ANALYZING THE BONE MINERAL DENSITY FOR OSTEOPOROSIS BASED ON GENERAL LABORATORY DATA AND EXAMINATIONS IN UNITED STATES POPULATION USING BIG DATA PROCESSING AND PREDICTIVE ANALYTICS

A Thesis

Presented to

The Faculty of the Department of Computer Science

California State University, Los Angeles

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

in

Computer Science

By

Safa Al Mahbub

August 2019

The thesis of Safa Al Mahbub is approved.

Dr. Mohammad Pourhomayoun, Committee Chair

Dr. Behzad Parviz

Dr. Raj Pamula, Department Chair

California State University, Los Angeles

August 2019

ABSTRACT

Analyzing The Bone Mineral Density For Osteoporosis Based On General Laboratory Data And Examinations In United States Population Using Big Data Processing And Predictive Analytics

By

Safa Al Mahbub

The inspiration behind the project is to use data science (DS) techniques and machine learning algorithms (MLA) to find trends in patients with Osteoporosis through easy, self-administered, and remote tests.

The standard approach to determining Osteoporosis is to measure a patient's Bone Mineral Density (BMD) level. This is done generally with a specialized machine that takes a dual-energy X-ray absorptiometry (DXA). Once a reading of a patient's BMD is taken, it is then compared to the average BMD of the patient's age group. If the BMD is 2.5 standard deviations below the average, then Osteoporosis is diagnosed.

A DXA machine takes up a lot of space and costs a couple of hundred dollars to get a scan. Along with finding trends in patients with Osteoporosis, another goal is to see if perhaps there is an accurate way to predict BMD without such a large machine that can be used in remote locations that don't have accesses to the machine.

ACKNOWLEDGMENTS

I would like to thank the NASA DIRECT-STEM program at California State University, Los Angeles for guiding me in the research process. I would also like to thank Kae Swada and Dr.Clark for helping us understand the effects of Osteoporosis and how we can all benefit from studying the topic.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1

Data Accumulation

1.1 Gathering the Datasets

The dataset that we used in this study comes from the National Health and Nutrition Examination Survey (NHANES) from 2007-2008 from the Inter-university Consortium for Political and Social Research (ICPSR) website (United States Department of Health and Human Services). The NHAHES data had a variety of different datasets, roughly broken into three categories: laboratory results, examination, and questionnaire. The inspiration behind this project was to determine BMD and Osteoporosis through very basic medical tests that could be done in a closed environment. For that reason, some of the data that was used was easily accessible data such as blood pressure, a standard blood test, a standard biochemical reading, etc.

With that in mind, there were some datasets that had to be excluded in the project due to a number of reasons. The data from NHAHES was gathered from the general public, and as such some laboratory test, examinations, and questionnaires were administered to some patients while others were omitted.

1.2 Data Preprocessing

The data was written as tab separated values (TSV) files, each with a linking - respondent sequence number (SEQN) id column. Using Python's TSV-to-dataframe function, the program easily transformed the TSV files into separate dataframe objects. However, the built-in function was not able to transform all of the data correctly. Most of the data that is listed are numeric values but were converted as strings. The data had to be transformed string to numeric, which required also getting rid of values with white spaces or not-a-number values.

A few of the features that were in the individual TSV files had some repetitive data that occurred in one of two ways. The first method is having the same data listed in different units of measurements. For these values, the unit of measurement that is more commonly used as the standard was kept while the duplicated data was ignored. The second type of repetitive data that occurred was having the same data be listed in different datasets under different names. For these values, (i.e. total cholesterol occurred in both "DS116 Laboratory Standard Biochemistry Profile" dataset as well as the "DS118 Laboratory Total Cholesterol" dataset) only one value was kept.

Once all the data was formatted properly and duplicate data was removed, the data was merged together using the SEQN number.

The data was heavily skewed towards people without osteoporosis. To even the data, we had to make a 50/50 split of people who had osteoporosis and those that did not.

1.3 Feature Selection through Correlation values

The resulting size of the dataframe was around 280 by 72, meaning our data included 72 feature elements for around 140 subjects with osteoporosis and around 140 without. To reduce the computational complexity of the model and also to eliminate redundant or useless features, a number of feature selection methods were applied. The most effective feature selection method was Correlation based supervised feature selection. Corr function in Python produces a separate dataframe where it lists the correlation of one feature to another. Using these with the target variable, we sorted the features and selected the best 20 of them to be used in future analytics models. Using 0.2 as the minimum level of noticeable correlation in this project, the following correlation coefficients were found between Total Femur BMD and other features.

Table 1 Correlation Coefficients Between Femur Bone Mineral Density and Data

DXXOFBMD : Total Femur BMD	
Gender	0.549605
Weight	0.503425
Age	0.466415
Height	0.457911
Arm Circumference	0.422449
Upper Leg Length	0.418028
Upper Arm Length	0.410224
Average Systolic Blood Pressure	0.297350
Red Blood Cell Count	0.286580
Creatinine found in Urine	0.279079
Body Mass Index	0.273511
Percentage of Hematocrit Blood	0.272137
Waist Circumference	0.247999
Subscapular Skinfold	0.207795

CHAPTER 2

Data Visualization for Trends in Bone Mineral Density

Using the features determined based on the correlation, we used the data visualization Python library, Seaborn, to see what relationship, or correlation, the features have with BMD. Following this was a cycle of editing the data to get rid of outliers to produce the best data visualization of said correlation. The following are the visualization trends that the results produced.

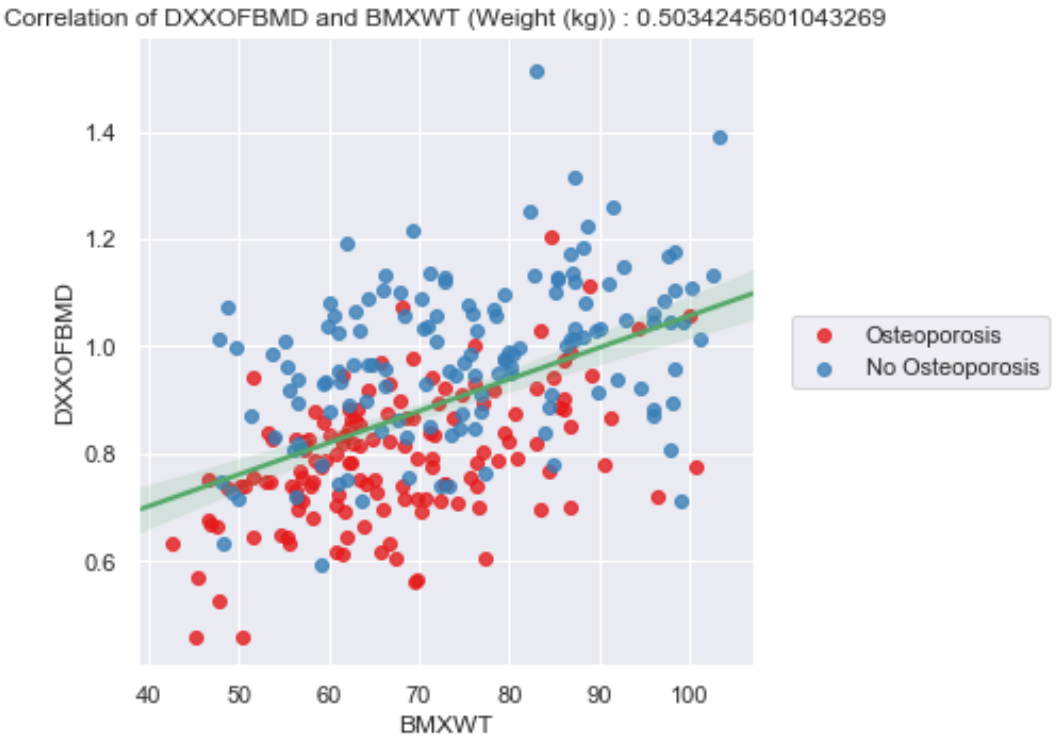

Figure 1. Correlation of BMD and Weight

From this data visualization, it appears that the less someone weighs, the lower their BMD is, which makes them more prone to Osteoporosis.

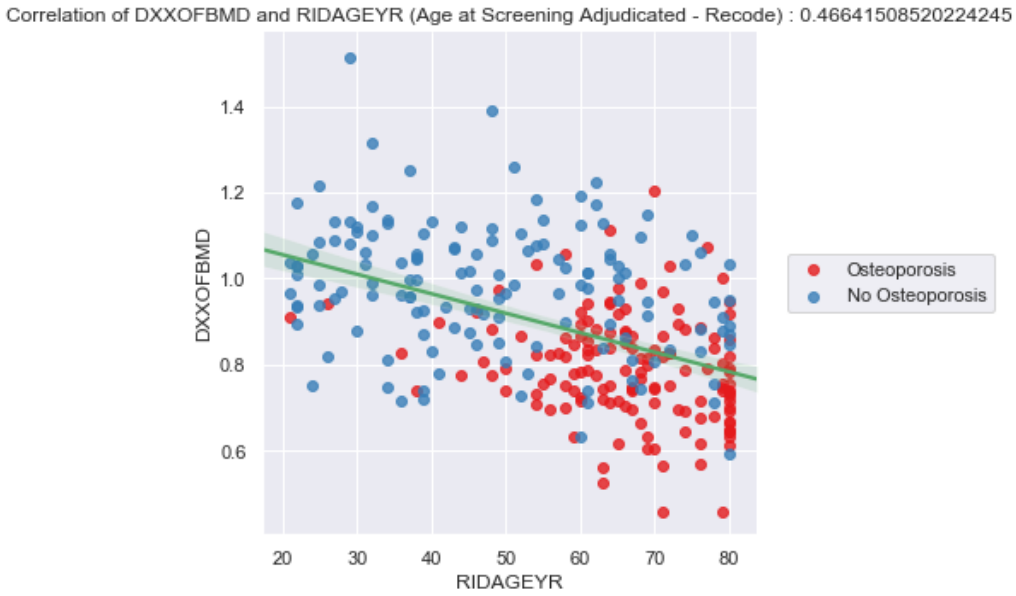

Figure 2. Correlation of BMD and Age

From this data visualization, it appears that the older someone is, the lower their BMD is, which makes them more prone to Osteoporosis.

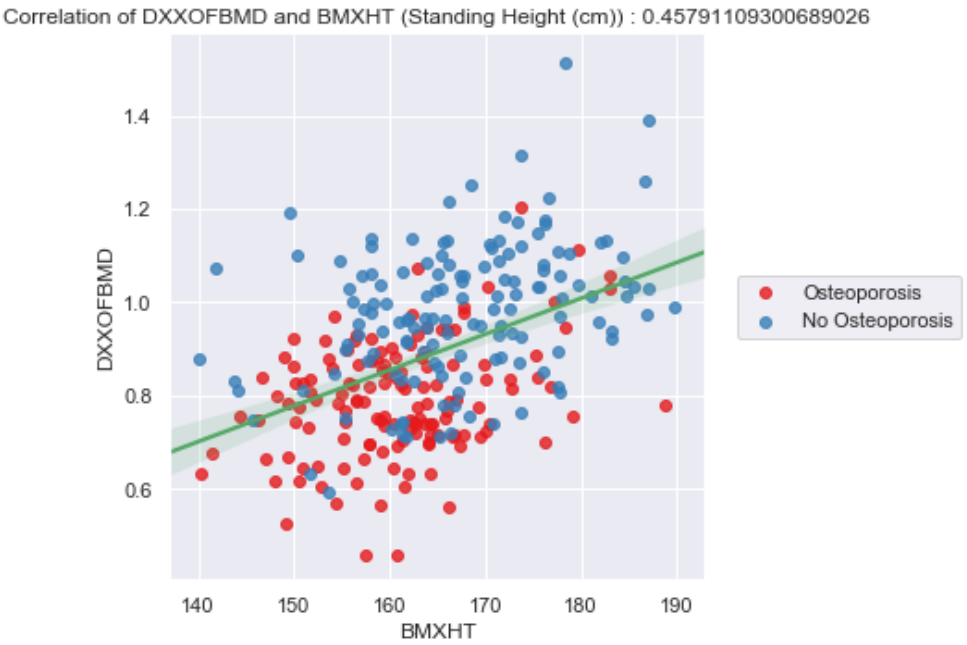

Figure 3. Correlation of BMD and Height

From this data visualization, it appears that the lower someone's height is, the lower their BMD is, which makes them more prone to Osteoporosis.

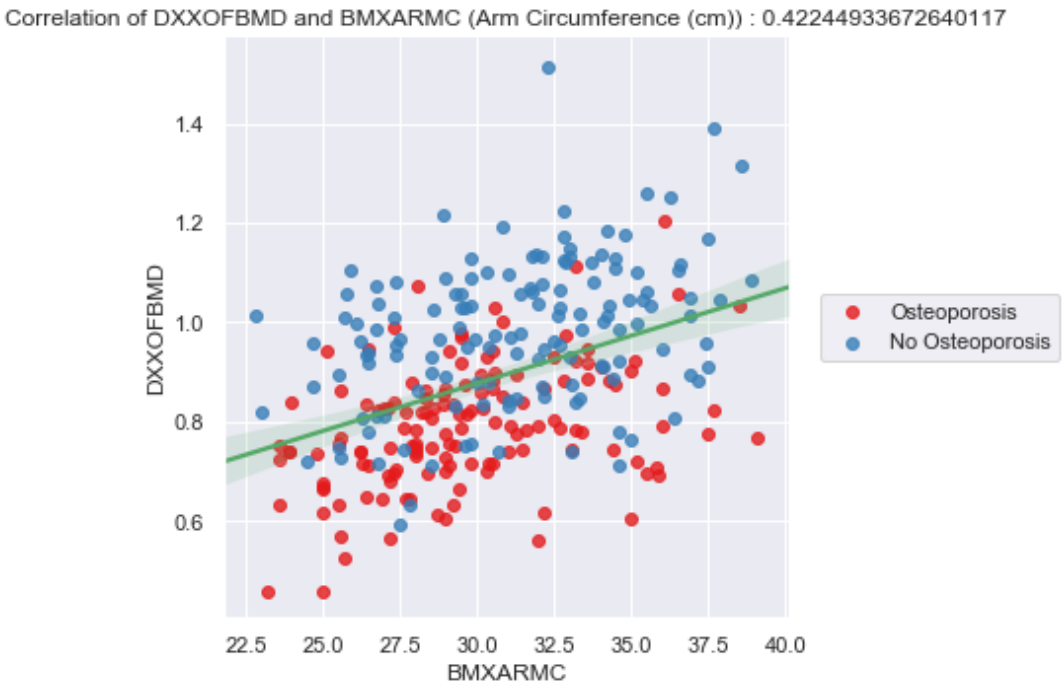

Figure 4. Correlation of BMD and Arm Circumference

From this data visualization, it appears that the smaller someone's arm circumference is, the lower their BMD is, which makes them more prone to Osteoporosis.

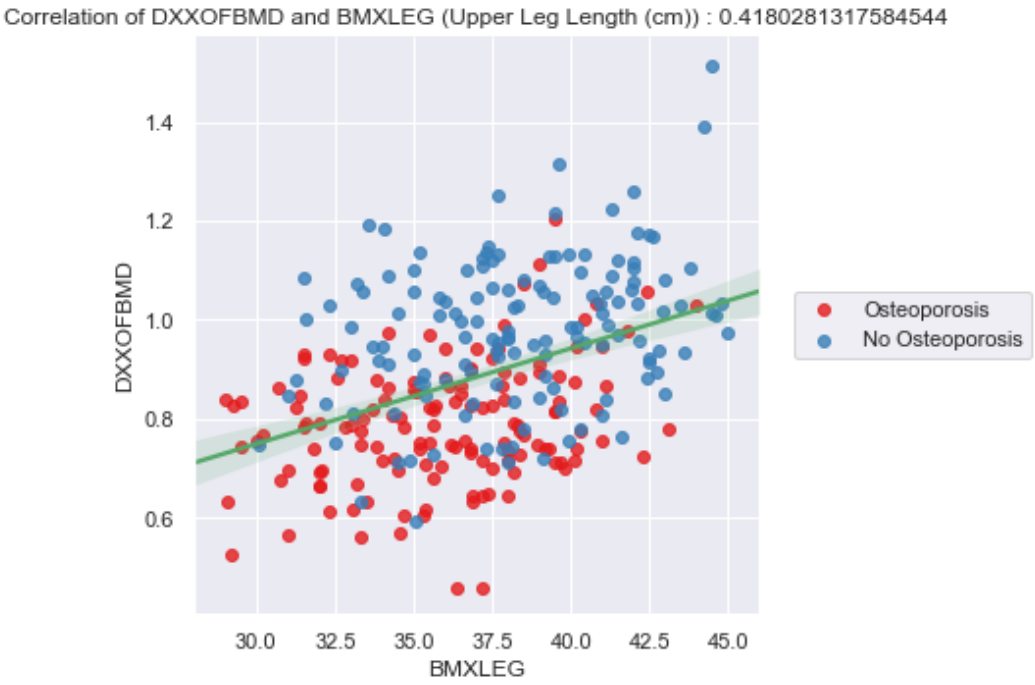

Figure 5. Correlation of BMD and Upper Leg Length

From this data visualization, it appears that the less someone's upper leg length, the lower their BMD is, which makes them more prone to Osteoporosis.

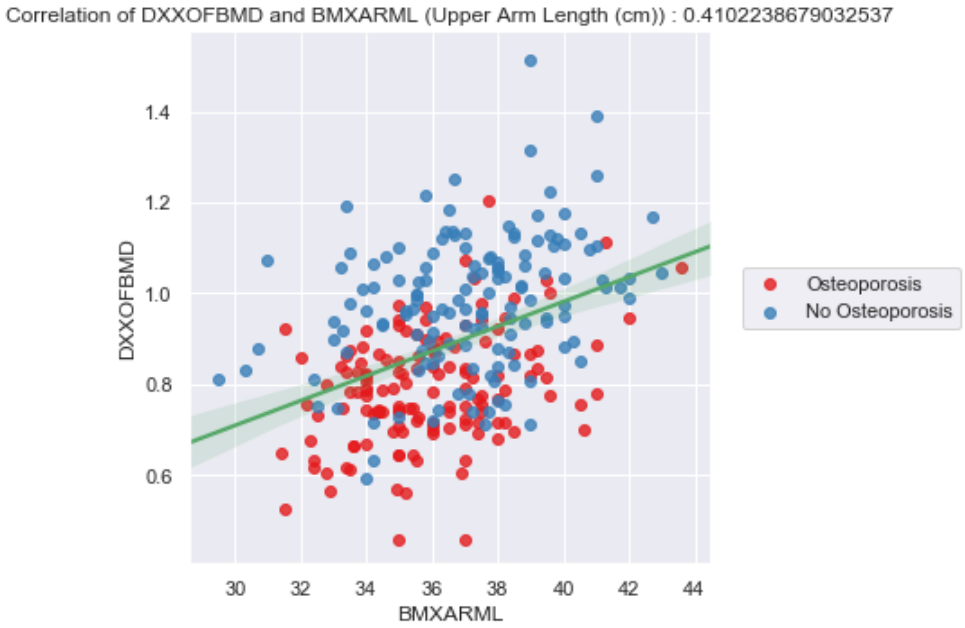

Figure 6. Correlation of BMD and Upper Arm Length

From this data visualization, it appears that the less someone's upper arm length, the lower their BMD is, which makes them more prone to Osteoporosis.

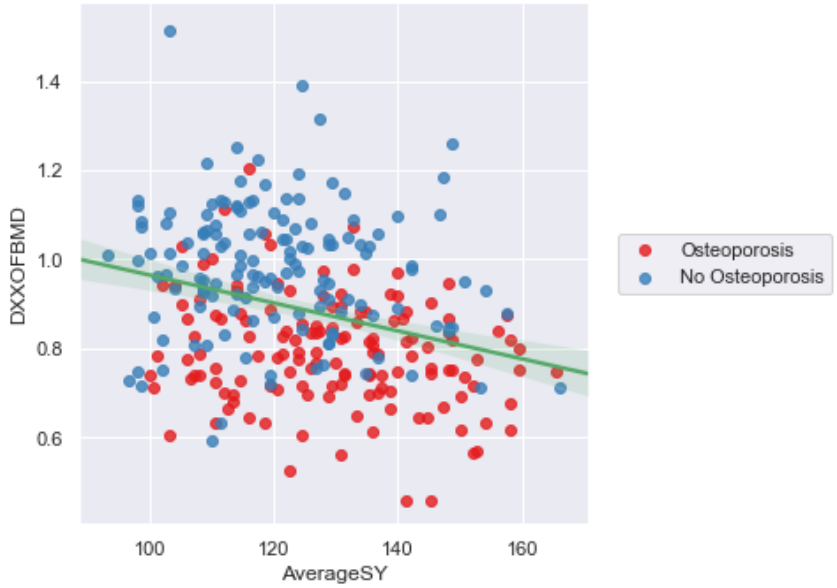

Figure 7. Correlation of BMD and Average Systolic Blood Pressure

From this data visualization, it appears that the higher someone's systolic blood pressure is, the lower their BMD is, which makes them more prone to Osteoporosis.

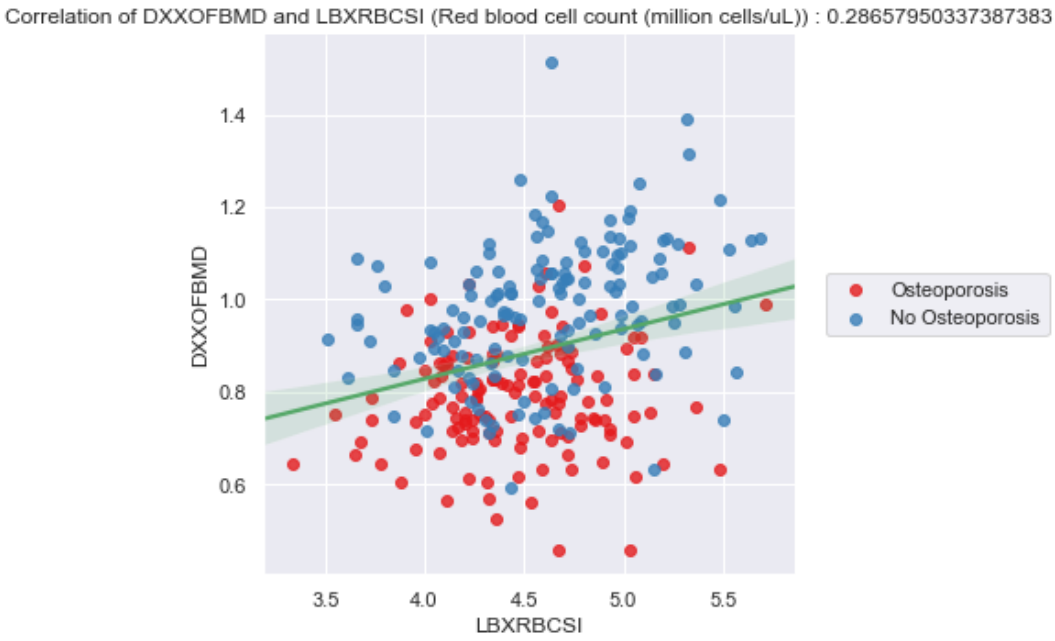

Figure 8. Correlation of BMD and Red Blood Cell Count

From this data visualization, it appears that if your red blood cell count is near 4.5 million, the lower their BMD is, which makes them more prone to Osteoporosis.

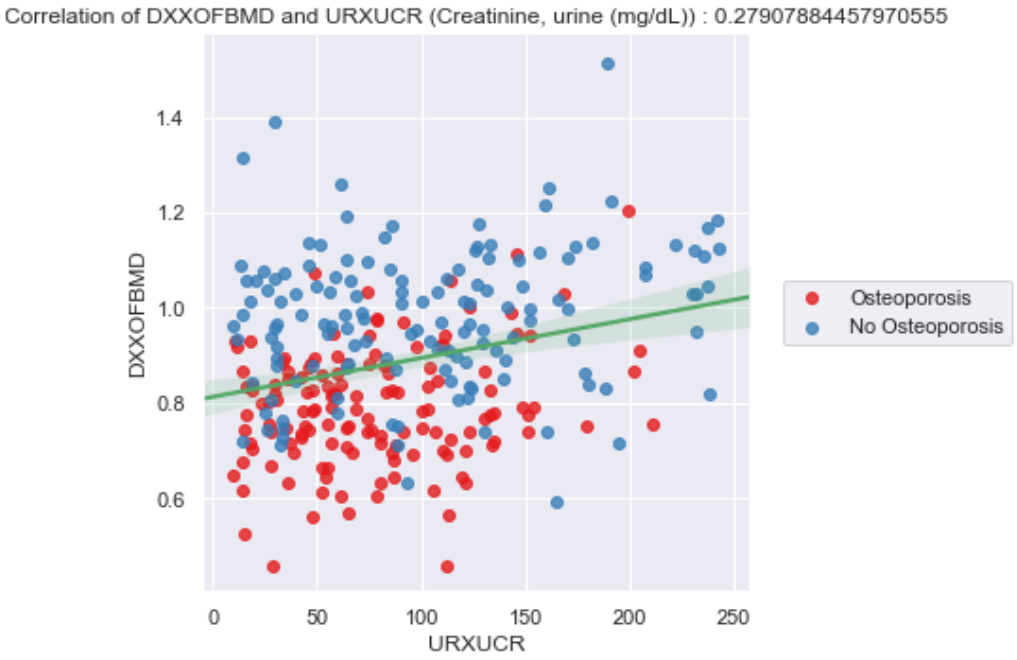

Figure 9. Correlation of BMD and refatinine found in Urine

From this data visualization, it appears that the less creatinine someone has in their urine test, the lower their BMD is, which makes them more prone to Osteoporosis.

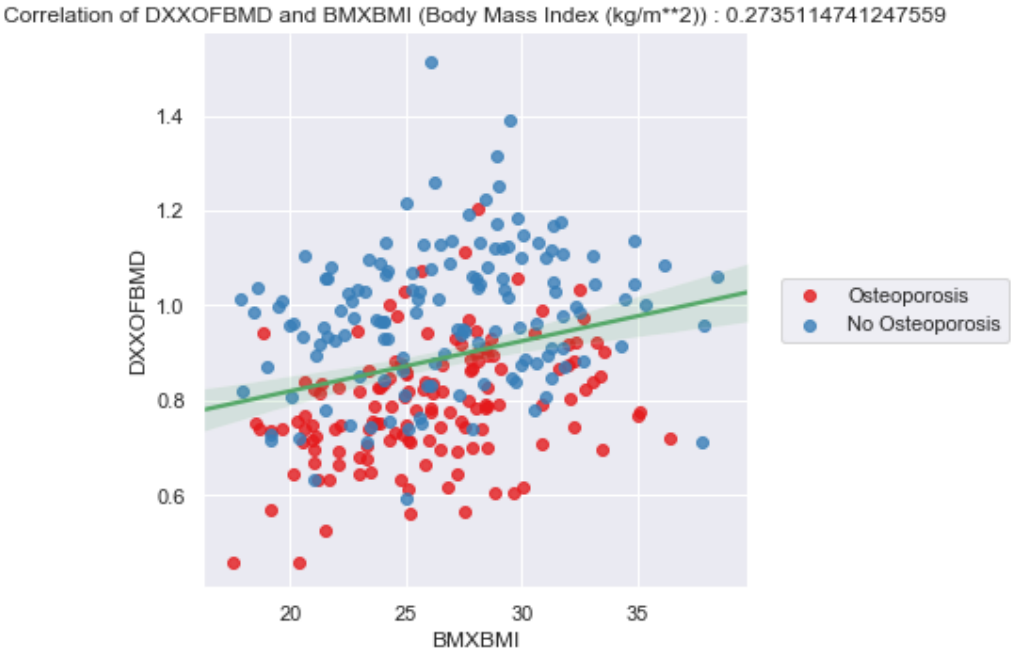

Figure 10. Correlation of BMD and Body Mass Index

From this data visualization, it appears that the less someone's body mass index is, the lower their BMD is, which makes them more prone to Osteoporosis.

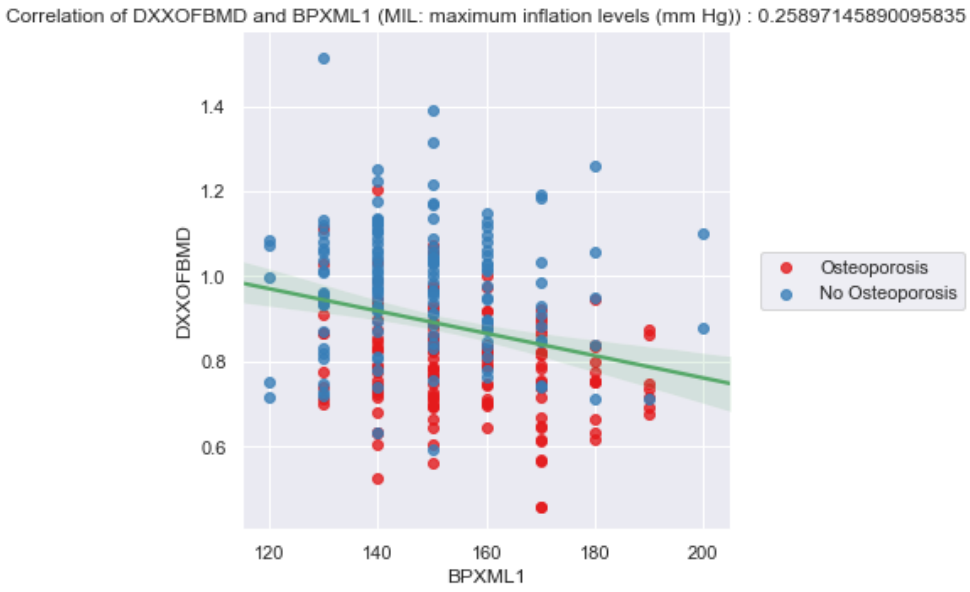

Figure 11. Correlation of BMD and Maximum inflaion levels

From this data visualization, it appears that the higher someone's maximum inflation level, the lower their BMD is, which makes them more prone to Osteoporosis.

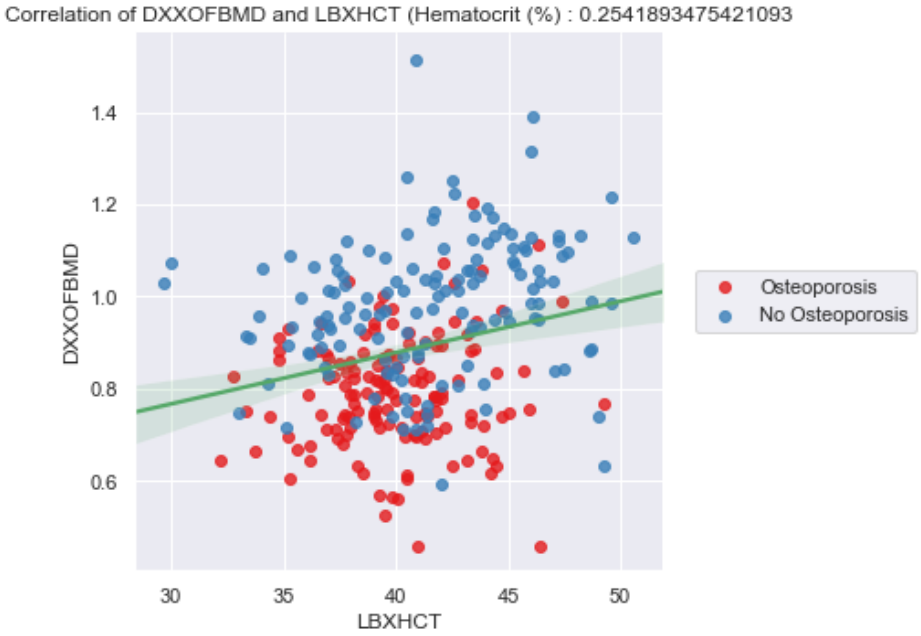

Figure 12. Correlation of BMD and Percentage of Hematocrit Blood

From this data visualization, it appears that the lower someone's hematocrit blood count percentage is, the lower their BMD is, which makes them more prone to Osteoporosis.

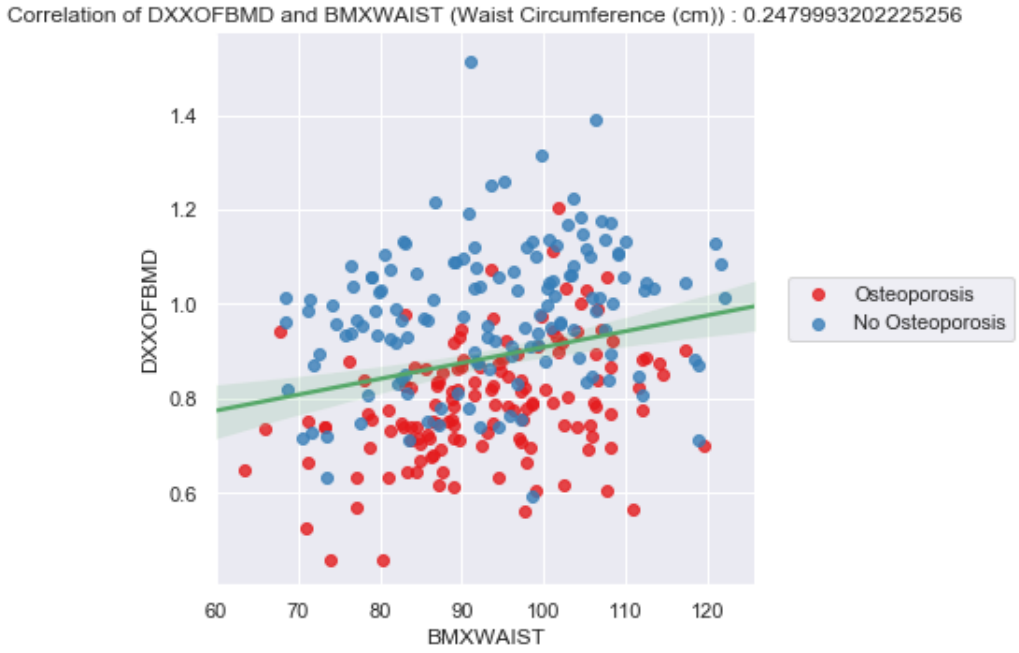

Figure 13. Correlation of BMD and Waist Circumference

From this data visualization, it appears that the lower someone's waist circumference is, the lower their BMD is, which makes them more prone to Osteoporosis.

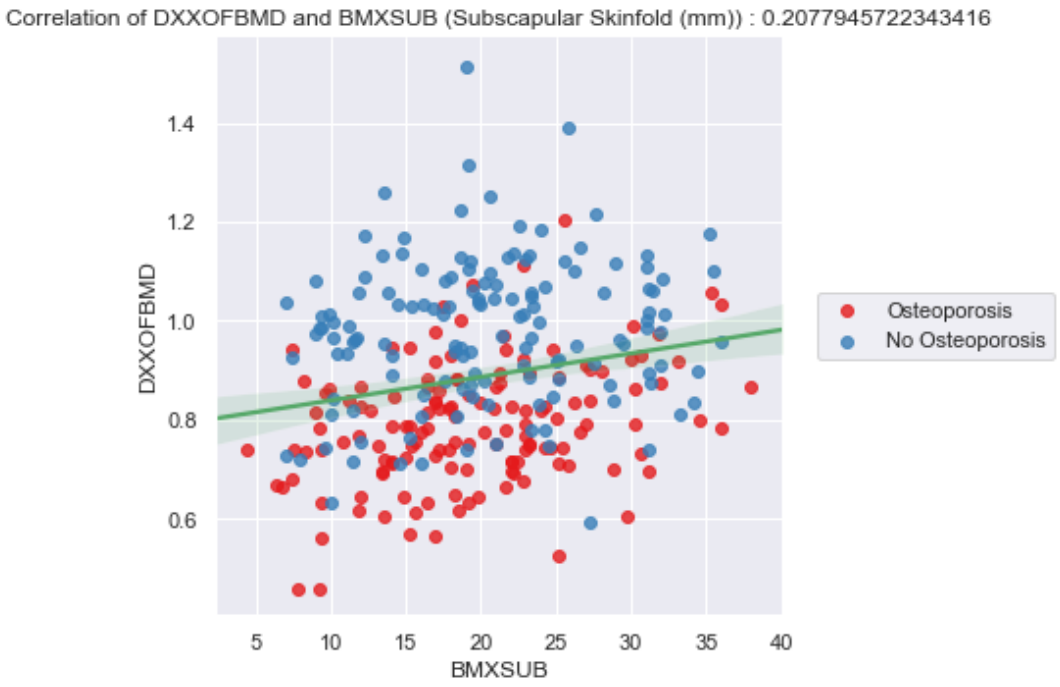

Figure 14. Correlation of BMD and Subscapular Skinfold

From this data visualization, it appears that the lower someone's subscapular skinfold is, the lower their BMD is, which makes them more prone to Osteoporosis.

CHAPTER 3

Regression Algorithms and Predicting BMD through Regression MLA

Once the best features that correlated to BMD were determined, we wanted to see if we could use them in a regression machine-learning algorithm to predict BMD and diagnose Osteoporosis. Three main machine-learning algorithms were tested: Linear Regression (LR), Random Forest Regression (RFR), and Artificial Neural Network (ANN). We also find the best number of features to use for each machine-learning algorithm by evaluating the root mean squared error (RMSE) as our accuracy reading.

The data was randomly split so that 80% of the data was used for training, and the other 20% of the data would be used in order to test the model. A fixed random seed was used for reproducibility. The accuracy of a predictive model is highly dependent on the data that is used to train. The 80% of the data used to train is randomly selected, and so it will train the model in a particular way that another 80% randomly selected data would not. To address this, cross-validation was done. Cross Validation is when you spilt the data, one part for testing, and one part for training. The data saved for training is split into k subsets. I used a 5-fold cross validation, and so the 80% data was split into 5 subsets. Each subset is used to train the model one at a time. Once the model is trained, it will predict values, which will then be compared to the testing data. The accuracy of the 5 different models will be averaged to get the average accuracy of the model, ensuring that we have the average case for the training data.

For the RFR two main parameters that were evaluated were the n_estimators parameter and the max_features parameter. A RFR is an ensemble MLA, meaning it takes the average of multiple smaller MLA to determine an outcome. A RFR creates a

variety of decision trees of different features and depths, each of which are used to determine the outcome, and are then averaged to reach a final outcome. The n_estimators parameter determines how many trees to make for the forest. So naturally, the more trees there are, the more accurate the prediction should be. However, after a certain amount, adding more trees minimally affects the accuracy of the RFR. For that reason, many different numbers of n_estimators were tested to find the minimum n_estimators required to get the lowest RMSE score. The other parameter that was tested was the max_features parameter, which determines how many features are used in creating the tree. The best two options given are None and "sqrt" so both were tested along side the n_estimators. It was found that the best combination of the two was 60 n_estimators with a max_features of 'sqrt'.

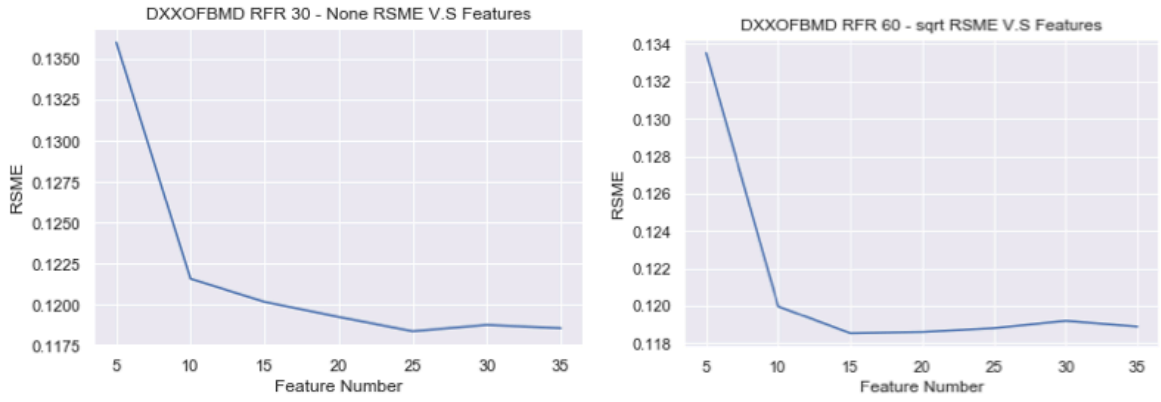

Figure 15. Comparison of RSME for Random Forest Regression with Different Variables

In order to ensure that the best ANN was produced for this test, a variety of different hyper parameters were tested. The scoring metric used to determine this was root-square-mean-error (RSME) and mean-average-percentage-error (MAPE). For this project, the following hyper parameters were tested: solver type, activation method, and alpha value.

There are different solver types for weight optimization that are applicable to the ANN model. The solver types that were compared were adam and lbfgs. At first, the default ‘adam’ was used, but this weight optimization method is mainly used for large datasets. The data set for this project is around 2015 by 72, and so ‘adam’ was not producing the desired results. The RSME score using ‘adam’ was relatively around the same for any number of features used which meant the data was either being overfit – the model was working very well for the training data but not for anything else and so results in errors for predictions- or it was being underfit – the model was not working well with the data. The second weight activation method tested was ‘lbfgs’. This produced the best results, having both the RSME and MAPE scores lower as more features were used to train the model.

The 4 different activation functions were used, tested with different alpha values to see which produced the best results. Both ‘identity’ and ‘relu’ produced models whose error increased as the number of features increased, so they were not the best activation function to use for the project. The activation function paired with alpha value that yielded the best results were that of ‘tan’ with alpha value 10 and ‘logistic’ with alpha value 1. The logistic’ with alpha value 1 was slightly better, so those parameters were used.

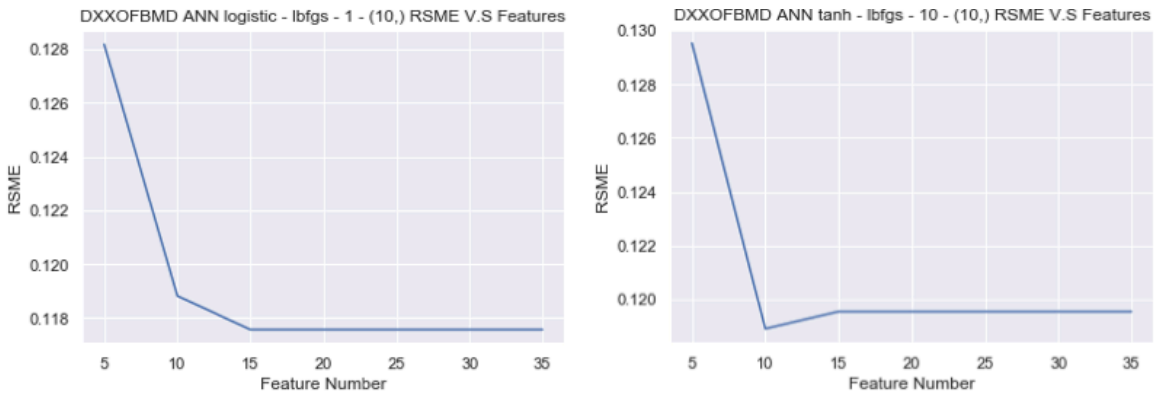

Figure 16. Comparison of RSME for Artificial Neural Network with Different Variables

Once the MLA's were fine-tuned, we were able to discover the best number of features to use for the MLA.

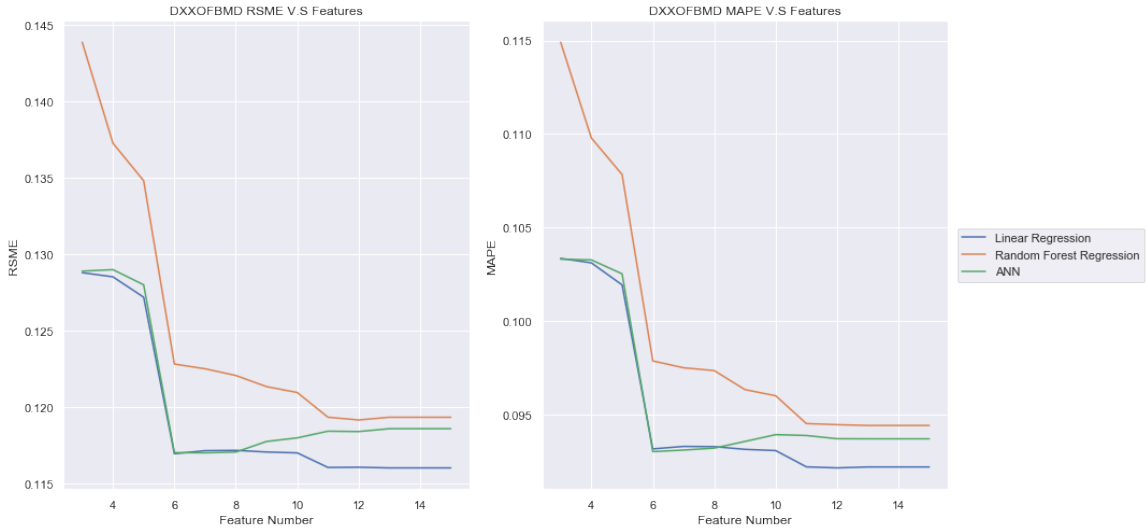

Figure 17. Comparison of different Machine Learning Algorithms

The optimal number of features to use in a LR is 11. The optimal number of features to use in a RFR is 12. The optimal number of features to use in an ANN is 6. As the number of features increases, the RSME and MAPE score starts to increase.

CHAPTER 4

Results and Conclusions

Osteoporosis is a medical condition that affects people of many different backgrounds. This medical condition usually occurs with older patients, but a medical phenomenon has occurred where astronauts in space have a higher chance of being diagnosed with this disease. Astronauts need to be in peak physical condition and are generally between the ages of 26-46 (NASA). It has been know that spaceflight, for reasons still yet determined, cause young and healthy adults to exhibit conditions of osteoporosis much earlier than usual. For this reason, it would be very helpful to predict if an astronauts was exhibiting symptoms of Osteoporosis in order to take preemptive measures against it.

Osteoporosis is generally diagnosed by a by a doctor with the help of a Dual X-ray machine. The machine is able to scan a part of a patient's bone, usually around the hip or femur, and determine the BMD of said patient. That patient's BMD is then compared to the average BMD of a healthy person in their age group. If their BMD is -2.5 standard deviations from the average, osteoporosis is diagnosed. This process requires the expertise of a medical professional and a big machine that needs to be operated, both of which is near impossible to take into Space to monitor an Astronaut's health. For this reason, the goal of this project is to see if there is a way to create a predictive model to monitor an Astronauts health to prevent osteoporosis.

The first goal of this project was to find different trends in osteoporosis in relation to laboratory data. Based on the correlation coefficient of each feature to another, it was determined that the highest correlated features to BMD are: gender, weight, age, height,

arm circumference, upper leg length, arm length, systolic blood pressure, red blood cell count, creatinine found in urine, body mass index, percentage of hematocrit blood, waist circumference, and subscapular skinfold. These features can be grouped into three categories: physical attributes, blood conditions, and materials found in urine. Based on the trends of the physical conditions, we can conclude that smaller individuals with lower weight have a higher chance of developing osteoporosis.

The second goal of this project was see what the best MLA is to determine BMD. Many extensive tests were taken to fine-tune the different parameters of the MLA in order to yield the best results. RSME and MAPE were used to determine the efficiency and correctness of the MLA. It was concluded that the best regression MLA to use to determine BMD is a LR using the following eight features: gender, weight, age, height, arm circumference, upper leg length, arm length, and the average of systolic blood pressure taken three times.

REFERENCES

"Bone Density Test." *Mayo Clinic*, Mayo Foundation for Medical Education and Research, 7 Sept. 2017, www.mayoclinic.org/tests-procedures/bone-density-test/about/pac-20385273. Accessed Nov. 2018.

"Bone Mass Measurement: What the Numbers Mean." *National Institutes of Health*, U.S. Department of Health and Human Services, (n.d.), www.bones.nih.gov/health-info/bone/bone-health/bone-mass-measure. Accessed Nov. 2018.

"Diagnosing Osteoporosis." *Diagnosing Osteoporosis | International Osteoporosis Foundation*, (n.d.), www.iofbonehealth.org/diagnosing-osteoporosis. Accessed Nov. 2018.

Kalatzis, Mortazavi B, and Mohammad Pourhomayoun. "Interactive Dimensionality Reduction for Improving Patient Adhearence in Remote Health Monitoring." *The 2018 International Conference on Computational Science and Computational Intelligence (CSCI'18)*, Dec. 2018.

NASA, NASA, (n.d.), astronauts.nasa.gov/content/faq.htm. Accessed Sept. 2018.

Sawada, Kae, et al. "Analyzing the Mutation Frequencies and Correlation of Genetic Diseases in Worldwide Populations Using Big Data Processing, Clustering, and Predictive Analytics." *2017 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2017, doi:10.1109/csci.2017.255.

Sawada, Kae, et al. "Analyzing the Potential Occurrence of Osteoporosis and Its Correlation to Cardiovascular Disease Using Predictive Analytics." *International Journal On Advances in Life Sciences*, vol. 10, 2018.

Sawada, Kae, et al. “Predictive Analytics to Determine the Potential Occurrence of Genetic Disease and Their Correlation: Osteoporosis and Cardiovascular Disease.” *Proceeding of The Tenth International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies*, 2018.

United States Department of Health and Human Services. Centers for Disease Control and Prevention. National Center for Health Statistics. National Health and Nutrition Examination Survey (NHANES), 2007-2008. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2012-02-22. https://doi.org/10.3886/ICPSR25505.v3

Vahedi, Mohammad Reza, et al. “Predicting Glucose Levels in Patients with Type1 Diabetes Based on Physiological and Activity Data.” *Proceedings of the 8th ACM MobiHoc 2018 Workshop on Pervasive Wireless Healthcare Workshop - MobileHealth18*, 2018, doi:10.1145/3220127.3220133.

Yoo, Sangseo, et al. “Interactive Predictive Analytics for Enhancing Patient Adherence in Remote Health Monitoring.” *Proceedings of the 8th ACM MobiHoc 2018 Workshop on Pervasive Wireless Healthcare Workshop - MobileHealth18*, 2018, doi:10.1145/3220127.3220131.