were correctly predicted to be “Low”. QDA has an AUC of 0.735 (Table 25).

Table 27

QDA Performance Measures After Balancing

Accuracy	Sensitivity	Specificity	AUC
0.737	0.819	0.689	0.735

5.2.5 K-Nearest Neighbors (KNN)

Accuracy is 76.0%, which means 76.0% of the predictions were correct. Sensitivity is 90.6%, which means 90.6% of the actual “Normal” BMD were correctly predicted to be “Normal”. The model has a specificity of 69.3%, which means 69.3% of the actual “Low” BMD cases were correctly predicted to be “Low”. KNN has an AUC of 0.758 (Table 26).

Table 28

KNN Performance Measures After Balancing

Accuracy	Sensitivity	Specificity	AUC
0.760	0.906	0.693	0.758

5.2.6 Logistic Regression

Accuracy is 73.3%, which means 73.3% of the predictions were correct. Sensitivity is 72.9%, which means 72.9% of the actual "Normal" BMD were correctly predicted to be "Normal". The model has a specificity of 73.6%, which means 73.6% of the actual "Low" BMD cases were correctly predicted to be "Low". Logistic regression has an AUC of 0.733 (Table 27).

Table 29

Logistic Regression Performance Measures After Balancing

Accuracy	Sensitivity	Specificity	AUC
0.733	0.729	0.736	0.733

5.2.7 Support Vector Machine (SVM)

SVM performed very well in classifying BMD level after balancing the data. Accuracy is 86.1%, which means 86.1% of the predictions were correct. Sensitivity is 91.3%, which means 91.3% of the actual "Normal" BMD were correctly predicted to be "Normal". The model has a specificity of 82.2%, which means 82.2% of the actual "Low" BMD cases were correctly predicted to be "Low". SVM has an AUC of 0.861 (Table 28).

Table 30

SVM Performance Measures After Balancing

Accuracy	Sensitivity	Specificity	AUC
0.861	0.913	0.822	0.861

5.2.8 Gradient Boosting

The gradient boosting model performed very well. Out of the eight algorithms, it is the second-best performing algorithm in classifying the BMD levels after the random forest. Accuracy is 92.2%, which means 92.2% of the predictions were correct. Sensitivity is 97.0%, which means 97.0% of the actual "Normal" BMD were correctly predicted to be "Normal". The model has a specificity of 88.3%, which means 88.3% of the actual "Low" BMD cases were correctly predicted to be "Low". Gradient boosting has an AUC of 0.928 (Table 29).

Table 31

Gradient Boosting Performance Measures After Balancing

Accuracy	Sensitivity	Specificity	AUC
0.922	0.970	0.883	0.928

5.3 Feature Selection after Balancing

In this part we will implement the machine learning algorithms to the balanced dataset, but with reduced number of variables. We will use random forest algorithm to find the most important variables, then we will train all the algorithms only using those selected variables. The above graph shows the most seven important features according to random forest. As expected, gender is and important factor in classifying BMD levels. We notice that all the seven variables are the same, except age, it is replaced with gender.

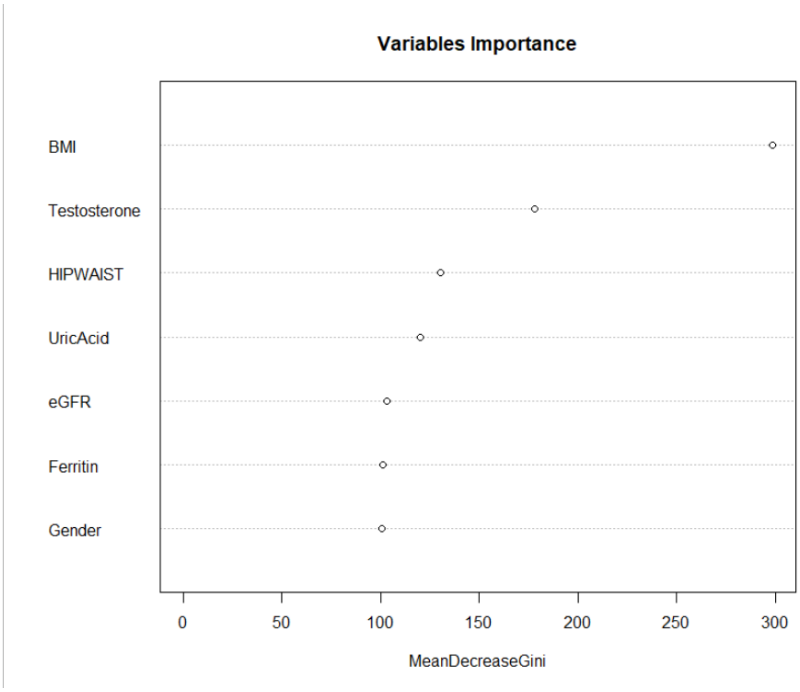

Figure 21 Top seven important variables by RF – balanced data.

In the coming section, we will demonstrate the results of the machine learning algorithms in classifying BMD levels using the balanced dataset with reduced set of variables.

5.3.1 Decision Tree

After feature selection, the results didn't change. Accuracy is 71.6%, which means 71.6% of the predictions were correct. Sensitivity is 67.8%, which means 67.8% of the actual "Normal" BMD were correctly predicted to be "Normal". The model has a specificity of 77.2%, which means 77.2% of the actual "Low" BMD cases were correctly predicted to be "Low". Decision tree has an AUC of 0.717 (Table 30).

Table 32

Decision Tree Performance Measures After Balancing

Accuracy	Sensitivity	Specificity	AUC

0.716	0.678	0.772	0.717

5.3.2 Random Forest

After feature selection, the results dropped slightly, but still vey outstanding. Accuracy is 95.4%, which means 98.9% of the predictions were correct. Sensitivity is 92.4%, which means 92.4% of the actual "Normal" BMD were correctly predicted to be "Normal". The model has a specificity of 92.4%, which means 92.4% of the actual "Low" BMD cases were correctly predicted to be "Low". Random forest has an AUC of 0.954 (Table 31).

Table 33

Random Forest Performance Measures After Balancing

Accuracy	Sensitivity	Specificity	AUC
0.954	0.989	0.924	0.954

5.3.3 Linear Discriminant Analysis (LDA)

After feature selection, the results dropped slightly. Accuracy is 71.7%, which means 71.7% of the predictions were correct. Sensitivity is 71.6%, which means 71.6% of the actual "Normal" BMD were correctly predicted to be "Normal". The model has a specificity of 72.0%, which means 72.0% of the actual "Low" BMD cases were correctly predicted to be "Low". LDA has an AUC of 0.718 (Table 32).

Table 34

LDA Performance Measures After Balancing

Accuracy	Sensitivity	Specificity	AUC

0.717	0.716	0.720	0.718

5.3.4 Quadratic Discriminant Analysis (QDA)

After feature selection, the results dropped slightly. Accuracy is 71.7%, which means 71.7% of the predictions were correct. Sensitivity is 71.6%, which means 71.6% of the actual "Normal" BMD were correctly predicted to be "Normal". The model has a specificity of 72.0%, which means 72.0% of the actual "Low" BMD cases were correctly predicted to be "Low". QDA has an AUC of 0.718 (Table 33).

Table 35

QDA Performance Measures After Balancing

Accuracy	Sensitivity	Specificity	AUC
0.717	0.716	0.720	0.718

5.3.5 K-Nearest Neighbors (KNN)

Accuracy is 78.0%, which means 78.0% of the predictions were correct. Sensitivity is 91.4%, which means 91.4% of the actual "Normal" BMD were correctly predicted to be "Normal". The model has a specificity of 71.3%, which means 71.3% of the actual "Low" BMD cases were correctly predicted to be "Low". KNN has an AUC of 0.779 (Table 34).

Table 36

KNN Performance Measures After Balancing

Accuracy	Sensitivity	Specificity	AUC
0.780	0.914	0.713	0.779

5.3.6 Logistic Regression

Accuracy is 70.3%, which means 70.3% of the predictions were correct. Sensitivity is 69.5%, which means 69.5% of the actual “Normal” BMD were correctly predicted to be “Normal”. The model has a specificity of 71.2%, which means 71.2% of the actual “Low” BMD cases were correctly predicted to be “Low”. Logistic regression has an AUC of 0.703 (Table 35).

Table 37

Logistic Regression Performance Measures After Balancing

Accuracy	Sensitivity	Specificity	AUC
0.703	0.695	0.712	0.703

5.3.7 Support Vector Machine (SVM)

SVM performance dropped significantly after reducing the number of variables. Accuracy is 75.1%, which means 75.1% of the predictions were correct. Sensitivity is 78.4%, which means 78.4% of the actual “Normal” BMD were correctly predicted to be “Normal”. The model has a specificity of 72.5%, which means 72.5% of the actual “Low” BMD cases were correctly predicted to be “Low”. SVM has an AUC of 0.750 (Table 36).

Table 38

SVM Performance Measures After Balancing

Accuracy	Sensitivity	Specificity	AUC
0.751	0.784	0.725	0.750

5.3.8 Gradient Boosting

Gradient boosting kept the same level of performance after reducing the number of variables. Accuracy is 89.2%, which means 89.2% of the predictions were correct. Sensitivity is 94.7%, which means 94.7% of the actual “Normal” BMD were correctly predicted to be “Normal”. The model has a specificity of 84.9%, which means 84.9% of the actual “Low” BMD cases were correctly predicted to be “Low”. Gradient Boosting has an AUC of 0.891 (Table 37).

Table 39

Gradient Boosting Performance Measures After Balancing

Accuracy	Sensitivity	Specificity	AUC
0.892	0.947	0.849	0.891

5.4 Comparison

Figure 20 shows a comparison between all models used in all sections. Balancing the Data has a great effect on the performance of the model. Random forest reached an excellent performance when we used the balanced data, all its performance measures are above 90%.

When applying feature selection, the results dropped slightly, but not significantly. Thus, we decided to go with reduced models, based on the parsimony rule.

The best algorithms in classifying the machine BMD levels are: Random Forest, Gradient Boosting and SVM respectively. While the most important variables are: “BMI”, “Testosterone”, “Hip-Waist ratio”, “Uric Acid”, “eGFR”, “Ferritin”, “Gender” and “Age”.

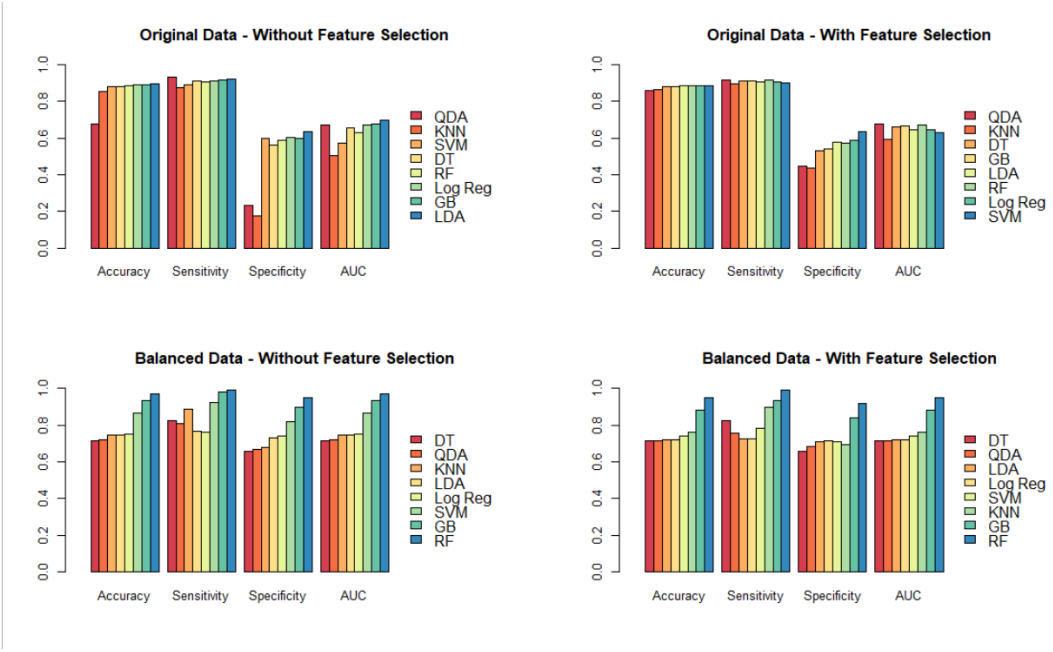

Figure 22 Comparison between ML used to classify BMD levels in all section.

CHAPTER 6: CONCLUSION

Osteoporosis is a skeletal disease characterized by the reduction in the mass of the bones and leads to more weak and fragile bones. It affects millions of people around the world. Osteoporosis eventually leads to fracture which severely affects patient's quality of life and in some cases leads to death. Moreover, osteoporosis could have a huge impact on the economic due to the long hospital stay that patients' need to heel from an osteoporotic fracture. Hence, diagnosing the disease early is crucial to the disease management care, which usually not feasible, since the disease symptoms' stay hidden until bones break. For this reason, osteoporosis is called "Silent Disease".

Osteoporosis is diagnosed with Bone Mineral Density (BMD), which is the amount of mineral in our bones consists of. BMD could be measured using many technologies; however, Dual X-ray Absorptiometry (DXA) is considered to be the "Gold Standard" in this field. DXA is an X-ray machine measures the amount of reduction in x-ray beams that passing through a certain space. In practice, the DXA output is not useful solely, thus it is transferred to *T-score* which is the patient's BMD compared to the BMD of a healthy adult.

BMD is categorized into four groups, which are: Normal (T-score > -1), Osteopenia (-1 > T-score > -2.5), Osteoporosis (-2.5 > T-score) and Severe Osteoporosis (-2.5 > T-score, plus an existing fracture). Osteopenia is a stage before osteoporosis and also known as *low BMD*.

This research aimed to apply machine learning algorithms to classify the BMD into Normal or Low based on Qatar Biobank data, and to identify the most accurate and precise algorithm in classifying the BMD levels, plus highlighting which variables has the highest influence on the BMD levels. Furthermore, to calculate the mean and standard deviation of BMD values in Qatar.

The reference mean of BMD and standard deviation were found to be 1.22 and 0.124 respectively. Explanatory analysis showed that the proportion observation with low BMD out of the total observations was quite small. Hence, two approaches were followed. First, machine learning algorithms were implemented on the raw data. Second, the algorithms were implemented on a balanced data were the proportion of people with low BMD was exactly 50%. In phase one analysis, the algorithms performed poorly due to the unbalanced structure of the data. On the other hand, in the second phase the results were impressive. The best performing algorithm was Random Forest (RF) with AUC of 95.4%, followed by Gradient Boosting with AUC of 89.1%. The variables with the highest influence on BMD levels were found to be "BMI", "Testosterone", "Hip-Waist ratio", "Uric Acid", "eGFR", "Ferritin", "Gender" and "Age".

Machine learning is a very hot topic with a lot of potentials. This research showed that artificial intelligence and machine learning approaches could make a revolution in bone health field by providing a very accurate information about people's bone with the minimal amount of money. The output of this research could be used by Primary Health Care Institution or any other health institution as an early detection tool for low BMD issues,

In the future, this research could be extended in several ways. Firstly, to test the effectiveness of the deep learning model such as neural networks in classifying BMD levels. Furthermore, embedding the feature selection process with each technique might enhance the algorithms performance. Lastly, use and compare different balancing schemes such as downsampling. The limitation of this study is data collection phase, a more concrete design could be implemented in the future.

REFERENCES

Abdulla, N., Alsaed, O. S., Lutf, A., Alam, F., Abdulmomen, I., Al Emadi, S. & Johansson, H. (2022). Epidemiology of hip fracture in Qatar and development of a country specific FRAX model. Archives of Osteoporosis, 17(1), 1-6.

Bone density test - Mayo Clinic. Mayoclinic.org. (2022). Retrieved 11 September 2022, from https://www.mayoclinic.org/tests-procedures/bone-density-test/about/pac-20385273.

Chen, Y., Yang, T., Gao, X., & Xu, A. (2022). Hybrid deep learning model for risk prediction of fracture in patients with diabetes and osteoporosis. *Frontiers of Medicine*, *16*(3), 496–506. https://doi.org/10.1007/s11684-021-0828-7

Cisco Secure Endpoint. (n.d.). Retrieved October 25, 2022, from https://www.cisco.com/c/dam/m/cs_cz/training-events/webinars/tech-club-webinars/secure-endpoint.pdf

Encyclopedia, M., & test, B. (2022). *Bone mineral density test: MedlinePlus Medical Encyclopedia*. Medlineplus.gov. Retrieved 11 September 2022, from https://medlineplus.gov/ency/article/007197.htm

Erjiang, E., Wang, T., Yang, L., Dempsey, M., Brennan, A., Yu, M., ... & Carey, J. J. (2021). Machine learning can improve clinical detection of low BMD: the DXA-HIP study. Journal of Clinical Densitometry, 24(4), 527-537.

Galassi, A., Martín-Guerrero, J. D., Villamor, E., Monserrat, C., & Rupérez, M. J. (2020). Risk Assessment of Hip Fracture Based on Machine Learning. *Applied Bionics and Biomechanics*, *2020*, 1–13. https://doi.org/10.1155/2020/8880786

Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). *An introduction to statistical learning: with applications in R*. Spinger.

Ibrahim, W. N., Younes, N., Shi, Z., & Abu-Madi, M. A. (2021). Serum uric acid level is positively associated with higher bone mineral density at multiple skeletal sites among healthy qataris. Frontiers in Endocrinology, 12, 653685.

Juan, Y. C., Chen, C. M., & Chen, S. H. (2015). A classifier fusion approach to osteoporosis prediction for women in Taiwan. Journal of Industrial and Production Engineering, 32(6), 360-368.

Kanis, J. A., McCloskey, E. V., Johansson, H., Cooper, C., Rizzoli, R., & Reginster, J. Y. (2013). European guidance for the diagnosis and management of osteoporosis in postmenopausal women. Osteoporosis international, 24(1), 23-57.

Kerkadi, A., Lathief, S., Khial, Y., Teleb, T., Attieh, G., Rahman, M. M., ... & Agouni, A. (2022). The relationship between bone mineral density and body composition among Qatari women with high rate of obesity: qatar Biobank Data. Frontiers in Nutrition, 9.

Khondaker, M. T. I., Khan, J. Y., Refaee, M. A., Hajj, N. E., Rahman, M. S., & Alam, T. (2020). Obesity in Qatar: a case-control study on the identification of associated risk factors. Diagnostics, 10(11), 883.

Kruse, C., Eiken, P., & Vestergaard, P. (2017). Machine learning principles can improve hip fracture prediction. Calcified tissue international, 100(4), 348-360.

Lgayhardt. (n.d.). Deep Learning vs. Machine Learning - Azure Machine Learning. Retrieved October 25, 2022, from https://learn.microsoft.com/en-us/azure/machine-learning/concept-deep-learning-vs-machine-learning

Ij, H. (2018). Statistics versus machine learning. Nat Methods, 15(4), 233.

Mehta, S. D., & Sebro, R. (2020). Computer-Aided Detection of Incidental Lumbar Spine Fractures from Routine Dual-Energy X-Ray Absorptiometry (DEXA) Studies Using a Support Vector Machine (SVM) Classifier. *Journal of Digital Imaging*, *33*(1), 204–210. https://doi.org/10.1007/s10278-019-00224-0

Nam, K. H., Seo, I., Kim, D. H., Lee, J. il, Choi, B. K., & Han, I. H. (2019). Machine Learning Model to Predict Osteoporotic Spine with Hounsfield Units on Lumbar Computed Tomography. *Journal of Korean Neurosurgical Society*, *62*(4), 442–449. https://doi.org/10.3340/jkns.2018.0178

Park, H. W., Jung, H., Back, K. Y., Choi, H. J., Ryu, K. S., Cha, H. S., Lee, E. K., Hong, A. R., & Hwangbo, Y. (2021). Application of Machine Learning to Identify Clinically Meaningful Risk Group for Osteoporosis in Individuals Under the Recommended Age for Dual-Energy X-Ray Absorptiometry. *Calcified Tissue International*, *109*(6), 645–655. https://doi.org/10.1007/s00223-021-00880-x

Shim, J.-G., Kim, D. W., Ryu, K.-H., Cho, E.-A., Ahn, J.-H., Kim, J.-I., & Lee, S. H. (2020). Application of machine learning approaches for osteoporosis risk prediction in postmenopausal women. *Archives of Osteoporosis*, *15*(1), 169. https://doi.org/10.1007/s11657-020-00802-8

WHO Scientific Group on Prevention, Management of Osteoporosis, & World Health Organization. (2003). Prevention and management of osteoporosis: report of a WHO scientific group (No. 921). World Health Organization.

Yoo, T. K., Kim, S. K., Kim, D. W., Choi, J. Y., Lee, W. H., Oh, E., & Park, E.-C.

(2013). Osteoporosis Risk Prediction for Bone Mineral Density Assessment of Postmenopausal Women Using Machine Learning. *Yonsei Medical Journal*, *54*(6), 1321. https://doi.org/10.3349/ymj.2013.54.6.1321

Yoshimura, M., Moriwaki, K., Noto, S., & Takiguchi, T. (2017). A model-based cost-effectiveness analysis of osteoporosis screening and treatment strategy for postmenopausal Japanese women. Osteoporosis International, 28(2), 643-652

APPENDIX

Appendix A1: Confusion Matrices for the Eight Machine Learning Techniques Used (Original Data – Without Feature Selection)

			Predicted	
			Normal	Low
Decision Tree	Actual	Normal	592	72
		Low	28	33
Random Forest	Actual	Normal	613	84
		Low	7	21
LDA	Actual	Normal	605	77
		Low	15	33
QDA	Actual	Normal	447	37
		Low	137	68
KNN	Actual	Normal	608	102
		Low	12	3
Logistic Regression	Actual	Normal	597	73
		Low	23	32
SVM	Actual	Normal	617	88
		Low	3	17
Gradient Boosting	Actual	Normal	599	78
		Low	21	27

Appendix A2: Confusion Matrices for the Eight Machine Learning Techniques Used (Original Data – With Feature Selection)

			Predicted	
			Normal	Low
Decision Tree	Actual	Normal	592	72
		Low	28	33
Random Forest	Actual	Normal	588	79
		Low	32	26
LDA	Actual	Normal	604	82
		Low	16	23
QDA	Actual	Normal	588	71
		Low	32	34
KNN	Actual	Normal	592	81
		Low	28	24
Logistic Regression	Actual	Normal	605	82
		Low	15	23
SVM	Actual	Normal	616	86
		Low	4	19
Gradient Boosting	Actual	Normal	603	77
		Low	17	28

Appendix A3: Confusion Matrices for the Eight Machine Learning Techniques Used (Balanced Data – Without Feature Selection)

			Predicted	
			Normal	Low
Decision Tree	Actual	Normal	491	233
		Low	113	382
Random Forest	Actual	Normal	568	6
		Low	36	609
LDA	Actual	Normal	434	147
		Low	170	468
QDA	Actual	Normal	363	80
		Low	241	535
KNN	Actual	Normal	363	80
		Low	241	535
Logistic Regression	Actual	Normal	442	164
		Low	162	451
SVM	Actual	Normal	481	46
		Low	123	569
Gradient Boosting	Actual	Normal	525	16
		Low	79	599

Appendix A4: Confusion Matrices for the Eight Machine Learning Techniques Used (Balanced Data – With Feature Selection)

			Predicted	
			Normal	Low
Decision Tree	Actual	Normal	491	233
		Low	113	382
Random Forest	Actual	Normal	554	6
		Low	50	609
LDA	Actual	Normal	431	171
		Low	173	444
QDA	Actual	Normal	398	150
		Low	206	465
KNN	Actual	Normal	371	35
		Low	233	580
Logistic Regression	Actual	Normal	432	190
		Low	172	425
SVM	Actual	Normal	414	114
		Low	190	501
Gradient Boosting	Actual	Normal	500	28
		Low	104	587