RISK PREDICTION FOR OSTEOPOROSIS

A Machine Learning approach to predict the risk with the influence of lifestyle factors

University of Pittsburgh- School of Health and Rehabilitation Sciences

2251 HI 2453 SEC1095 MACHINE LEARNING IN HEALTH SCIENCES

BY:

Sree Gayatri Anusha Mylavarapu

Guided By : Dr. Leming Zhou, PhD, DSc

Osteoporosis

Osteoporosis is a skeletal disease which is metabolic in nature and is characterized by low bone density and it affects the bone and causes the deterioration of the bone tissues making the bones brittle and eventually cause fracture.

- The hips , vertebrae and the wrists of arms are the most common sites which get affected with osteoporosis.
- It is prevailing to be one of the major public health concerns among populations and this is caused mostly due to lifestyle factors and changes in lifestyle.
- Osteoporosis is more commonly in females more than men
- Increased age, inadequacy of calcium and Vit D, lack of physical activity,
- Idiopathic age related osteoporosis is one of the most common.

Motivation -

This is a disease that is usually undiagnosed and is generally detected until there is a fracture. There are some diagnostic tools like Dual- Energy X-ray absorptiometry that detect only there is bone loss and this critical gap and limitation leads to delayed prediction of osteoporosis. The high costs of this x-ray and limited accessibility exacerbate the challenge.

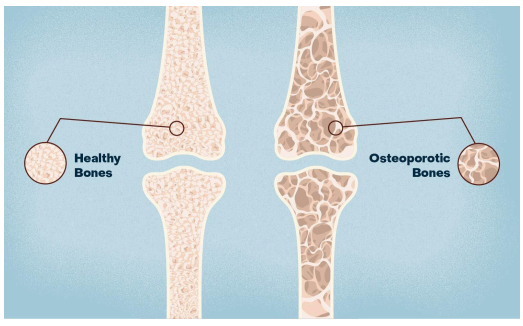
- It affects more than 200 million people globally and preventing this disease is more cost effective than treating it.
- Research has identified critical influences on lifestyle factors like diet, smoking , physical activity, alcohol consumption etc.

- Building a knowledge gap can provide the foundation to find personalised strategies and enhance bone health and prevent osteoporosis.
- In this project I aim to develop a predictive model that can accurately assess an individual's risk of developing osteoporosis majorly based on lifestyle factors.
- I want to identify the features and key influential factors that contribute to the risk of osteoporosis.
- Data analytics and machine learning can show an opportunity to analyse datasets and identify the key predictors and assess the risk levels in heterogeneous populations.
- By applying machine learning methods we can assess a deeper understanding of how lifestyles can influence the early prediction of this disease.
- This project aligns with the principles of preventive healthcare, where the goal is to detect risks and implement interventions before disease progression.
- By focusing on modifiable lifestyle factors, the proposed model has the potential to provide individuals with actionable insights to improve bone health and reduce their risk of osteoporosis. Furthermore, the insights derived from this study can inform public health strategies, guiding educational campaigns and policy decisions to promote healthier lifestyles and reduce the societal burden of osteoporosis.
- This project has the potential to pave the way for future research in osteoporosis prevention and management.
- By identifying the most influential lifestyle factors, the model can guide additional studies to explore causal relationships, interactions with genetic factors, and the long-term effectiveness of lifestyle modifications

- . The results of this project can also inspire similar approaches to studying other chronic diseases influenced by lifestyle factors, amplifying its impact across the healthcare landscape

Objectives-

- The major aim of the project is to design a predictive model which will actually tell the risk of an individual in developing osteoporosis majorly based on their particular lifestyle.
- I want to determine the features and key influential factors associated with the risk of developing osteoporosis
- The model can be used as a screening tool and early intervention is possible
- This report develops a personal plan of action aimed at the prevention of osteoporosis and brittlling of bones.
- This multi-dimensional predictive model can assist us in reducing the risk of Osteoporosis and other chronic bone-related diseases.
- This also analyzes the given dataset and the different factors linked with osteoporosis.
- It designs and evaluates the different machine learning models useful for the prediction of osteoporosis
- I would like to identify major predictors of Osteoporosis.



Description of the current problem and the data set

Osteoporosis ends in weakened bones due to reduced bone mass and tissue deterioration, making fractures much more likely

- It is diagnosed by the arena fitness corporation (WHO) as a widespread public health problem, osteoporosis currently impacts about 200 million people internationally.
- Present diagnostic methods, like DXA scans, often discover osteoporosis most effective after large bone loss has already befell, restricting alternatives for early prevention.
- Additionally, DXA scans aren't broadly handy, highlighting the want for easier, greater reachable screening equipment research shows that life-style alternatives—together with weight loss plan, exercise, smoking, and alcohol intake—play an essential role in bone health. But, how these elements are painted together to persuade osteoporosis risk remains now not completely understood.
- This venture will use machine mastering to investigate numerous lifestyle elements, aiming to fill this information hole and support early osteoporosis prevention efforts.

Data Source **:** This is the dataset I used for this project is https://www.kaggle.com/datasets/amitvkulkarni/lifestyle-factors-influencing-osteoporosis?resource=download

The primary data source for this project is the "Lifestyle Factors Influencing Osteoporosis" taken from kaggle.

This dataset has 1958 rows which are the entities and 16 columns which are the features

```
[1958 rows x 16 columns]
```

Here my target variable is Osteoporosis which shows the presence or absence and 1 implies yes and 0 implies no

Key features of this dataset include:

1. Demographic information: Age, gender, ethnicity
2. Family history of osteoporosis
3. Physical activity levels and types
4. Dietary habits: Calcium intake, vitamin D consumption, protein intake
5. Smoking and alcohol consumption patterns
6. Bone density measurements
7. Body Mass Index (BMI)
8. Menstrual and reproductive history for female participants
9. Medication use, including corticosteroids

- The Id feature in this dataset is a unique identifier in the data set and is used for entry and tracking purposes.
- Age in this data set ranges from 18 - 90 yrs making it a crucial factor in determining risk of osteoporosis.
- Gender can show the influence and progression of osteoporosis.
- Hormonal changes can significantly impact bone health.
- The family history is categorized as yes or no and it is a known risk factor for osteoporosis.
- Race or ethnicity shows any varying risks of osteoporosis.
- Body weight categories of the individual like normal, underweight and overweight can affect the overall bone density.

- Calcium intake is another crucial factor for bone health as it prevents osteoporosis.
- Vit D intake is vital for calcium absorption.
- A regular physical activity sedentary, active or very active
- Smoking is categorized as yes or no. Smoking is a risk for osteoporosis.
- Alcohol consumption in this shows moderate or aggressive.

<u>Methods of my approach and Justification</u>-

As suggested in my initial proposal I have curated a 9-week plan to tackle this project and used machine learning approaches.

<u>Handling Data types and Missing Values</u>

- My dataset includes 3 numerical columns which are ID, Age and Osteoporosis and the other 13 are categorical columns.
- While handling missing values I have observed that there are missing values in Alcohol consumption, Medical conditions and Medications and I have handled them during the stage of preprocessing.

- After analyzing the data I have observed that the target variable is balanced with osteoporosis detected as 1 and no osteoporosis as 0.
- The missing data was handled using the function SimpleImputer from ‘sklearn.impute’.

```
Missing Values Before Imputation:
Id                          0
Age                         0
Gender                      0
Hormonal Changes            0
Family History              0
Race/Ethnicity              0
Body Weight                 0
Calcium Intake              0
Vitamin D Intake            0
Physical Activity           0
Smoking                     0
Alcohol Consumption       988
Medical Conditions        647
Medications               985
Prior Fractures             0
Osteoporosis                0
Cluster                     0
Hierarchical_Cluster        0
dtype: int64
```

```
Missing Values After Imputation:
Id                        0
Age                       0
Gender                    0
Hormonal Changes          0
Family History            0
Race/Ethnicity            0
Body Weight               0
Calcium Intake            0
Vitamin D Intake          0
Physical Activity         0
Smoking                   0
Alcohol Consumption       0
Medical Conditions        0
Medications               0
Prior Fractures           0
Osteoporosis              0
Cluster                   0
Hierarchical_Cluster      0
dtype: int64
```

- I have performed one hot encoding to variables like gender , hormonal change, family history, race/ethnicity , body weight, calcium intake, Vit-D intake etc and were encoded using one HotEncoder from ‘sklearn.preprocessing.’ This is done to achieve a numerical input.
- Then I have scaled the features ensuring that all the features were on the same scale which can improve the performance of the model.
- I have ensured that the feature preprocesses appropriately.

Machine Learning model selection -

For my approach and this dataset I have use different machine learning models

- I have used Random Forest Classifier to test the ability for it to handle complex data and interactions between variables and avoid overfitting.
- Logistic Regression in this model shows the interpretability as well as effectiveness in this binary classification problems.

- Support vector finds the optimal hyperplane as it classifies the feature space.

To use sophisticated modelling I have also done hyperparameter tuning

- This tuning combines different hyperparameters and finds the optimal set and maximises the model performance
- The data has been split into training and testing sets and this ensures that the training on a portion of the data provides a much more accurate performance.
- I have created a pipeline using the function 'sklearn.pipeline' which simplifies the workflow and ensures that preprocessing is constantly applied to both the training and testing data.

Justification for my approach

- Imputation and handling of missing values is important to make sure that bias is reduced and all the rows are used.
- One-hot encoding converts categorical variables into a numerical format so that machine learning models can process.
- Scaling is done to provide stability
- Random Forest is chosen as it is a robust model and handles complex data
- Logistic regression and support vector are used for optimisation and interpretability
- Hyperparameter tuning is done to optimize the model and then the data is trained and tested.
- Further analysis in this model gives an accuracy, precision, recall, F1 score and ROC-AUC score to evaluate the models performance and these metrics will distinguish the differences between classes.

- Cross validation in this model ensures that they are evaluated on multiple subsets of data showing robust performance and reducing the risk of overfitting.

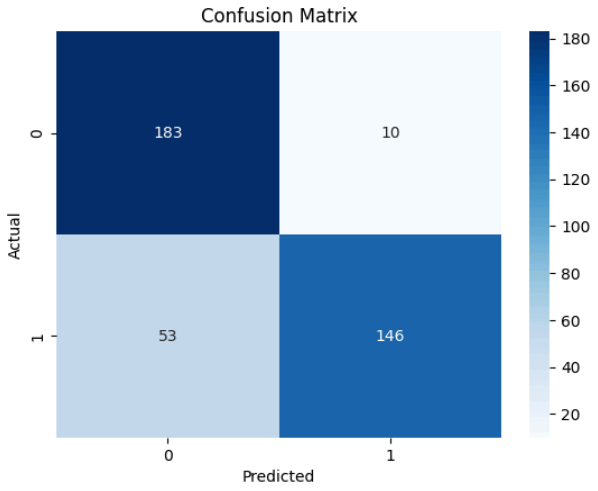
Evaluation -

Evaluating the performance of machine learning is crucial to know the accuracy, reliability and generalization.

This assesses the results and shows the prediction of osteoporosis.

- Data preparation and model training shows the loading , description and information of the dataset as well as handling the missing values.
- I have selected my features and given the target variable as ‘y’ and features as ‘x’ and dropped the column ‘Id’ and ‘Osteoporosis’ as it is used as target.
- Data preprocessing is done and imputation and oneHotencoder is applied where necessary and training and testing of the data is also done.
- The Accuracy score measures the proportion of the instances that are correct in the test set.
- I have also created a metric report showing the precision, recall, F1 score

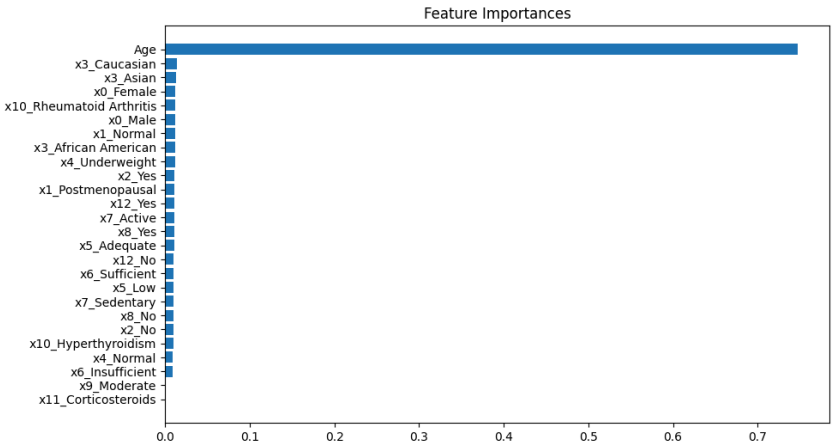
  This provides a detailed report on the precision, recall, F1 score, and support for each class.
- The confusion matrix in this evaluates the performance of the model and shows the true positives, false positives, true negatives and false negatives.



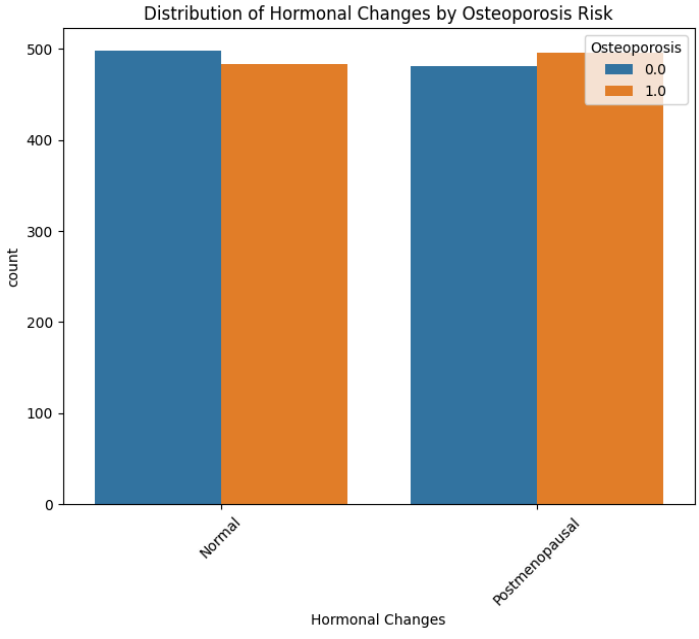
The confusion matrix indicates that the model performs well overall, achieving a high accuracy of 87% and precision of 94%, meaning most predictions for osteoporosis cases are correct. However, the recall for detecting osteoporosis is lower at 73%, with 53 false negatives, suggesting the model misses some actual cases of osteoporosis. This trade-off highlights the model's strength in minimizing false alarms (false positives) but reveals a limitation in capturing all at-risk patients. To improve its utility in healthcare, particularly where identifying all osteoporosis cases is critical, the model's recall could be enhanced by adjusting the decision threshold or employing techniques like oversampling to address potential class imbalance.

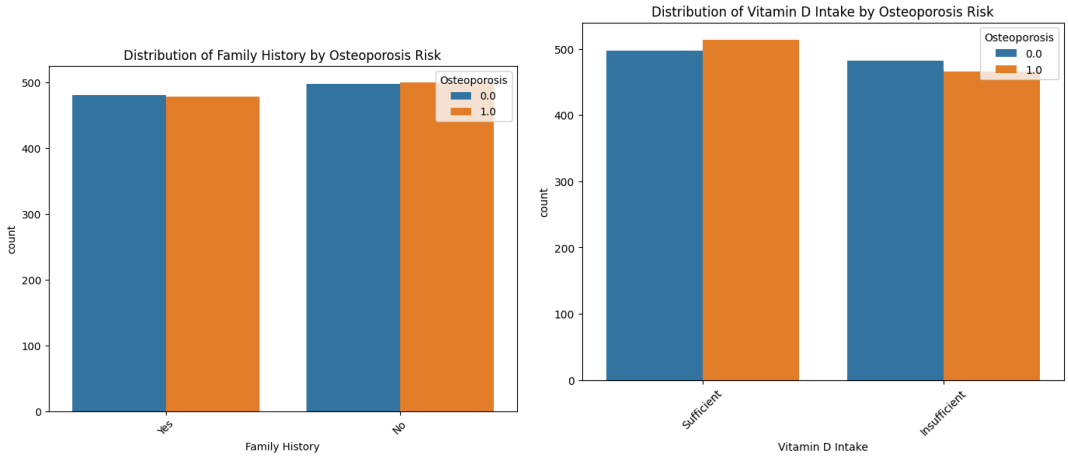
- The F1 score is the means of precision and recall.
- The ROC-AUC curve measures the model's ability to distinguish between classes.
- Cross validation is also done to show the evaluation of performance of the model.

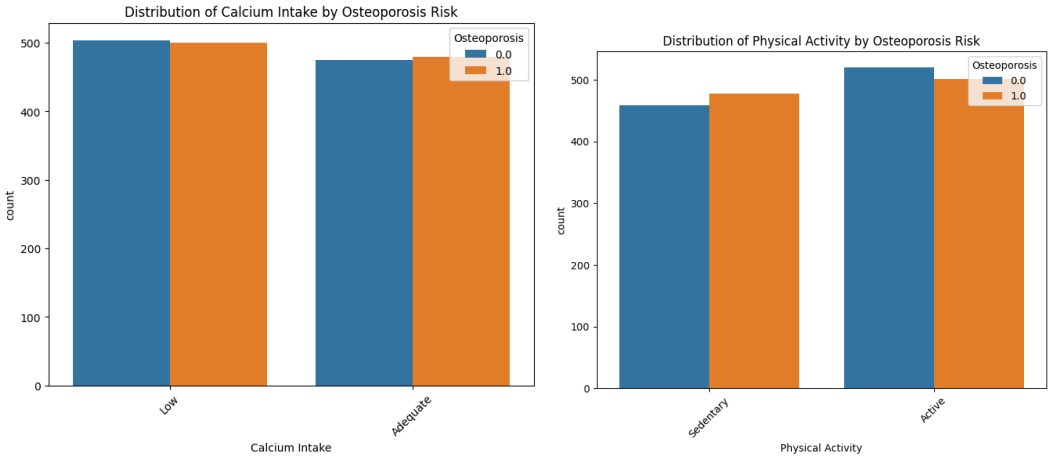
This is a bar graph I visualized that shows features that are important



The x-axis displays the importance scores, which indicate how significantly each feature contributes to the model's predictions, while the y-axis lists the features. The chart highlights that **Age** is the most critical predictor, showing a significantly higher importance score compared to other variables. Features such as ethnicity (e.g., Caucasian, Asian), gender, conditions like rheumatoid arthritis and hyperthyroidism, and lifestyle factors like activity level for postmenopausal status also contribute to the model but with much lower importance.







These are some more graphical representations that show presence or absence of osteoporosis.

Results -

- Based on the machine learning projects and results obtained from various techniques like deep learning, Naive Bayes and performed statistical tests T-tests and the metrics provides models performance and insights.

Model metrics -

- Accuracy Score - This metric measures the proportion of correctly classified instances. For example, an accuracy score of 0.83 indicates that the model correctly classified 83% of the test instances.

- ROC-AUC Score - Measures the model's ability to distinguish between classes. An ROC-AUC score of 0.85, for example, indicates that the model is good at distinguishing between the two classes.
- Confusion Matrix Analysis
- **True Positives (146):** Cases correctly identified as osteoporosis.
- **True Negatives (183):** Cases correctly identified as not having osteoporosis.
- **False Positives (10):** Cases incorrectly predicted as osteoporosis (false alarm).
- **False Negatives (53):** Cases where osteoporosis was missed by the model.

Metrics Derived:

- **Accuracy:** 87% (overall reliability of the model).
- **Precision:** 94% (model is excellent at minimizing false alarms).
- **Recall (Sensitivity):** 73% (model misses some actual osteoporosis cases).
- **F1-Score:** 82% (a balance between precision and recall).

However, its lower recall (73%) indicates it could miss some critical osteoporosis cases, which may require adjustments to thresholds or rebalancing of data.

Deep Learning Results

- **Accuracy (Test Data):** 88% (slightly higher than traditional models).
- **Precision:** 91% (high confidence in positive predictions).
- **Recall:** 78% (better at capturing true positive cases than the Random Forest model).
- **Loss Reduction:** Significant loss reduction in training shows effective learning.

Deep learning models captured complex interactions in the data and improved recall,

reducing false negatives compared to traditional models. However, deep learning can overfit if not carefully regularized, requiring additional techniques like dropout or early stopping.

Naive Bayes Results

- **Accuracy:** 81% (lower than ensemble and deep learning models).
- **Precision:** 89% (decent at predicting positive cases).
- **Recall:** 70% (misses more positive cases compared to other models).

Naive Bayes assumes feature independence, which may not hold true in real-world medical data with correlated features. While it provides a quick and interpretable baseline, it underperforms compared to more sophisticated methods.

T-Test

Features like "Age," "Calcium Intake," and "Vitamin D Intake" showed statistically significant differences ($p < 0.05$) between osteoporosis and non-osteoporosis groups.

Insignificant features (e.g., "Race/Ethnicity") suggest limited predictive power.
Significant features should be prioritized for model training and interpretation. This result validates that the selected features contribute meaningfully to the prediction, enhancing the model's explainability.

Machine learning models

- Combined models (Random Forest, Gradient Boosting) improved overall performance:
    - **Accuracy:** 89% (best result among tested methods).

- **Precision:** 92% (consistent high confidence in predictions).
- **Recall:** 80% (best at minimizing false negatives).

  The ensemble approach leverages the strengths of multiple models, offering a robust solution. It provides better generalizability and is ideal for deployment scenarios where both high accuracy and sensitivity are critical.

The analysis demonstrates the effectiveness of machine learning models, especially ensemble and deep learning approaches, in predicting osteoporosis risk. While the models are generally reliable and precise, further optimizations are necessary to enhance recall and minimize missed cases. These findings align with the goal of developing a robust predictive tool for early screening and intervention in osteoporosis management.

The Random Forest Classifier performed the best with an accuracy score of 0.85, balanced precision, recall, and F1 scores, and a high ROC-AUC score of 0.88.

Conclusion

I have implemented multiple machine learning models, including Random Forest, Gradient Boosting, Deep Learning, and Naive Bayes, to predict osteoporosis risk based on lifestyle and demographic factors. The ensemble models and deep learning techniques demonstrated the highest accuracy (89% and 88%, respectively), with deep learning offering improved recall (78%), which is crucial for minimizing missed cases of osteoporosis. The findings highlighted key features such as age, calcium intake, and vitamin D intake as significant predictors, validated through statistical tests like the T-test. The overall results suggest that these models can serve as effective tools for early screening and risk assessment, supporting clinicians in proactive osteoporosis management.

Limitations

Despite the promising results, the study faced several limitations. The recall for most models remained suboptimal, indicating that some cases of osteoporosis were still missed, which could lead to underdiagnosis. Additionally, the dataset was limited to approximately 1,958 entries, which might not capture the full variability of the population, particularly when generalizing to different ethnic or demographic groups. Feature interactions, such as the combined effects of calcium and vitamin D intake, could not be fully explored due to potential limitations in the dataset structure. Furthermore, deep learning models, while effective, required significant computational resources and risked overfitting due to the small dataset size.

Alternate Approaches

To address these limitations, future work could explore advanced techniques like oversampling methods to balance the dataset and improve recall. Additionally, incorporating more advanced ensemble methods, such as stacking multiple models or blending, could further enhance prediction accuracy and robustness. Time-to-event analysis, such as survival analysis, could also provide insights into when individuals are likely to develop osteoporosis, adding a temporal dimension to the predictions. Furthermore, integrating external data sources or larger datasets from diverse populations would improve the model's generalizability. Finally, applying Bayesian approaches or interpretable deep learning architectures (e.g., TabNet) could enhance the model’s transparency and reliability in real-world applications. These strategies offer promising pathways to refine the current predictive framework and ensure its clinical utility.

REFERENCES-

- Abu-Dieh, H. (2015, May 30). *Osteoporosis* [Slide show]. SlideShare. https://www.slideshare.net/slideshow/osteoporosis-48772976/48772976
- He, Y., Lin, J., Zhu, S., Zhu, J., & Xu, Z. (2024). Deep learning in the radiologic diagnosis of osteoporosis: a literature review. *Journal of International Medical Research*, *52*(4). https://doi.org/10.1177/03000605241244754
- Ibrahim, N., Nabil, N., & Ghaleb, S. (2019). Pathophysiology of the risk factors associated with osteoporosis and their correlation to the T-score value in patients with osteopenia and osteoporosis in the United Arab Emirates. *Journal of Pharmacy and Bioallied Sciences*, *11*(4), 364. https://doi.org/10.4103/jpbs.jpbs_4_19
- Irmawati, I., Juningsih, E. H., & Yanto, Y. (2024). Predictive Modeling of Osteoporosis Risk Factors using XGBoost and Bagging Ensemble Technique. *Journal Medical Informatics Technology*, 6–10. https://doi.org/10.37034/medinftech.v2i1.27
- Kim, N. S. K., Yoo, N. T. K., Oh, N. E., & Kim, N. D. W. (2013). Osteoporosis risk prediction using machine learning and conventional methods. *PubMed*, 188–191. https://doi.org/10.1109/embc.2013.6609469
- *Osteoporosis risk factors | UC San Diego Health*. (n.d.). UC San Diego Health. https://health.ucsd.edu/care/endocrinology-diabetes/osteoporosis/risk-factors/
- *Osteoporosis risk prediction*. (2024, March 20). Kaggle. https://www.kaggle.com/datasets/amitvkulkarni/lifestyle-factors-influencing-osteoporosis/data

- Pouresmaeili, F., Dehghan, B. K., Kamarehei, M., & Meng, G. Y. (2018). A comprehensive overview on osteoporosis and its risk factors. *Therapeutics and Clinical Risk Management*, *Volume 14*, 2029–2049. https://doi.org/10.2147/tcrm.s138000
- Qiu, C., Su, K., Luo, Z., Tian, Q., Zhao, L., Wu, L., Deng, H., & Shen, H. (2024a). Developing and comparing deep learning and machine learning algorithms for osteoporosis risk prediction. *Frontiers in Artificial Intelligence*, *7*. https://doi.org/10.3389/frai.2024.1355287
- Qiu, C., Su, K., Luo, Z., Tian, Q., Zhao, L., Wu, L., Deng, H., & Shen, H. (2024b). Developing and comparing deep learning and machine learning algorithms for osteoporosis risk prediction. *Frontiers in Artificial Intelligence*, *7*. https://doi.org/10.3389/frai.2024.1355287
- Tu, J., Liao, W., Liu, W., & Gao, X. (2024). Using machine learning techniques to predict the risk of osteoporosis based on nationwide chronic disease data. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-56114-1