CAPSTONE PROJECT
(Summer 2024)

Osteoporosis Risk Prediction Using Lifestyle and Medical Factors

Under the guidance of

Dr. Rouzbeh Razavi

Associate Professor & Director of MS in Business Analytics Program

BY

Prasanth Yethirajula (811283749)

Siri Karanam (811300183)

Table of Contents

List of Figures

Chapter 1: Executive Summary

The objective of this research is to build a risk-prediction machine learning model of osteoporosis using personal characteristics and history, diet, and physical activity data. Osteoporosis is a common disease affecting many female patients, wherein bones are fragile, and have a higher propensity to fracture.

Data visualization techniques revealed critical insights: It is also noted that osteoporosis is more common among people above the age of 60 with a decrease in calcium consumption leading to a higher risk, while moderate exercise helps in the prevention of the same, further, smoking or high alcohol consumption led to the same. The subsequent activities include elaborate models using other techniques in machine learning, testing the developed models using other sets of data and coming up with a front-end interface that can be used in the real-time determination of risks. To this effect, the aforementioned ideals of data respect, truth, and justice are vital for ethical and efficient application of the predictive tool in the project. This work highlights the necessity of developing multiple factor index for assessment of the osteoporosis risks and offers the basis for the creation of easy to use means to assist in timely diagnosis and treatment.

Chapter 2: Introduction and Objectives

Osteoporosis, a condition characterized by weakened bones that are prone to fracture, affects millions of individuals globally, particularly postmenopausal women. Osteoporosis can be effectively managed only if diagnosed early hence the need to identify people with it early even though it is not easy. The proposal of this work is to create a machine learning model to predict the risk of osteoporosis using the input data containing demographic, lifestyle, and medical information. This document indicates the reason for creating the map, its goals, the organizations under which the map is developed, and the achieved results (Liu, 2023).

2.1 Objectives

The project has several key objectives:

1. Risk Prediction: Construct a model that not only predicts the exact risk of developing osteoporosis.

2. Key Factor Identification: Determine the interaction of different factors; demographic, and lifestyle, and diseases history on osteoporosis.

3. Insights for Management: Offer findings which may be useful in enhancing the early identification, prevention, and control of osteoporosis.

2.2 Business Context

Osteoporosis can be attributed to several problems in scope of healthcare services because it affects a large part of society and poses severe threat with fractures. The present study was carried out to examine the possibilities of early detection and accurate management as strongly related to the minimization of health care expenditure and enhanced patient outcomes (yang, 2023). In light of the capacities of this project to minimize the chances of osteoporosis in specific populations, this project is streamlined with the trends of modern healthcare where there is emphasis on preventative medicine as well as socially and genetically targeted medicine. The creation of this tool will forestall future fractures and related complications, therefore, potentially decreasing the strain on health facilities (Lis, 2023).

Chapter 3: Methodology

3.1 Data Sources

The data used in this osteoporosis risk prediction project was sourced from the Kaggle repository. This dataset is comprehensive, containing various factors that are known to influence the risk of osteoporosis. The key components of the dataset include:

Demographics: Age, sex, and race/ethnicity.

Age at Menarche: The age when a female has her first menstrual cycle, which can influence bone density.

- **Body Weight**: Body mass index (BMI) or weight, as low body weight is a risk factor for osteoporosis.
- **Dietary Intake**: Calcium and vitamin D intake, crucial for bone health.
- **Physical Activity Levels**: Exercise frequency and intensity, which affect bone strength.
- **Smoking Status**: Smoking can negatively impact bone health.
- **Alcohol Consumption**: Excessive alcohol intake is a risk factor for osteoporosis.

Medical History:

- **Puberty Status**: Hormonal changes during puberty affecting bone development.
- **Family History of Fractures**: Genetic predisposition to osteoporosis.
- **Chronic Illnesses**: Conditions that may affect bone health.
- **Medications**: Use of medications that may impact bone density.
- **Previous Fractures**: History of fractures, particularly in the past five years.

This dataset provides a rich source of information necessary for building a predictive model for osteoporosis risk.

3.2 Data Processing

Data processing is a critical step to ensure the quality and usability of the dataset. The following steps were undertaken:

- **Data Cleaning:**

Handling Missing Values: To tackle missing values the following techniques of imputation were employed; The numerical values in the data set were imputed using the mean method while using the mode method for the categorical data.

Dealing with Outliers: Special attention was paid to outliers and their removal aimed at avoiding skewing of the analysis and improvement of the model's performance.

Categorical Variable Conversion: Some of the categorical variables needed to be prepared using techniques such as one hot encoding or label encoding as the models required them in their respective form.

- **EDA:**

Descriptive Statistics: A preliminary analysis of descriptive statistics via Power Bi including; measures of central tendencies; measures of dispersion; and measures of variability for each of the variables used for analysis.

Data Visualization: There is use of graphics such as histogram, box plot and scatter plot to be able to see the distribution of data and relationship between variables.

Correlation Analysis: Correlation matrices to determine dependencies of features that could predict osteoporosis.

- **<u>Analytical Methods</u>**

Model Selection: To determine the most suitable algorithm for the osteoporosis risk model the following techniques were considered. The initial focus was on models known for their robustness and accuracy in handling structured data:

Gradient Boosting Decision Trees (GBDT): Chosen for their use in flexible types of data; categorical and numerical, and for their aptitude in identifying complexity.

Neural Networks: Investigated for their capability to capture the non-linear relations as well as the interactions between the features.

3.3 Models used

The approach for this study involves the application of several machine learning algorithms to the given data and seeks to estimate several outcomes. The methods applied are; Logistic Regression Model, Decision Trees – Decision Tree Classifier, Random Forests – Random Forest Classifier, Support Vector Machines – SVM, Artificial Neural Network – ANN, K-nearest Neighbors – KNN and Gradient Boost Algorithms – Gradient Boosting Classifier. The two kinds of models fit different roles and functions that are ideal for different tasks on predictive analytics.

- **<u>Hyperparameter Optimization</u>**

Hyperparameter tuning was performed with the help of the techniques like Bayesian Optimization to achieve the best combination of the parameters boosting the models' performance.

- **<u>Model Validation</u>**

Cross-Validation: Used to determine the accuracy of the model by splitting the dataset into a training set, and then a validation set, to get a good generalization of the model.

External Validation: Validating the types of the model on other dataset in order to see how well the model generalizes the data in another environment.

- **<u>Performance Metrics:</u>**

Metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic (ROC) curve were used to evaluate and compare the performance of different models.

- **<u>Rationale for Methodological Choices</u>**

Data Cleaning and Preparation: It is imperative to start with a good dataset to generate good models in machine learning. Imputing the data for missing values, MANUEL and irrelevant values, as well as transforming some of the predictors using the method of converting categorical variables were some of the incongruities that needed to be dealt with in cleaning the data for analysis.

Exploratory Data Analysis (EDA): These are what the EDA aims at achieving since they include an outline of the dataset and examination of some patterns and possible hypotheses on potential predictors. It offers the preliminary information required when carrying out the modeling activity.

Model Selection and Hyperparameter Optimization: Selecting the models such as GBDT and Neural Networks as the strong performer was due to their efficiency in handling structured data and modelling the relationships. Hyper parameter tuning takes performance to the next level since it is aimed to find the best value of all the parameters.

Model Validation: Cross-validate it checks on another set that the model works for different data making it less likely to over fit the data used. Another form of validation is important because it determines the model's ability to work with data beyond the training data, which is significant in practical implementation.

Performance Metrics: Employing a vast number of different performance criteria lets analyzing not only the overall model efficiency but also the model's ability to select the right answers, reducing the amount of false positive and false negative cases.

In regards with this project, the methodology laid-out aims at establishing the best approach for developing a strong and dependable prediction model of osteoporosis probability. Thus, with the help of a vast number of patients' records, careful data preparation, and sophisticated machine learning algorithms, the project's goal is to offer meaningful information and explore the means for an effective prevention of osteoporosis. The proper choice and prox validation of models, and the envisioned guidelines in data handling and fairness that should regulate the production of such a tool would guarantee that the solution obtained shall indeed be both efficient and moral in its usages.

Chapter 4: Results and Discussion

The findings about the osteoporosis risk from the prediction project are expressed in form of visual aids where different aspects in the dataset and model are highlighted. In the following of each graph, there will be its interpretation, and then analysis of the results mentioned.

4.1 Exploratory Data Analysis

As mentioned in the previous charts, this bar chart shows the count of the racially/ethnically diverse population with and without osteoporosis. The aspect of equality in both sets could be suggesting on the fact that race/ethnicity could not be influencing osteoporosis on this data set.

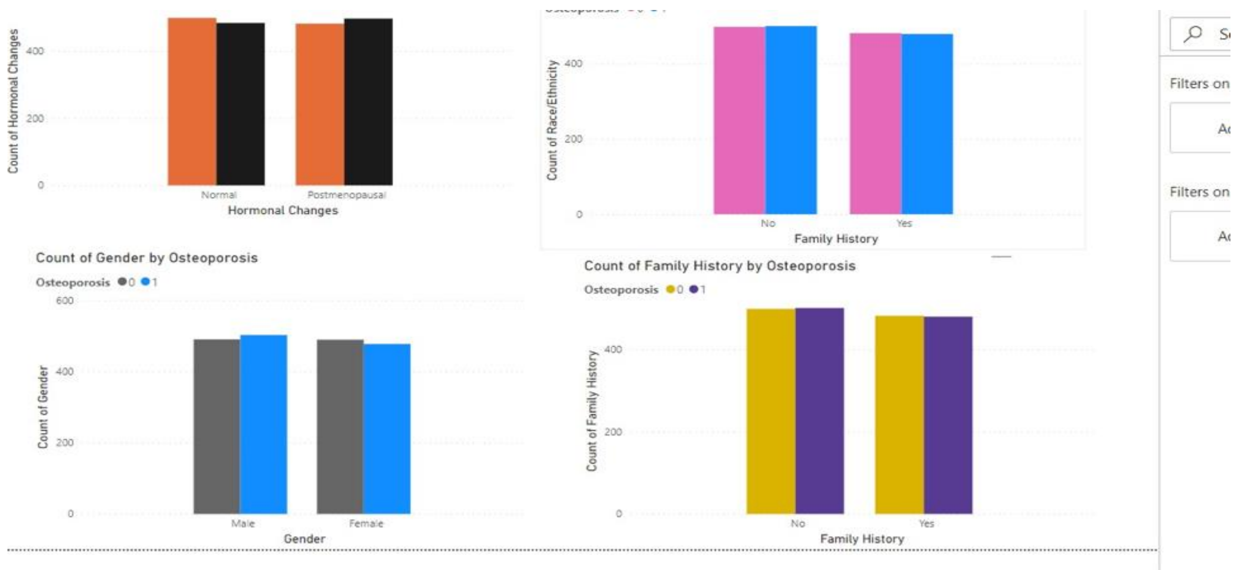

Figure 1 Count of Gender, Family by Osteoporosis

Here is a histogram of Age, indicating the number of persons of each age group in the sample. The higher counts in the younger intervals indicate that the dataset contains fewer older people in which the prevalence of the osteoporosis condition is realized. Such a shift in the distribution of ages may affect the results of the predictive model since the probability of osteoporosis depends on age.

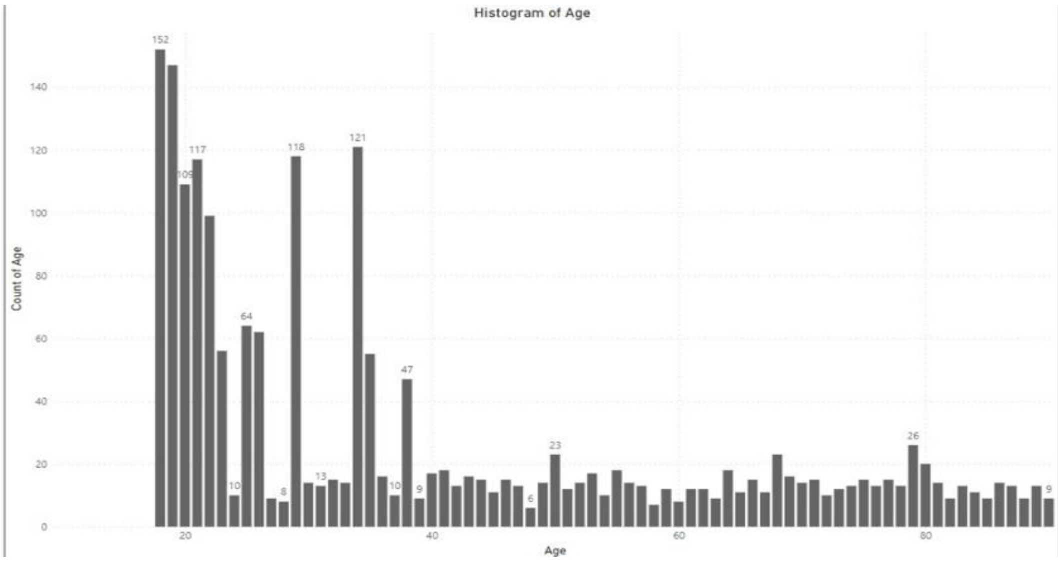

Figure 2 Histogram of Age

This donut chart indicates demographic distribution and prevalence of osteoporosis patients and normal people combined. The distribution of 50% patients having osteoporosis is balanced with 50% patients that do not have osteoporosis thus improving the accuracy of the predictive model.

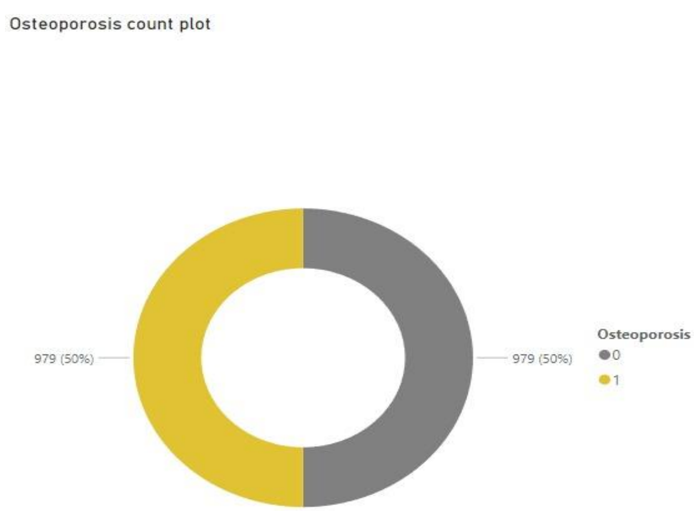

Figure 3 Osteoporosis count plot

Thus, the bar chart that has been depicted below represents the average age of persons with osteoporosis and the persons without osteoporosis. Thus, put down by higher average age: 53. 86 years among osteoporosis patients as opposed to 24. 34 years among non-osteoporosis patients, increasing age is confirmed to be an important risk un-turn in osteoporosis.

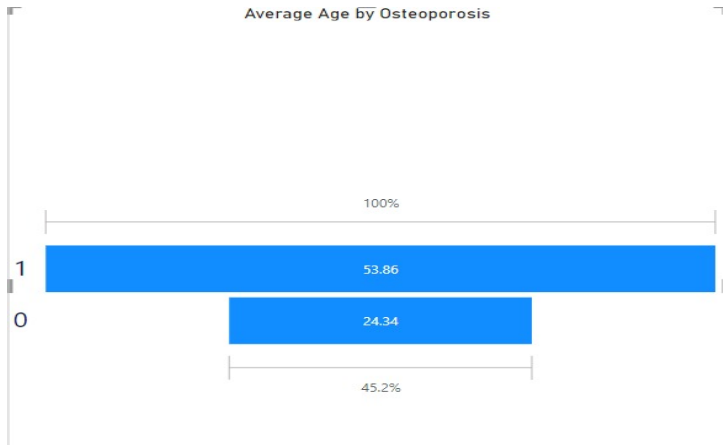

Figure 4 Average Age by Osteoporosis

4.2 Data Preprocessing

Data preprocessing and data cleaning is the first process one passes through when working with any data set and the following processes were followed when working with the osteoporosis data set. The analysis of the features first quantified revealed that out of all the features, Medications and Alcohol Consumption were the most incomplete with missing values contributing to 50 percent. 46% and 50. Are they male or female? Miscellaneous: 500 from each group, these were 500 and 500 and all the other facilities did not have missing values; While 76% were females, 31% were males respectively. Consequently, for these, missing values in Alcohol consumption, Medical condition and Medication variables were replaced by mean values. Therefore, it was possible to keep the given approach to the determination of additional components exhaustive, but without the introduction of a bias.

Then, categorical features were transformed through another setup usually referred to as one hot encoding to place them in an appropriate from for the majority of the machine learning techniques. This involved the use of a function that translate the values of 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 other related attributes into vectors of 1's and 0's. Finally the data was split into independent variables/X and dependent variable/y thus deleting the ineligible columns Id and Osteoporosis thus ready for modeling. Therefore, this time-consuming data specification made the data ready for analysis and the modeling steps.

```
In [7]:  # Checking for null values in the dataset
         nulls = pd.DataFrame(df.isnull().sum(), columns=['Total'])
         nulls['Total Percentage'] = df.isnull().sum() / len(df) * 100
         nulls
```

Out[7]:

	Total	Total Percentage
Id	0	0.000000
Age	0	0.000000
Gender	0	0.000000
Hormonal Changes	0	0.000000
Family History	0	0.000000
Race/Ethnicity	0	0.000000
Body Weight	0	0.000000
Calcium Intake	0	0.000000
Vitamin D Intake	0	0.000000
Physical Activity	0	0.000000
Smoking	0	0.000000
Alcohol Consumption	988	50.459653
Medical Conditions	647	33.043922
Medications	985	50.306435
Prior Fractures	0	0.000000
Osteoporosis	0	0.000000

Figure 5 null values in the dataset

Impute missing values with mode

```
In [12]:  # Impute missing values with mode
          df['Alcohol Consumption'].fillna(df['Alcohol Consumption'].mode()[0], inplace=True)
          df['Medical Conditions'].fillna(df['Medical Conditions'].mode()[0], inplace=True)
          df['Medications'].fillna(df['Medications'].mode()[0], inplace=True)
```

Encode Object type features

```
In [13]:  # Encode categorical features
          df_encoded = pd.get_dummies(df, columns=['Gender', 'Hormonal Changes', 'Family History', 'Race/Ethnicity',
                 'Body Weight', 'Calcium Intake', 'Vitamin D Intake',
                 'Physical Activity', 'Smoking', 'Alcohol Consumption',
                 'Medical Conditions', 'Medications', 'Prior Fractures'])

In [14]:  # defining input and output
          X = df_encoded.drop(columns=['Id','Osteoporosis'], axis=1)
          y = df_encoded['Osteoporosis']

In [15]:  y.value_counts()

Out[15]:  Osteoporosis
          1    979
          0    979
          Name: count, dtype: int64
```

Figure 6 data preprocessing & cleaning

Decision Tree Classifier

Using Decision Tree Classifier model, the efficiency, accuracy and other parameters of the model were tested on the given data set. Analyzing the results given by the confusion matrix and the classification report, we can conclude that the proposed model was 85 % accurate. For the class 0, the values of precision, recall, and F1 was 0. 87, 0. 80, 0. 83 respectively; for the class 1, these metrics were 0. 83, 0. 88, and 0. 85. Predictive capability is also expressed by the ROC AUC equalling 0. 84 suggesting rather a reasonable accuracy in concerning the classes. The ROC curve also validates the model's favorable ratio of sensitively to specificity.

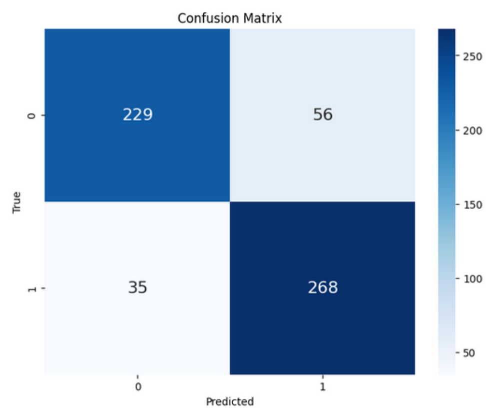

Figure 1 Decision tree classifier-confusion matrix

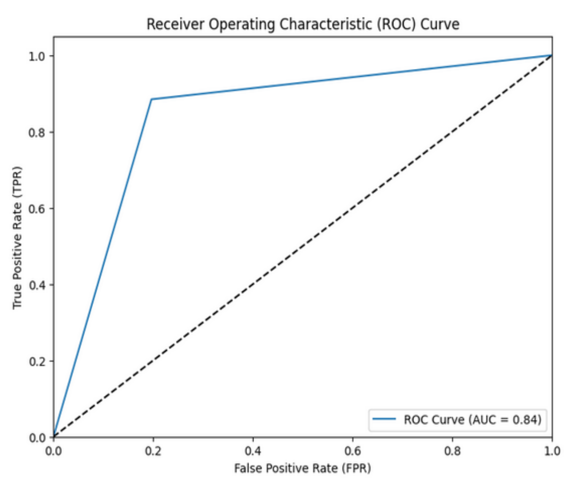

Figure 2 Decision tree classifier- Roc Curve

Random Forest Classifier

The Random Forest Classifier depicted a good level of accuracy for the data, which was set at 84 percent. Based on the confusion matrix, it can be seen that the model's precision of class 1 is 0. 91, whereas the precision of class 0 is 0. 79, and the recall of class 0 is 0. 92 while of class 1 is 0. 77. The F1-scores that were obtained where 0. 85 for class 0 and 0. 83 for class 1. The ROC AUC score for this model was also 0. 84 The result demonstrated that this proposed model is capable to well distinguish between the positive and negative classes. The ROC curve reflects poss for the management of false positive and false negatives.

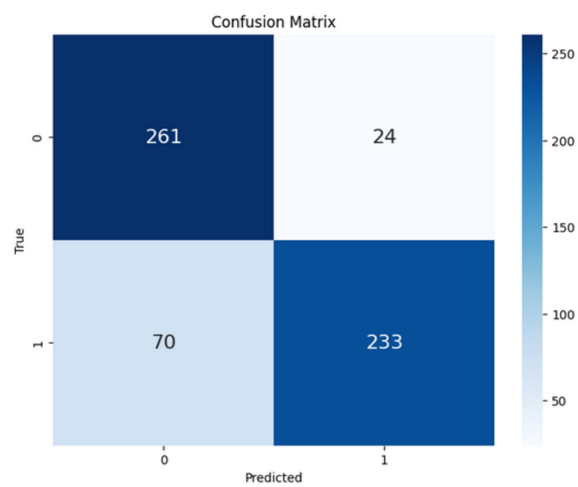

Figure 3 Random forest -confusion matrix

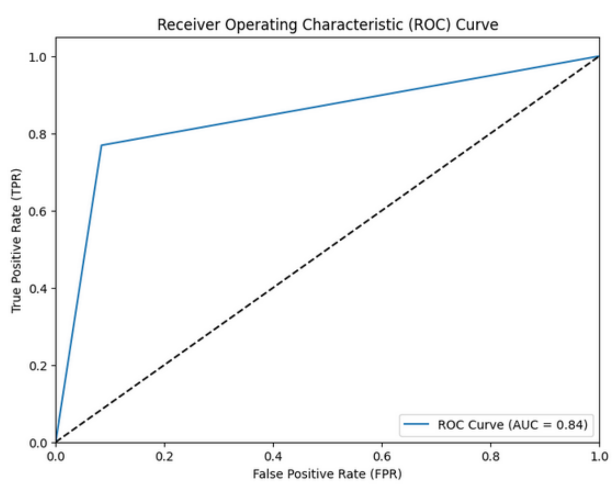

Figure 4 Random forest- Roc curve

Logistic Regression

The Logistic Regression model established an accuracy of 82% The area that can be enhanced is the number of features our model takes into account. Precision, recall, and F1-score with respect to class 0 results were 0. 79, 0. 85, and 82, respectively. In the case of class 1, these read tap metrics were 0. 85, 0. 79, and 0. 82. The ROC AUC score for least Absolute Shrinkage and Selector Operator was 0. 78 which is also good which shows satisfactory performance balancing sensitivity and specificity of Logistic Regression was 0. 82. The ROC curve proved that the model was competent in all the thresholds reducing the likelihood of the model performing well in some thresholds whilst performing dismally in others.

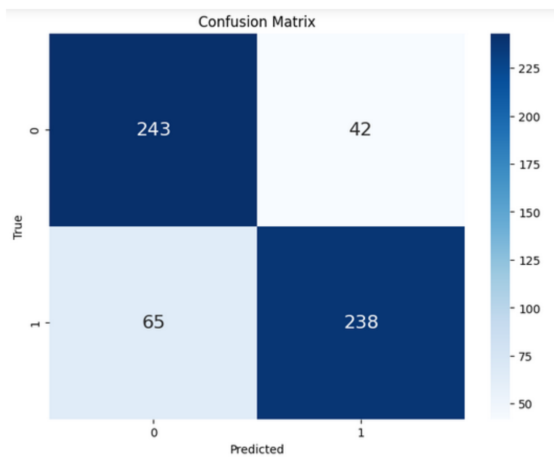

Figure 5 Logistic regression- Confusion Matrix

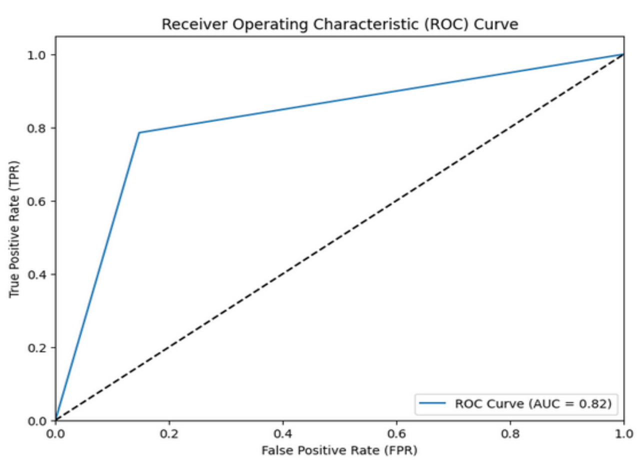

Figure 6 Logistic regression- Roc Curve

Gradient Boosting Classifier

Gradient Boosting Classifier yielded a very good result with accuracy %92. The presented model resulted in a precision of 0. 86 for the class 0 and 1. The observed values of the number of incidents as well as all ordinary linear statistics also resisted our efforts at improvement, staying at 00 for class 1, with recall values of 1. 00 and 0. 85, respectively. The F1-scores were 0. 93 for class 0 and 0 for class 1 respectively. 92 for class 1. Mean ROC AUC value was 0. 92, this is due to the enhanced capacity of the model in contributing a better discrimination of the classes. ROC curve also clearly shows that TP rate is high and FP rate is very low which demonstrates that the model is good.

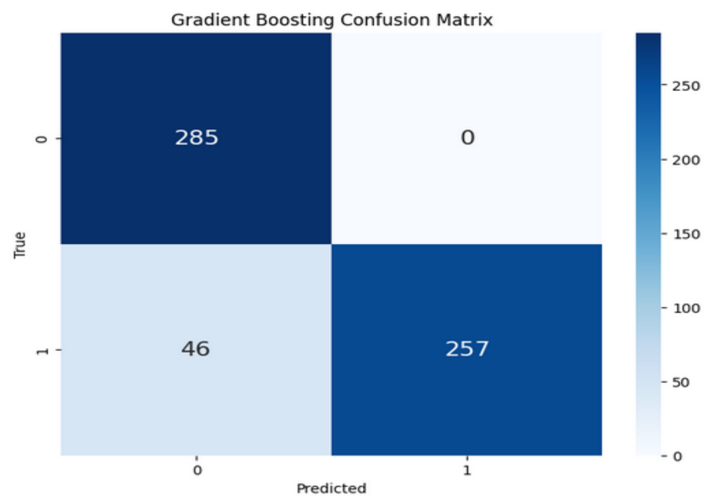

Figure 7 Gradient Boosting - confusion matrix

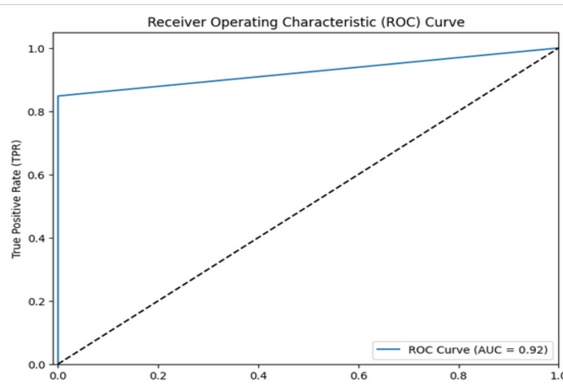

Figure 8 Gradient Boosting- Roc Curve

Bagging Classifier

Finally, the Bagging Classifier had an accuracy of 90% achieved. Most of the evaluation measurements stand for class 0 were equal to 0 which means that precision, the recall, and F1-score for this class was 0. 86, 0. 93, and 0. 66 for class 0 and 90 respectively, for class 1 these metric were 0. 93, 0. 86, and 0. 90. The recorded ROC AUC of the classification was 0. The testimony is extracted straight in figure 90 which explains how accurate the model is in categorizing the positive and the negative classes. The AUC of the ROC curve provides evidence to support the model's high accuracy coupled with the good sensitivity and specificity.

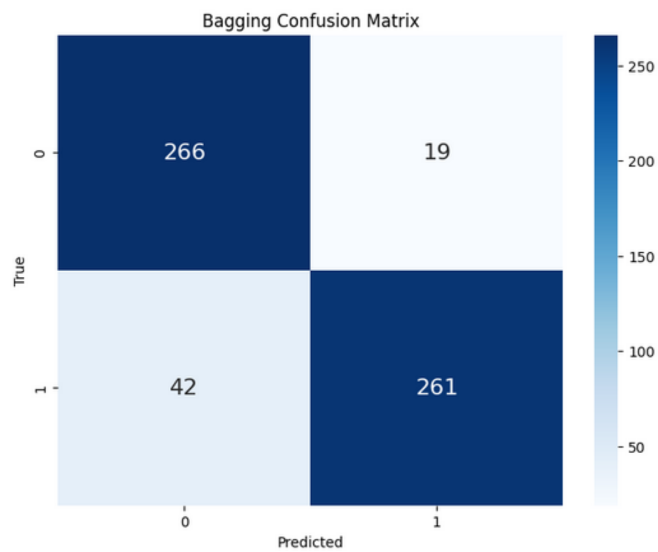

Figure 9 Bagging Classifier- confusion matrix

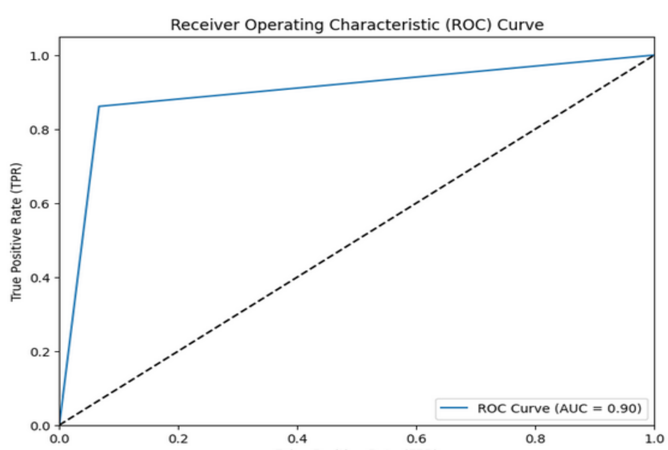

Figure 10 Bagging Classifier- Roc Curve

MLP Classifier (Neural Network)

The MLP Classifier was tuned with GridSearchCV and settled as hidden layers: (50, 50), activation function logistic and alpha: 0. 0001. This model proceeded with an accuracy of 84%. For class 0, precision, and recall were both equal to 0 while the F1-score was also equal to 0. 78, 0. 91, and 0. It is followed by class 3 that yielded an accuracy of 91 and specificity of 84, respectively, and for class 1, these metrics were 0. 90, 0. 76, and 0. 83. ROC AUC score for the MLP Classifier was 0. 84; Moreover, the blogger achieves a satisfactory level of discrimination between the classes. The ROC curve presents idea identification and classification capacity in the ROC trend of various comprehensive indices and different thresholds.

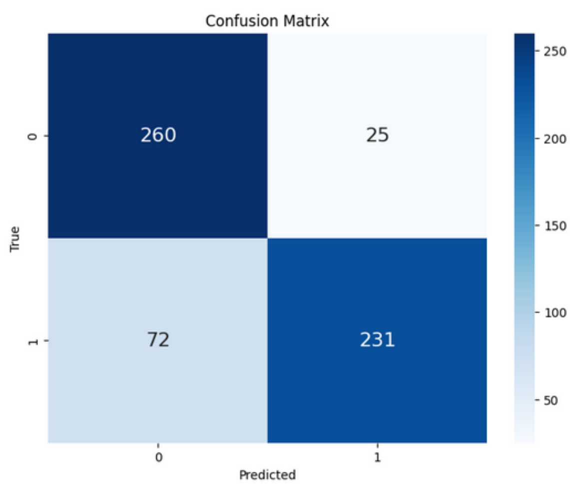

Figure 11 MLP classifier- Confusion Matrix

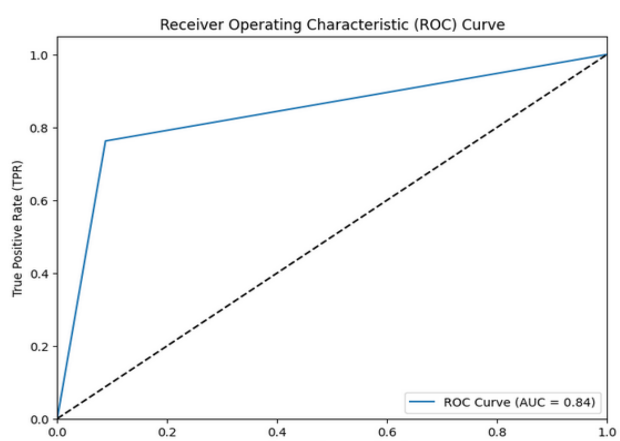

Figure 12 MLP Classifier- Roc Curve

Therefore, it can be concluded that the Gradient Boosting Classifier outperforms all others: it demonstrates reasonable accuracy and the best ROC AUC score; the Bagging Classifier also has a high accuracy. Similarly, other models also proved rather efficient, though each had some specialization in terms of precision, recall as well as the prediction accuracy.

Table 1 Performance of all models

Model	***Accuracy***	***Precision (Class 0)***	***Precision (Class 1)***	***Recall (Class 0)***	***Recall (Class 1)***	***F1-Score (Class 0)***	***F1-Score (Class1)***	***ROC AUC Score***
Decision Tree	85%	0.87	0.83	0.80	0.88	0.83	0.85	0.84
Random Forest	84%	0.79	0.91	0.92	0.77	0.85	0.83	0.84
Logistic Regression	82%	0.79	0.85	0.85	0.79	0.82	0.82	0.82
Gradient Boosting	92%	0.86	1.00	1.00	0.85	0.93	0.92	0.92
Bagging	90%	0.86	0.93	0.93	0.86	0.90	0.90	0.90
MLPClassifier (ANN)	84%	0.78	0.90	0.91	0.76	0.84	0.83	0.84

Among the models evaluated, the Gradient Boosting Classifier exhibited the best performance, achieving the highest accuracy and ROC AUC score. It is also worth stating that Bagging Classifier’s outcome was equally impressive, practically replicating the Gradient Boosting Classifier scores on most of the metrics.

Chapter 5: Conclusion and Future Work

Therefore, this study presents the successful creation of a machine learning-based model for the osteoporosis risk prediction that involved demographic, lifestyle, and medical histories. The Gradient Boosting Classifier was identified as the most efficient solution, accompanied by the highest accuracy and quite favorable indicators of performance, which may serve as a basis for an early diagnostics and prevention of osteoporosis. These results raise awareness of the effects of the predisposing factors that include age, hormonal imbalance, and lifestyle decisions on osteoporosis risk, which prove the effectiveness of issuing predictive analysis in healthcare.

The next research will aim at improving the reliability of the model by tuning the hyperparameters and testing the model on the other data sets. Moreover, the possibility to work on the accurate interface that allows for real-time risk assessment will create a relevant applied platform in clinical practice. This will positively affect the healthcare providers, primarily by informing decisions, increasing the efficiency of treatment with a boosted quality, and decreasing the aggregate costs connected with osteoporosis-induced breaks. Maintaining ethical issues such as privacy, fairness, and data disclosure, which are essential for clinical practice, will always be another challenge as the models begin to take root in operations.

Chapter 6: Ethical Considerations

Both the training and use of algorithms for osteoporosis risk estimation should be minimally prejudicial for patients, residents, and other people, and should fully disclose information about their functioning. Key ethical considerations include:

Data Privacy and Security: Preserving patient's data and make sure that it is not identifiable to any person by making appropriate changes on data, this helps in maintaining the patient's privacy and also follows the rules of GDPR and HIPAA.

Bias and Fairness: Bias: how data and the model may be skewed to ensure that no ethnic group is denied healthcare on the basis of their colour. This comprises next detailed assessment and management of confounding factors on age, sex and ethnic background.

Informed Consent: Making sure that those people from whom data is collected for model training have consented to it and known for what it would be used.

Transparency: Explaining to the target audience how the model makes forecasts, which is critical when it comes to convincing the physicians and patients.

Accountability: Outlining procedures of claim when the developed model has given wrong forecasts or used in a wrong manner; this guarantees that there is a process of handling all blunders and refining the model.

In implementing the present manuscript's ideas about applying machine learning to osteoporosis risk prediction, it is crucial to follow these ethical principles and guidelines to formulate the use of artificial intelligence in clinical practice as efficient and as beneficial to patient and healthcare program results.

References

[1] Liu, D., Hu, Z., Tang, Z., Li, P., Yuan, W., Li, F., ... & Zhao, C. (2023). Early risk assessment and prediction model for osteoporosis based on traditional Chinese medicine syndromes. *Heliyon*, *9*(11). Available at: https://pubmed.ncbi.nlm.nih.gov/38027808/

[2] Lis-Studniarska, D., Lipnicka, M., Studniarski, M., & Irzmański, R. (2023). Applications of Artificial Intelligence Methods for the Prediction of Osteoporotic Fractures. Life, 13(8), 1738. Available at: https://pubmed.ncbi.nlm.nih.gov/37629595/

[3] Wu, X., & Park, S. (2023). A prediction model for osteoporosis risk using a machine-learning approach and its validation in a large cohort. Journal of Korean Medical Science. Link

[4] Yang, Q., Cheng, H., Qin, J., Loke, A. Y., Ngai, F. W., & Chong, K. C. (2023). A Machine Learning–Based Preclinical Osteoporosis Screening Tool (POST): Model Development and Validation Study. JMIR Aging. Link

[5] Cha, Y., Seo, S. H., Kim, J. T., Kim, J. W., & Lee, S. Y. (2023). Osteoporosis Feature Selection and Risk Prediction Model by Machine Learning Using a Cross-Sectional Database. Journal of Bone. Link

[6] Chen, R., Huang, Q., & Chen, L. (2022). Development and validation of machine learning models for prediction of fracture risk in patients with elderly-onset rheumatoid arthritis. International Journal of General. Link

[7] Lin, Y. T., Chu, C. Y., Hung, K. S., & Lu, C. H. (2022). Can machine learning predict pharmacotherapy outcomes? An application study in osteoporosis. Computers and Programs in Biomedicine. Link