Do demographics, or other variables, affect the incidence of osteoporosis?

Bagheri, A.*,* Coppi, L., Hernandez, C., Khan, M., Negm, A. and Wai, Y.

# Section 1: Introduction

We decided to work with the Osteoporosis dataset available at <https://www.kaggle.com/datasets/amitvkulkarni/lifestyle-factors-influencing-osteoporosis>. The dataset is organized as a csv file where every row is a patient or person and several columns representing independent variables possibly affecting the risk of developing osteoporosis, and a final column called “Osteoporosis” where 2 possible values are present: 0 if the patient doesn’t have osteoporosis, and 1 if the patient has the disease.

The objective of our project is to analyze the dataset to determine which variables, or combination of variables, have a greater or lower possibility than average to be correlated with osteoporosis. Furthermore, we aim to create a simple predictive model of osteoporosis based on the more relevant of these variables.

We decided to assign the independent variables to groups based on similarity; then we assigned a group to each one of 5 students. The 6th student would then develop the predictive model.

The variables have been organized in the following way:

1. Age, gender, and race/ethnicity: we have called this group “demographic’. It’s assigned to Cris.
2. Family history (of osteoporosis), prior fractures and medications (corticosteroids): assigned to Luca.
3. Alcohol and smoking usage: assigned to Yonnie.
4. Calcium and Vitamin D intake, and hormonal changes: assigned to Ali.
5. Bodyweight, physical activity, and medical condition: assigned to Amro.
6. The modeling task has been assigned to Muneeb.

Our main hypothesis is that demographics have a strong impact on the incidence of osteoporosis; however, we are studying all the variables.

The next sections of this document will go over the work each member carried on the subset of variables they were assigned. Each section will draw conclusions on their part of the study. A separate discussion file will be submitted as a short report for the open discussion. The Appendix at the end of the document contains figures cited throughout the document.

# Section 2: Demographics (Cris)

When we submitted the proposal for this project, we wanted to identify if there was a particular segment at higher risk of Osteoporosis and what a better way to describe a segment than by its demographics as their main characteristics. In our dataset we can find 3 demographic variables: Age, Gender and Race/ Ethnicity, which we decided to group together to perform this first analysis. They will be analyzed separately as we assume they are not related to each other.

**Objective:**

To validate if the probability of people who suffer osteoporosis vary or change depending on their demographic variables.

If there is an important difference (over 10% of change) we also want to measure and identify the correlation between the demographic variable and the variable osteoporosis to determine if it could be used in a predictive model where the demographic variables are considered as independent, and the osteoporosis variables would be the dependent one.

**Data Preparation**

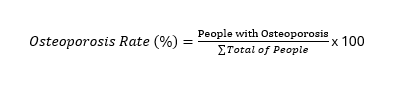
After loading the dataset, the data used for this analysis was filtered in a different subset that only considered the variables relevant for this part, they are:

* Age: integer variable with 1958 non null entries that range from 18 to 90
* Race/Ethnicity: Object(String) with 1958 non null entries, it is a categorical/nominal variable where the possible values are: Caucasian, African American, and Asian
* Gender: object (string) with 1958 non null entries, it is a categorical/nominal variable where the possible values can be either Female or Male.
* Osteoporosis: Object(Integer)with 1958 non null entries. It is an integer used as a dummy variable to represent the absence (0) or presence (1) of Osteoporosis in the subject.
* ID: Integer variable with 1958 non null entries. Values ranged from 1 to 1958 as an identifier for each entry, used only for counting purposes.

The data was fairly clean and completed for this section and did not require too much preparation, however, considering the wide range of values presented for the variable Age it was decided to create a new column called “Age Group” that would group the Age and create a Categorical/ Ordinal variable that could allow us to organize and analyze more effectively the data. Using the. loc function along with a list of conditions, the groups created were: 1.[18-30], 2.[31-45], 3.[40-60], 4. [60-75], 5.[76-90].

**Analysis**

This analysis was based on descriptive statistics by observing the occurrence of osteoporosis in our dataset according to our different variables, this analysis was performed separately for Age, Gender and Race. We used the **Osteoporosis Rate** to measure the probability of acquiring osteoporosis in each group of every variable**.** This rate was calculated using the following formula:



In order to perform this calculation:

1. We first grouped the data at two levels: The variable we were analyzing (Age, Gender or Race, one at the time) and Osteoporosis by using the .groupby() function.
2. then we pulled the total of people by counting all our entries using the ID variable and the .count() function
3. Next, through the function .unstack() we separated the number of people with osteoporosis from the people without this condition for each category.
4. With all this we finally had all the parameters needed in our formula and then created a new column to display the result of that calculation.

The results reveal that the occurrence of osteoporosis is quiet even for all groups of our variables. For Gender, in women osteoporosis was presented 49.38 % of the times, while men had it 50.60%. In terms of race/ethnicity, the rate was also very balanced for all groups: African American had an osteoporosis occurrence of 50.52% while Asian and Caucasian had it 49.76% and 49.69% respectively. This shows us that these different groups of participants are somehow well represented but they probably don’t influence the risk of osteoporosis; For instance, if we consider gender, almost 50% of our dataset is related to women and 50% for men and both groups have similar chances to develop osteoporosis. There is no group for Gender that suggests it has been more susceptible to suffer osteoporosis. This same result applies for Race/Ethnicity.

On the Other hand, Age and for our Analysis (Age Group) is a different story, there seems to be a relation between Age and the rate of people that suffer Osteoporosis visible in Figure1. The older our group gets, higher is the presence of the disease, the difference in the osteoporosis rate is relevant up to a point where everyone from 45 years old (our third group) has osteoporosis. This raises up a flag because it either indicates an extraordinary relationship between age and osteoporosis or it indicates our sample of study is probably biased since it is hard to believe it is a reliable representation of reality, which means using Age in our model could make us to overestimate our predictions. This last result led us to dig a little bit deeper in our analysis though descriptive statistics to verify the Age distribution of our data set and correlation of Age with Osteoporosis.

Through plot.hist() we observed in Figure 2 that our population is not normally distributed, there is a strong skew of 0.9087 that indicates that most of our records concentrate before 32 years old, while the maximum range is 90. Having this in mind, we evaluated the correlation between Age and Osteoporosis using .corr(), the result was 0.69, a fair indicator of positive correlation ( the highest the age, most closely the subject is to osteoporosis), however when we analyze this correlation among the different age groups defined with function .corrwith(), the list generated provide a NaN correlation value for our three older groups, which suggest there is no variability in our data. We confirm this using plt.scatter displayed in figure 3.

To finalize our analysis, given the correlation results for Age and osteoporosis we ran OLS regression with the statsmodel library to see if by using age as our independent variable we could model to estimate osteoporosis as our dependent variable. The results displayed in Figure 4 indicate with R-squared that -with our model- 47.8% of our dependent variable can be explained with Age. In addition, our P value of 0.000 suggests an almost certain probability of rejecting a null hypothesis, which means our independent variable does influence our dependent one.

**Conclusion**

* Variables Gender and Race/ Ethnicity do not influence the risk of Osteoporosis as the probabilities to suffer Osteoporosis within all their groups are very similar -almost 50% for all of them- just like flipping a coin.
* For our dataset, the Age variable is positively correlated with Osteoporosis and could be used in a model as an independent variable to estimate Osteoporosis as a dependent one with a degree of trust. However, our observations reveal that in this dataset osteoporosis is a certain result, while this explains why a relationship was identified, it could also raise a flag about the variability of our data and the possibility it is biased.
* The dataset of our analysis has a positive skewness in Age distribution, on the other hand there is no variability in the older groups of our sample, meaning even if a correlation was found, observations and results are probably not an accurate representation of reality. Or the dataset should be used to probably answer other questions.

# Section 3: Family History, Prior Fractures and Medications (Luca)

**Objective**

The objective of this section is to determine if and how strong of a correlation exists between the variables family history, prior fractures and medications, and osteoporosis. I will also explore whether specific groups have an incidence of osteoporosis that differs from the dataset average.

**Data Preparation**

After loading the csv file where the dataset is organized, I selected the four columns representing the variables I was interested in: Family History (of osteoporosis), Medications, Prior Fractures (independent variables) and Osteoporosis (dependent variable).

Checking the dataset for integrity, I noticed that the column “Medications” contained a large number of missing values. Upon inspecting the original data file, I realized that the ‘None” entries in the column had been read as NaNs in pandas. Thus, I converted the missing values in the “Medications” columns to a “None” string, representing a patient who doesn’t use corticosteroid medications.

**Analysis**

The three independent variables are strings (objects in pandas), indicating they are categorical variables. All of them have only two possible values each. Because all the variables are categorical, there is no info to obtain from using describe() or other pandas methods to analyze the distribution of the dataset, except for the dependent variable, Osteoporosis, which is a numerical value of either 1 (patient has osteoporosis) or 0 (patient doesn't have osteoporosis). Osteoporosis has a mean of exactly 0.5, indicating half of the patients have the disease, and half don't. When checking the situation of specific groups, a group average greatly different from 0.5 would indicate that the group definition likely has an impact on osteoporosis incidence.

With that in mind, we can use the pandas *groupby* method to check whether a variable affects osteoporosis. No transformations are required since the independent variables are all categorical. Family history doesn't appear to have an effect on osteoporosis, since the two groups (Yes and No family history) have an average osteoporosis value close to 0.5, indicating about half of the subjects in each group have osteoporosis. Taking corticosteroids appears to very slightly increase the outcome of developing osteoporosis; the difference among the two groups (taking or not this specific medication) is very small, and barely greater than a standard error of the mean. There is no clear effect of prior fractures either. It's interesting to also note how the count of the groups it's quite even for all three variables.

The next step of my analysis is to study whether patient groups arising from a combination of variables have an incidence of osteoporosis that differs from the mean of 0.5. For this purpose, we can use the *groupby* method on several variables at the same time.

Among people with no family history, those who use corticosteroids have a 53.7% incidence of osteoporosis, compared to 46.5% of those who don't. The difference is small but it's greater than the standard error of the mean, thus it could indicate an effect of these factors combined. Other groupings of two don't reveal groups that have a strong difference in osteoporosis incidence.

Grouping the data by all the three variables shows two additional groups that standout, even though the differences from the overall mean are still small:

1. People with no family history, who use corticosteroids and have had prior fractures, have greater incidence of osteoporosis (56.3%).
2. People with family history, who use corticosteroids, and had no prior fractures, have a smaller incidence of osteoporosis (46.2%).

**Conclusion**

Because the combination of the three variables offers a better relationship with osteoporosis than the single variables or groups of two, we can create a new variable by combining the three existing variables. We can assign this variable with the apply method to a new column of the dataframe.

This new combined variable ("Diseases") offers a better relationship with osteoporosis. The correlation is still small.

# Section 4: Alcohol and Smoking (Yonnie)

# **Objectives:** The primary objective of this data analysis is to explore the factors associated with osteoporosis risk and to identify potential predictors of the disease. The analysis aims to answer the following questions: How do lifestyle factors such as alcohol consumption and smoking influence osteoporosis risk? What trends and patterns exist in the data that may help predict osteoporosis risk?

**Data Preparation:** After loading the CSV file, I focused on the Alcohol Consumption and Smoking columns. ‘NaN’ in the "Alcohol Consumption" column was replaced with "None" to indicate non-drinkers by using: *df['Alcohol Consumption'].fillna('None', inplace=True).* In this case, NaN is not a missing value, instead, it is one of the values in the column. Therefore, NaN should not be dropped.

The numbers of ‘None’ and ‘Moderate’ in the Alcohol Consumption column, and the numbers of ‘Yes’ and ‘No’ in the Smoking column were counted by using: *alcohol\_counts = df['Alcohol Consumption'].value\_counts()*

*smoking\_counts = df['Smoking'].value\_counts()*

The dataset was converted from categorical variables into numerical representations. Specifically, the "Smoking" and "Alcohol Consumption" columns were mapped to numerical values:

* "Yes" in the "Smoking" column was mapped to 1, indicating a smoker.
* "No" in the "Smoking" column was mapped to 0, indicating a non-smoker.
* "Moderate" in the "Alcohol Consumption" column was mapped to 1, indicating moderate alcohol consumption.
* "None" in the "Alcohol Consumption" column was mapped to 0, indicating no alcohol consumption.

**Data analysis:** 988 individuals answered ‘None’ and 970 individuals answered ‘Moderate’ for alcohol consumption. Approximately 50% of individuals (982) reported being smokers, while the other 50% (976) were non-smokers.

The distribution of smoking and alcohol consumption columns was visualized using count plots. Both plots in *Figure 5* showed an equal distribution between smokers and non-smokers and between non-drinkers and moderate drinkers.

The correlation matrix was calculated to examine the relationship between alcohol consumption and smoking with osteoporosis risk. In *Figure 6*, the correlation coefficient between smoking and osteoporosis risk is -0.02, meaning a negative correlation. The correlation coefficient between alcohol consumption and osteoporosis risk is 0, meaning no correlation between the two variables.

**Conclusions:** The analysis of the dataset focused on the impact of lifestyle factors such as alcohol consumption and smoking on osteoporosis risk. The data revealed that alcohol consumption and smoking were equally distributed among the participants, with approximately 50% reporting moderate alcohol consumption and the other 50% reporting no alcohol consumption. Similarly, the distribution between smokers and non-smokers was also evenly split. This suggests that these lifestyle factors, at least in the levels of exposure reported by the survey participants, are not important predictors of osteoporosis risk. These findings emphasize the need for further research, potentially looking into other lifestyle, genetic, or environmental factors that could more significantly impact osteoporosis risk.

# Section 5: Calcium, Vitamin D and Hormonal Changes (Ali)

**Objectives:**

The objective of looking at the Calcium, Vitamin D and Hormonal Changes is to find any correlations, weak or strong, to how these variables affect an individual’s chances of developing Osteoporosis. These variables are established by the data source, are the relevant information that predict osteoporosis, and include demographics data. I hope to look deeper into these three variables and discover any correlations with high accuracy and with other variables that would affect the result of these three variables, like age and gender, in mind. I will first analyze the data to make sure that there is no missing or corrupted data, then investigate the three columns and find the ratios of Osteoporosis for each. My hypothesis is that for vitamin D and Calcium, low intake predicts Osteoporosis, and hormonal changes, postmenopausal predicts Osteoporosis.

**Data Preparation:**

The first step was to check that the data is clean. I analyzed all the columns to see how many Na/NaN values each had. Although the online source, Kaggle, said that there was no missing data, my code shows that the Alcohol Consumption column had 988 Na’s, Medical Conditions column had 647 Na’s and the Medications column had 985 rows missing values. After a closer look, I was assured that all three column’s null values are not actually missing, but they represent useful information. This can be ignored or changed but I decided to keep it as Na values. I used python’s Pandas and Numpy libraries to come to these conclusions. After loading the data into a dataframe, we can count the number of Na values.

**Analysis:**

The first step of the analysis was to group each variable and find the number of people with Osteoporosis and the number of people without Osteoporosis in each category of the columns that I had been assigned to (Calcium intake, Vitamin D Intake and Hormonal changes). Each group has its own categories: Calcium intake has Adequate and Low, Vitamin D Intake has Insufficient and Sufficient and Hormonal changes has Postmenopausal and Normal. I then divided the number of Osteoporosis by the total for each group in each column to get the percentage. This shows the percentage of those who have Osteoporosis given each category. For example, 50.209644% of those who had an adequate amount of Calcium intake had Osteoporosis! All three columns are graphed, showing the amount of Osteoporosis, non-Osteoporosis, and the total in graphs. The percentages are also shown as outputs in my code.

After this analysis, the results are not very indicative of Osteoporosis, because they are all less than 2% difference! There is definitely something wrong with the data, as it is showing that highest rate of Osteoporosis in Calcium Intake is those who have Adequate intake at a 50.209644% rate, Sufficient for Vitamin D intake at 50.840752% and Postmenopausal for Hormonal Changes at 50.767656%.

After speaking with teammates, we came to the conclusion that the data is not collected properly. Those who worked with the Age variables mentioned that everyone over 40 years of age had Osteoporosis and even in the younger ages, the data was skewed in favor of those who had Osteoporosis, but it was not 100%. This has likely happened because those who collected this information only asked those who went to the doctor for Osteoporosis. The data is not a good sample of the population. This also makes sense with my conclusion, because for all three variables, there is no relevant difference between all the categories in each variable.

To further analyze the data with this new information, I conducted analysis for people aged less than 40 and another analysis on a subset of the data with 50% from those who have osteoporosis, and 50% from those who do not.

For these new analyses I reused the code written previously by creating a function. The function “getAnalysis(MyDF)” gets a dataframe as an argument. This function was then called 3 times. The first was using the original dataframe. The second, a new altered DataFrame was set as the argument that had every data point with age less than 40. Finally, the third call was called with a dataframe where 50% of its data points had Osteoporosis.

Looking at the analyses of all three dataframes, the conclusion was still inconclusive. The differences were not large enough for each variable for it to be considered affecting Osteoporosis.

**Conclusions:**

The data set does not conclude anything about whether vitamin D Intake, Calcium Intake and Hormonal Changes affect Osteoporosis. This is most likely because the data was collected from those who had Osteoporosis, or were tested for it. My analysis shows very little difference between the amount of calcium, vitamin D intake and Osteoporosis. The amount is very negligible even when only looking at those younger than 40 years old and creating a subset of the original data to have 50% Osteoporosis and 50% non-Osteoporosis results. The same is true for hormonal changes. The data suggests that postmenopausal has a stronger indication of Osteoporosis but with only a 1-2% difference! This is not strong, and we cannot conclude anything.

The data collected should include more people of different ages who also do not have Osteoporosis so that there is something this data can be compared to.

# Section 6: Bodyweight, Physical Activity, and Medical Condition (Amro)

**Objective:**

The aim of this analysis is to comprehensively explore the potential influence of body weight, physical activity levels, and medical conditions on the incidence of osteoporosis. By delving into these variables, we aim to discern any patterns or correlations that may exist with osteoporosis risk. Additionally, we endeavor to identify whether specific combinations of these factors exhibit a stronger association with osteoporosis incidence, with the goal of enhancing predictive models in osteoporosis risk assessment.

**Data Preparation and Preprocessing:**

Upon loading the dataset, our attention was directed toward four pivotal variables: Body Weight, Physical Activity, Medical Conditions, and Osteoporosis. Recognizing the categorical nature of these variables, each with two possible values, our dataset underwent meticulous cleansing to ensure data integrity.

The 'Osteoporosis' column in the DataFrame is converted to binary format (0 or 1), assuming it originally contained categorical data. Only for the heatmap, a subset of columns relevant to the analysis is selected: 'Body Weight', 'Physical Activity', 'Medical Conditions', and 'Osteoporosis'. Categorical variables ('Physical Activity' and 'Medical Conditions') are converted to numerical format using one-hot encoding. This converts categorical variables into binary vectors, making them suitable for analysis. The correlation matrix is calculated using the corr() method in Pandas, which computes the pairwise correlation of columns in the data frame. A heatmap of the correlation matrix is plotted using Seaborn and Matplotlib libraries. This heatmap visualizes the correlations between different attributes, with higher absolute values indicating stronger correlations. The color scale ('coolwarm') is used to represent positive and negative correlations, with blue indicating negative correlations, and red indicating positive correlations. Annotations are added to the heatmap to display correlation coefficients. Finally, a bar chart is created to visualize the incidence of osteoporosis based on different combinations of body weight, physical activity, and medical conditions.

For bar graphs, the data remained in the original format except for medical conditions, which have “None” as a value. It was replaced with “No Med Con” to avoid confusion in the code and to differentiate between patients who have medical and non-medical conditions. The code then creates a bar plot to visualize the osteoporosis incidence for each combination of Body Weight, Physical Activity, and Medical Conditions. It uses the plt.bar() function to create the plot, with the x-axis representing the index of each combination and the y-axis representing the osteoporosis incidence percentage. The bars are colored sky blue for better visualization.

**Analysis:**

Our analysis began by scrutinizing individual variables against the possibility of having osteoporosis, yielding nuanced insights. For body weight analysis, 'Underweight' individuals exhibited a 50.7% incidence rate of osteoporosis, compared to 47.8% for those categorized as 'Normal' weight. Furthermore, regarding physical activity, 'Sedentary' individuals demonstrated a slightly higher incidence rate of 49.5% compared to 46.9% for 'Active' individuals. Also, individuals with reported medical conditions showcased a substantially higher incidence rate of 51.3% compared to 46.4% for those without such conditions.

In the second part of the analysis, we compared the combination of the three parameters—body weight, physical activity, and medical conditions—against the possibility of having osteoporosis. The combination included all the possibilities of the subcategory of each variable against the risk of osteoporosis occurrence. The classification included sorting the combinations of subcategories of body weight, physical activity, and medical conditions respectively against the risk of having osteoporosis. The combinations of subcategories were posted from lowest to highest risk as follows:

1. Normal - Sedentary - No Med Cond: 44.03%
2. Normal - Active - Rheumatoid Arthritis: 44.94%
3. Underweight - Active - No Med Cond: 45.86%
4. Normal - Active - Hyperthyroidism: 46.96%
5. Normal - Sedentary - Hyperthyroidism: 49.36%
6. Normal - Sedentary - Rheumatoid Arthritis: 50.00%
7. Underweight - Active - Rheumatoid Arthritis: 50.94%
8. Underweight - Active - Hyperthyroidism: 51.57%
9. Underweight - Sedentary - No Med Cond: 52.08%
10. Normal - Active - No Med Cond: 54.01%
11. Underweight - Sedentary - Hyperthyroidism: 54.40%
12. Underweight - Sedentary - Rheumatoid Arthritis: 56.92%

The analysis refers to the lowest possibility of the risk of osteoporosis occurring for the first group of patients who have “Normal” body weight, “Sedentary” physical activity, and no prior medical conditions. The highest group that faces the risk of osteoporosis is the latest one of the patients whose body weight is Underweight with physical activity classified as Sedentary and their medical conditions classified as Rheumatoid Arthritis. The difference in risk percentage of osteoporosis between those two groups (the first and last) is 12.89%.

These results suggest that while individual factors such as body weight, physical activity, and medical conditions may not independently correlate strongly with osteoporosis incidence, their interactions can have some influence on the likelihood of developing the condition. This underscores the importance of considering multiple variables simultaneously when assessing osteoporosis risk.

**Conclusion:**

In conclusion, our analysis sheds light on the intricate interplay between body weight, physical activity levels, and medical conditions concerning osteoporosis incidence. Through meticulous examination of individual variables and their combinations, we uncovered nuanced insights into risk factors associated with this condition. While individual factors such as body weight, physical activity, and medical conditions may not independently exhibit strong correlations with osteoporosis incidence, their combined effects unveil some associations.

Our findings reveal that certain combinations of these factors demonstrate varying degrees of risk differences, emphasizing the importance of considering multifactorial influences in osteoporosis risk assessment. Our analysis shows that individuals classified as underweight, leading sedentary lifestyles, and presenting with medical conditions such as rheumatoid arthritis face a notably higher risk of osteoporosis, at 56.89%. Conversely, those categorized as having normal body weight, leading sedentary lifestyles, and lacking medical conditions show comparatively lower risk percentages of 44.03%. The varying degrees of risk percentages between subcategories of the three examined parameters have correlations and some differences, and they help provide a bigger picture than analyzing only the individual parameters.

# Section 7: Modeling (Muneeb)

## 7.1 Data Preprocessing

During the preprocessing phase of the predictive modeling project, several crucial steps were undertaken to ensure the data were ready for analysis and model building. Firstly, the removal of the ID column was performed as it was deemed unnecessary for modeling purposes. Subsequently, an inspection for any missing values across the dataset was conducted. Upon careful observation, it was noted that NaN values in specific columns such as 'Alcohol Consumption', 'Medical Conditions', and 'Medications' did not signify actual missing data but rather instances where the medical condition was absent. Consequently, all NaN values in these columns were replaced with 0, while non-NaN values were replaced with 1. This encoding schema allowed for the effective representation of absence (0) or presence (1) of certain conditions or behaviors.

Moreover, to avoid any potential errors during modeling, certain columns were renamed for clarity and consistency. Following this, additional categorical columns such as 'FamilyHistory', 'BodyWeight', 'CalciumIntake', 'VitD', 'PhysicalActivity', 'AlcoholConsumption', 'Fractures', and 'Smoking' were scrutinized. For each of these columns, a similar replacement strategy was employed, converting the categorical values into binary form (0 or 1) based on their presence or absence.

## 7.2 Train-Test Split

After ensuring all necessary columns were appropriately encoded, the `train\_test\_split` function was utilized to partition our dataset into distinct subsets, facilitating the training of the model on one set while reserving another for evaluating its performance. This step is crucial to assess the model's performance on unseen data. Subsequently, a RandomForestClassifier model from scikit-learn was employed to train on the training data. Once trained, this model was used to make predictions on the testing data.

Upon making predictions, the model's accuracy was evaluated. The accuracy score provided insights into how well the model performed overall, while the distribution of the target variable in the testing set helped understand the class balance and potential biases in the data. The accuracy score of the predictive model on the testing data is 0.85. This indicates that the model correctly predicted the target variable for approximately 85% of the samples in the testing set. Additionally, a confusion matrix was plotted to visualize the model's performance further, offering insights into true positives, true negatives, false positives, and false negatives.

To shed light on the significance of variables in predicting the target variable, feature importance serves as a crucial concept and reveals compelling insights. It showed that Age emerged as the highest predictor of osteoporosis, attributing 63% of the predictive power, followed by Physical Activity, Calcium Intake, and Alcohol Consumption, which accounted for 32% collectively. This breakdown underscores the pivotal role of age in predicting osteoporosis risk, followed by lifestyle factors like physical activity and dietary habits.

## 7.3 Exploring K-Nearest Neighbors Algorithm for Classification Tasks

Firstly, the dataset was examined using the `describe()` function to gain insights into its statistical properties. The features were scaled using Min-Max scaling to bring them to a similar range between 0 and 1. This step aimed to prevent features with larger scales from dominating the distance calculations. With the dataset preprocessed, the KNN classifier was trained by setting n\_neighbors to 3. The `fit()` function trained the model on the scaled training set, allowing it to learn the inherent patterns in the data. Following training, predictions were made on the scaled test dataset using the `predict()` function. Leveraging the nearest neighbors identified during training, the algorithm predicted the class labels for the test instances. The performance of the model was evaluated using accuracy, computed by comparing predicted labels against ground truth labels from the test set using the `accuracy\_score()` function. Upon evaluation, the KNN classifier achieved an accuracy score of approximately 0.8699, indicating its ability to correctly predict the majority of class labels in the test set. This demonstrates the efficacy of the KNN algorithm in classification tasks, providing a reliable and interpretable solution.

## 7.4 Regression using Random Forest

The Random Forest Regression algorithm is a powerful machine learning technique that combines the predictions of multiple decision trees to produce robust and accurate regression models. In this analysis, a Random Forest Regressor was employed with 200 decision trees to predict the target variable based on the input features. The model was trained on a subset of the data and then evaluated using the test set. The predictions made by the model were compared with the actual values in the test set using two key metrics: Root Mean Square Error (RMSE) and R-squared (R²) score. The RMSE, which measures the average deviation of the predicted values from the actual values, was found to be 0.286, indicating that, on average, the model's predictions were within 0.286 units of the true values. Additionally, the R² score, which represents the proportion of variance in the target variable that is explained by the model, was calculated to be 0.672. This indicates that the model explains approximately 67.2% of the variance in the target variable, suggesting a reasonably good fit. Furthermore, it's notable that the RMSE of the model (0.286) is smaller than the standard deviation of the target variable in the test set (0.500). This indicates that the model's performance is better than a naive prediction that simply predicts the mean value of the target variable. In conclusion, based on the evaluation metrics and the comparison with the standard deviation of the target variable, it can be inferred that the Random Forest Regression model developed in this analysis performs well and provides reliable predictions for the given dataset.

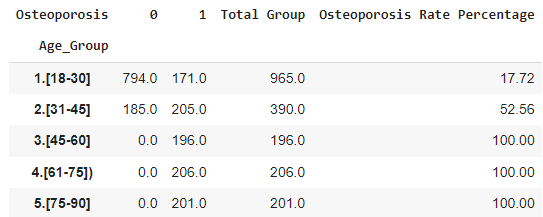
## 7.5 Linear Regression

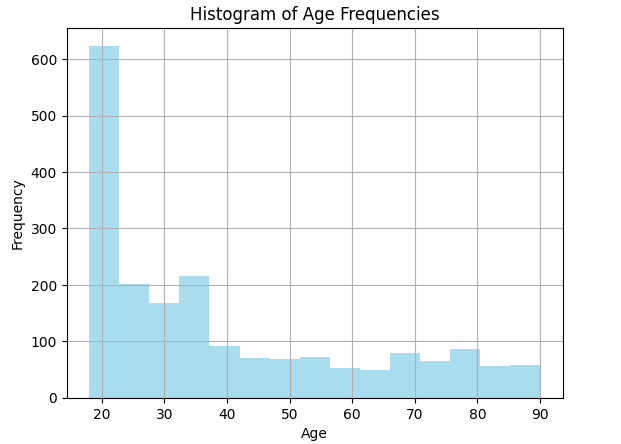
Conducted linear regression analysis after appropriately scaling my features using MinMaxScaler. The process of feature scaling is evidently crucial, particularly when dealing with features of varying scales or units. By doing so, I ensured that the algorithm converges more efficiently while preventing any single feature from unduly influencing the learning process. Following the scaling of the data, I proceeded to fit it to a linear regression model and made predictions on the test set. Subsequently, I assessed the model's performance using two fundamental metrics: root mean squared error (RMSE) and R-squared (R²). The obtained RMSE value, approximately 0.355, indicates the average prediction error of the model, while the R² value of around 0.496 suggests that the model explains roughly 49.6% of the variance in the dependent variable. These metrics collectively offer a comprehensive understanding of the model's predictive capabilities and its ability to generalize to unseen data. Overall, the results suggest that the model performs reasonably well in predicting the dependent variable based on the given features. However, there's always room for enhancement and refinement, and further iterations of the model may lead to even better performance and accuracy.

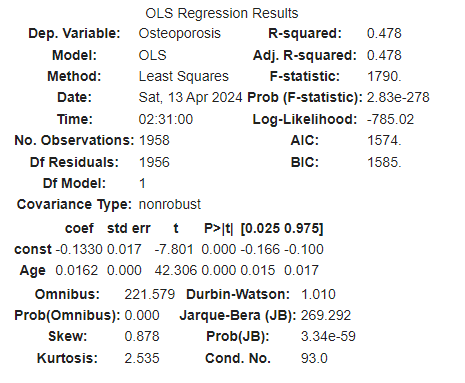
## 7.6 Conclusion

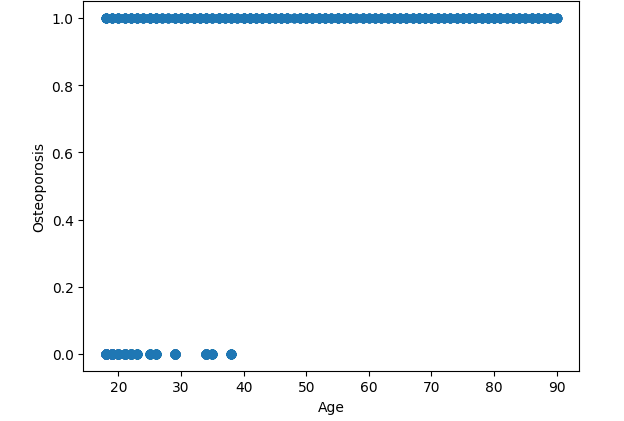
In conclusion, the predictive modeling project underwent comprehensive preprocessing to ensure data readiness, including feature encoding, renaming, and scaling. Through the utilization of various algorithms, including RandomForestClassifier, K-Nearest Neighbors, Random Forest Regression, and Linear Regression, the analysis achieved notable results. The RandomForestClassifier demonstrated a solid accuracy of 85%, while the K-Nearest Neighbors algorithm showcased a higher accuracy of approximately 86.99%, indicating its efficacy in classification tasks. Furthermore, the Random Forest Regression model exhibited robust predictive performance, explaining around 67.2% of the variance in the target variable with a reasonably low RMSE. Finally, the Linear Regression model provided insights into the data's linear relationships, explaining roughly 49.6% of the variance. Overall, these findings highlight the effectiveness of employing various machine learning techniques in predicting osteoporosis risk, paving the way for further refinement and enhancement in future iterations of the model.

# Section 8: APPENDIX

**Figure 1: Osteoporosis Rate by Age Group** **Figure 2: Histogram of Age Frequency** 



**Figure 3: Age/Osteoporosis Scatter Plot** **Figure 4:OLS Regression Results**



# 

**Figure 5: Distribution of Smoking and Figure 6: Correlation between**

**Alcohol Consumption Alcohol Consumption and Smoking**

**with Osteoporosis**

