Causal Analysis Of Osteoporosis Dataset

Introduction

The aim of this analysis is to identify the factors affecting osteoporosis based on various demographic, lifestyle, and health-related features. This analysis includes EDA, logistic regression, feature importance from a Random Forest model, and causal inference using the DoWhy library to estimate the causal effect of Age on osteoporosis while controlling for confounders.

Dataset Overview

The dataset contains 1958 rows and 16 columns, which are: Id, Age, Gender, Hormonal Changes, Family History, Race/Ethnicity, Body Weight, Calcium Intake, Vitamin D Intake, Physical Activity, Smoking, Alcohol Consumption, Medical Conditions, Medications, Prior Fractures, and Osteoporosis.

Data Preprocessing

Missing data was found in 3 columns:

- The 'Alcohol Consumption' column, with 988 missing values, was filled with the value "Unknown".
- The 'Medical Conditions' and 'Medications' columns, with 647 and 985 missing values, were filled with "None".

Categorical variables were cleaned by standardizing their format (lowercasing and stripping extra spaces). Additionally, common typos were corrected:

- The 'Vitamin D Intake' column had 'suffi' and 'fficient' corrected to 'sufficient'.
- The 'Physical Activity' column had 'sedent' corrected to 'sedentary' and 'acti' to 'active'.
- The 'Alcohol Consumption' column had 'modre' corrected to 'moderate'.
- The 'Medical Conditions' column had 'hyperthyroidm' corrected to 'hyperthyroidism'.

Exploratory Data Analysis (EDA)

Several graphs were plotted, the following was revealed:

- **Osteoporosis Distribution:** The plot revealed a balanced distribution between people with and without osteoporosis. (Figure 1).
- **Age Distribution:** A box plot showed that individuals with osteoporosis tend to be older, ranging from 40 to 70 years old. (Figure 2).
- **Gender Distribution:** The plot indicated that osteoporosis affects both genders, with a slightly higher incidence in men. (Figure 3).

- **Hormonal Changes and Osteoporosis:** Postmenopausal individuals were more likely to have osteoporosis than those with normal hormones (Figure 4).
- **Family History and Osteoporosis:** A family history of osteoporosis was less associated with the condition than no family history (Figure 4).
- **Body Weight and Osteoporosis:** Underweight individuals with osteoporosis were more than those without, while normal-weight individuals with osteoporosis were less than those without but more than underweight individuals with osteoporosis (Figure 4).
- **Calcium Intake and Osteoporosis:** Inadequate calcium intake was more linked to osteoporosis than adequate intake (Figure 4).
- **Vitamin D Intake and Osteoporosis:** Insufficient vitamin D intake was less associated with osteoporosis than sufficient intake (Figure 4).
- **Physical Activity and Osteoporosis:** Sedentary individuals with osteoporosis were less than active individuals with osteoporosis but more than those without (Figure 4).
- **Smoking and Osteoporosis:** Smokers had a lower occurrence of osteoporosis than non-smokers (Figure 4).
- **Alcohol Consumption and Osteoporosis:** Moderate alcohol consumption was slightly more linked to osteoporosis than unknown consumption levels (Figure 4).
- **Prior Fractures and Osteoporosis:** Prior fractures were strongly linked to osteoporosis (Figure 4).
- **Correlation Analysis:** Correlation was analyzed for 'Osteoporosis', 'Age', 'Calcium Intake', and 'Vitamin D Intake', with 'Age' showing the highest correlation of 0.69. (Figure 5).

Logistic Regression Model

A logistic regression model was built to predict the likelihood of osteoporosis based on these selected features: Age, Gender, Family History, Physical Activity, Calcium Intake, Vitamin D Intake, Smoking, Alcohol Consumption, Medical Conditions, Medications, Prior Fractures, Hormonal Changes, and Body Weight. The results were as follows:

- **Accuracy:** 82.3% (The model correctly predicted the presence or absence of osteoporosis for 82.3% of the cases in the test set.)
- **Precision, Recall, and F1-Score:** These metrics were balanced for both classes (0 = No osteoporosis, 1 = Osteoporosis), suggesting that the model performs well in predicting both categories.
    - For class 0, the precision was 0.79, recall was 0.86, and F1-score was 0.82.
    - For class 1, the precision was 0.86, recall was 0.79, and F1-score was 0.82.
- **ROC Curve and AUC:** The ROC AUC value was 0.90, indicating that the model is excellent at distinguishing between the two classes, as it has a high true positive rate and low false positive rate across various thresholds. (Figure 6).

Random Forest Model and Feature Importance

The Random Forest model identified 'Age', 'Medical Conditions', 'Gender', 'Family History', and 'Prior Fractures' as the most important features for predicting osteoporosis, based on their high feature importance scores. And 'Smoking' as the least important feature. (Figure 7).

Linear Regression Analysis of Age and Osteoporosis

Age was selected as the treatment variable, as it was identified as the most important feature related to osteoporosis in the random forest analysis. The causal analysis used linear regression to estimate the effect of Age on osteoporosis. The results show that as age increases, the likelihood of developing osteoporosis increases by 34.58% (the backdoor method). This effect is highly significant, with a p-value of 0.0. The model controlled for several confounders, including 'Gender', 'Family History', 'Body Weight', 'Calcium Intake', 'Vitamin D Intake', 'Physical Activity', 'Prior Fractures', 'Medical Conditions', and 'Alcohol Consumption'.

Conclusion

Through this analysis, key factors affecting osteoporosis were identified, including age, medical conditions, family history, and prior fractures. Among these, age was found to be the most significant factor influencing osteoporosis. The logistic regression model showed strong predictive power, with an AUC of 0.90. The Random Forest model provided additional insights into feature importance, further validating the relevance of these factors. The causal analysis confirmed that age has a significant effect on the likelihood of developing osteoporosis, with a robust effect size of 34.58%.

Appendix

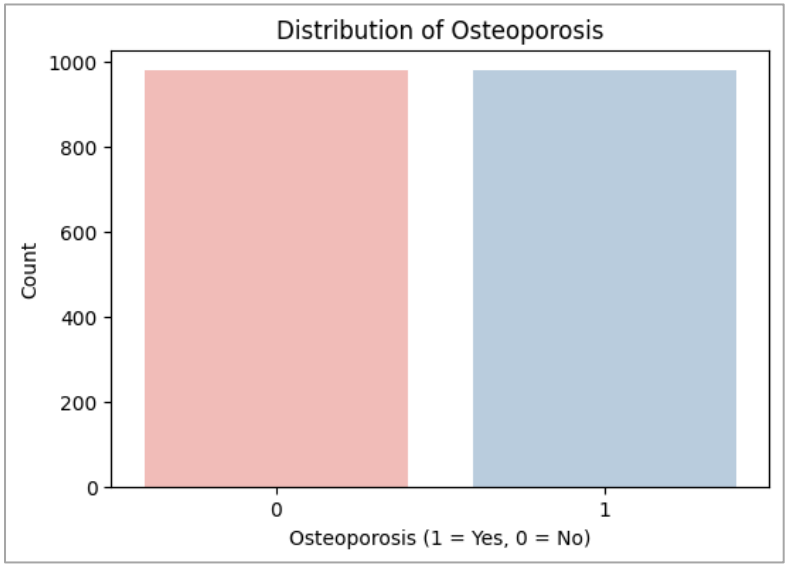

Figure 1

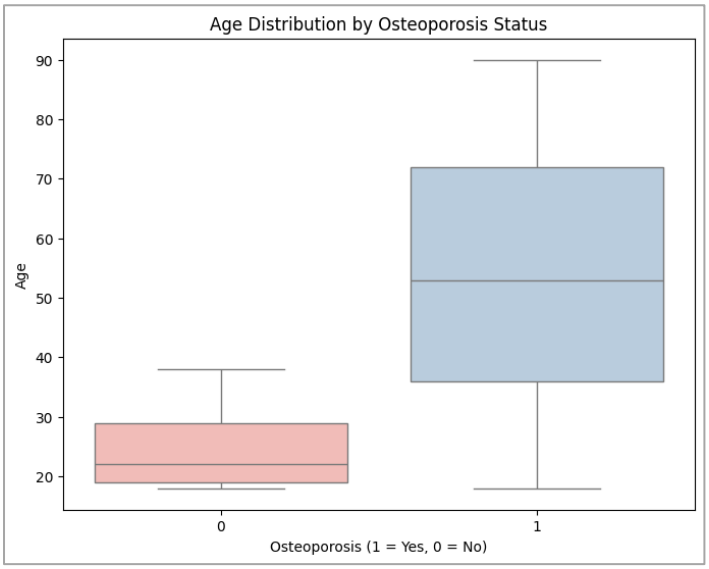

Figure 2

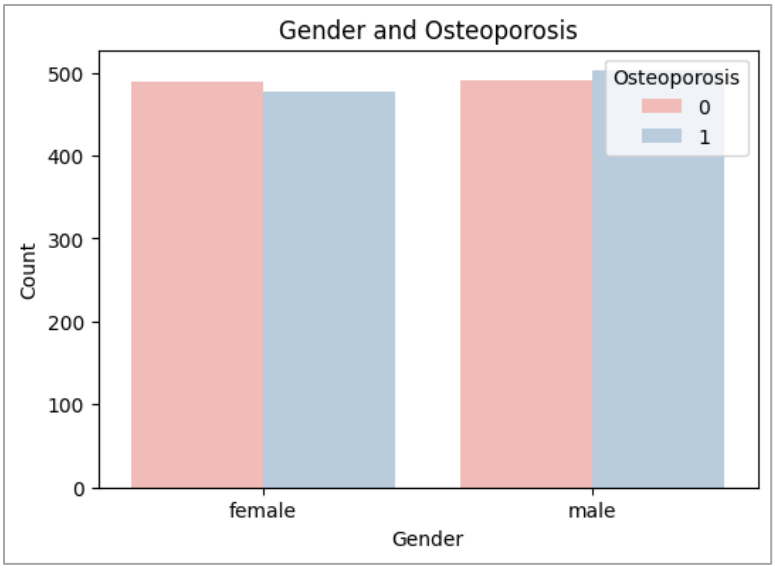

Figure 3

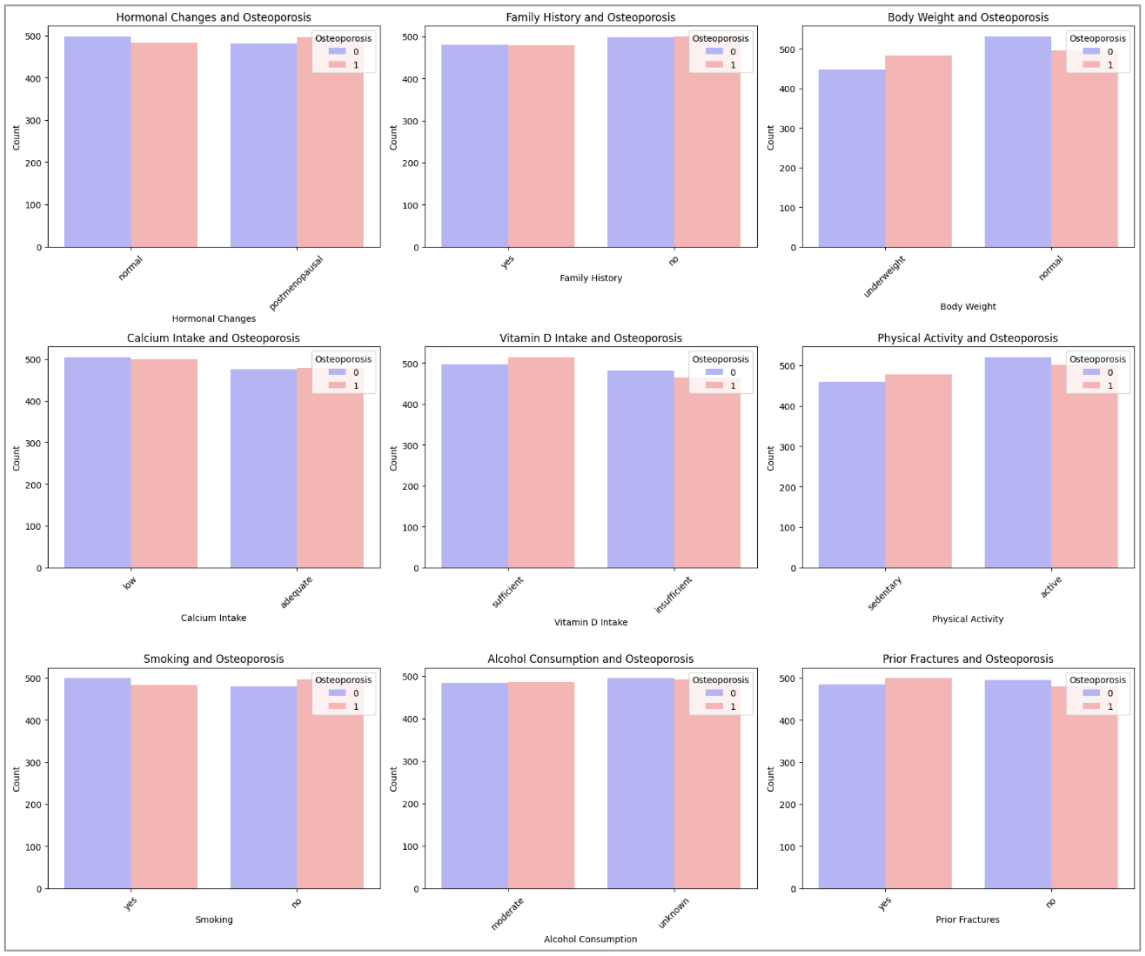

Figure 4

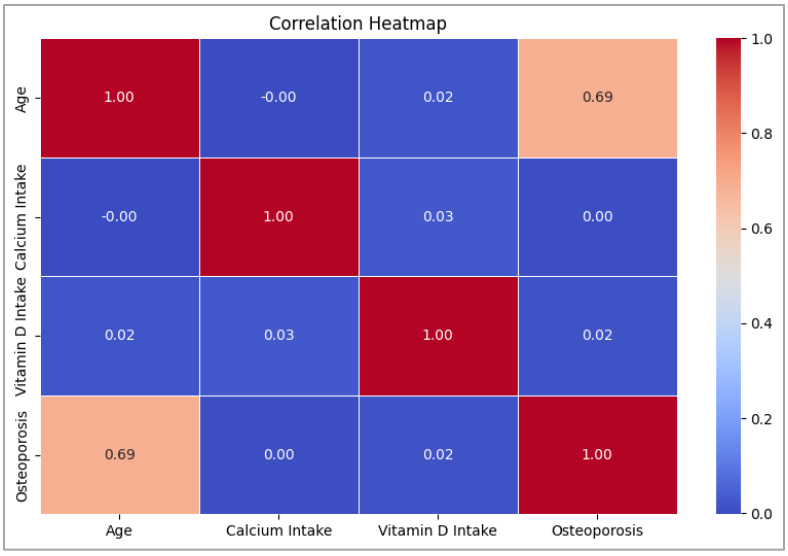

Figure 5

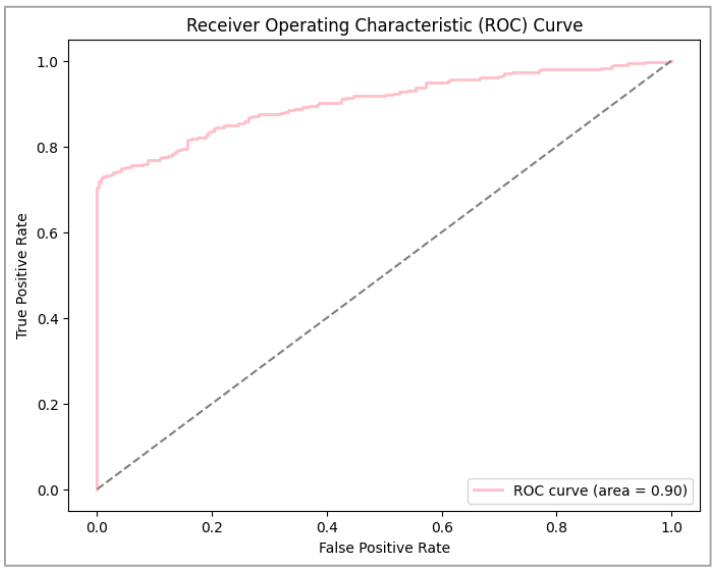

Figure 6

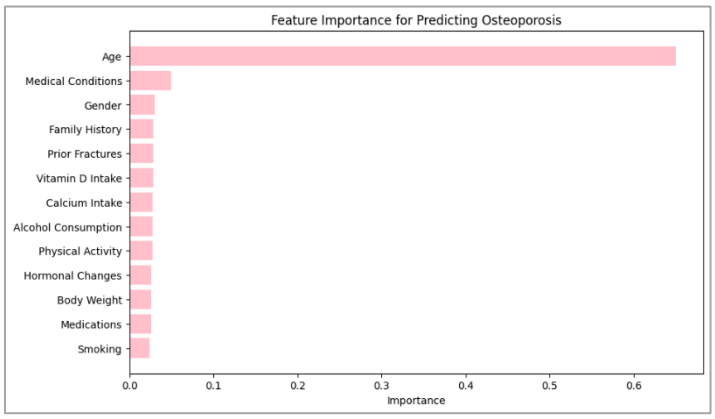

Figure 7