KG REDDY
College of Engineering
& Technology
AN AUTONOMOUS INSTITUTION

code unnati

edunet
foundation

SAP

OSTEOPOROSIS RISK PREDICTION

PROBLEM STATEMENT

Osteoporosis is a condition characterized by weakened bones and an increased risk of fractures. It often remains undiagnosed until significant damage occurs, leading to severe health consequences. Globally, over 200 million people are affected, with a particularly high incidence among post-menopausal women and the elderly. Early detection is crucial for effective management and prevention of fractures. However, access to diagnostic tools like DEXA scans is limited due to high costs and inadequate infrastructure, especially in rural and resource-poor settings. This situation underscores the urgent need for a low-cost, accessible, and reliable predictive system that can identify individuals at risk of osteoporosis and enable timely intervention.

PROJECT OVERVIEW

This project involves the development of a machine learning-driven diagnostic tool for osteoporosis risk assessment, utilizing basic demographic and medical data. The primary goal is to democratize early risk identification using artificial intelligence, thereby reducing reliance on costly diagnostic tests in the initial stages. By analyzing key indicators such as age, gender, BMI, lifestyle choices, and medical history, the model predicts whether an individual is at risk of developing osteoporosis.

Workflow Summary:

1. Data Collection & Preprocessing: Gathering relevant data and cleaning it for analysis.
2. Feature Engineering and Encoding: Transforming raw data into a format suitable for machine learning.
3. Splitting Dataset into Training/Testing Sets: Dividing the data to evaluate model performance.
4. Training Multiple Classifiers: Implementing various machine learning algorithms to find the best model.
5. Evaluating Performance: Assessing models based on accuracy, precision, recall, F1-score, and ROC-AUC.
6. Visualizing Confusion Matrix and ROC Curve: Providing insights into model performance through visual tools.

SOLUTION OFFERED

The final solution is a smart, machine learning-based osteoporosis prediction system designed to provide:

1. Fast and Early Risk Assessment: Utilizing basic personal and health details for quick evaluations.
2. Performance Visualization: Offering detailed metrics and plots, including ROC curves and confusion matrices.
3. Multi-Model Comparison: Allowing users to identify the best algorithm for the given dataset.
4. Scalable Deployment Capability: Ready for integration into digital health platforms or hospital databases.

Features of the System:

1. Input Fields: Age, Gender, Weight, Height, BMI, Alcohol intake, Smoking habits, Physical activity, Family history, etc.
2. Output: Prediction label indicating Risk or No Risk, along with a probability score.
3. Visual Insights: ROC curve and confusion matrix for performance evaluation.
4. Evaluation Metrics: Accuracy, Precision, Recall, F1-score, and AUC-ROC for comprehensive assessment.

WHO ARE THE END USERS?

This project caters to a wide audience in both clinical and public health domains:

1. Primary Healthcare Providers: Assisting in early diagnosis in non-specialist clinics.
2. Hospitals and Diagnostic Centers: Facilitating digital screening before confirming diagnoses with imaging.
3. Rural Healthcare Programs: Extending preventive care in under-equipped regions.
4. Patients & Elderly Individuals: Providing a personal screening tool to assess risk and consult healthcare professionals accordingly.
5. Insurance and Health Tech Startups: Integrating risk profiling into digital health assessments or policy generation.

TECHNOLOGY USED TO SOLVE THE PROBLEM

1. MACHINE LEARNING ALGORITHMS:

The project experimented with 11 classification algorithms to identify the most accurate and reliable model.

1. Decision Tree: A simple, interpretable tree-based classifier.
2. Random Forest: An ensemble of decision trees using bagging for improved accuracy.
3. Gradient Boosting: A boosting method that builds weak learners sequentially to enhance performance.
4. XGBoost: An optimized version of gradient boosting with regularization to prevent overfitting.
5. K-Nearest Neighbors (KNN): An instance-based learning algorithm that classifies based on proximity to training examples.
6. AdaBoost: An ensemble boosting method that adjusts weights of misclassified instances to improve accuracy.
7. Naive Bayes: A probabilistic model based on Bayes' theorem, suitable for classification tasks.
8. Support Vector Machine (SVM): A margin-based classifier that uses kernel functions to handle non-linear data.
9. Extra Trees: A randomized tree ensemble method that reduces variance and improves robustness.
10. Stacking Classifier: Combines multiple base learners through a meta-model for enhanced predictions.
11. Bagging Classifier: Aggregates predictions from bootstrapped datasets to improve stability and accuracy.

Model Evaluation Metrics

1. Accuracy: Measures the overall correctness of the model.
2. Precision: Indicates the true positive rate among all positive predictions.
3. Recall (Sensitivity): Assesses the model's ability to detect all positive cases.
4. F1-Score: Represents the harmonic mean of precision and recall, balancing both metrics.
5. ROC-AUC: Evaluates the model's ability to differentiate between classes, with higher values indicating better performance.
6. Confusion Matrix: Displays the distribution of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for a comprehensive view of model performance.

2. PYTHON PACKAGES & LIBRARIES

1. pandas, numpy: Essential for data wrangling and handling.
2. scikit-learn: Provides tools for modeling, evaluation, and preprocessing.
3. xgboost: Implements the gradient boosting framework for enhanced model performance.
4. matplotlib, seaborn: Used for data visualization to present insights effectively.
5. joblib: Facilitates model saving and loading for deployment purposes.

3. VISUALIZATION OUTPUTS

1. Confusion Matrix: A heatmap that provides insights into classification performance.
2. ROC Curve: Illustrates the trade-offs between true positive rates and false positive rates at various thresholds.
3. Feature Importance: Highlights the significance of different features in tree-based models, aiding interpretability.

4. DATASET OVERVIEW

a. Features: 15 columns including Age, Sex, BMI, and various Lifestyle factors
b. Target: A binary classification indicating the presence or absence of osteoporosis.
c. Format: The dataset is cleaned and label-encoded for effective analysis.
d. Source: Data is derived from health surveys and patient records, structured for machine learning applications.

RESULTS FOR OSTEOPOROSIS RISK PREDICTION

Model Performance Comparison:

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Decision Tree	0.7993	0.7729	0.8647	0.8162	0.7968
Random Forest	0.8435	0.9203	0.7624	0.8339	0.8928
Gradient Boosting	0.9167	1.0000	0.8383	0.9120	0.9239
XGBoost	0.8827	0.9209	0.8449	0.8812	0.9261
K-Nearest Neighbors	0.8537	0.9189	0.7855	0.8470	0.8980
AdaBoost	0.8690	1.0000	0.7459	0.8544	0.9145
Naive Bayes	0.8520	0.9154	0.7855	0.8455	0.8993
Support Vector Machine	0.8452	0.9454	0.7426	0.8318	0.9000
Extra Trees	0.7976	0.8333	0.7591	0.7945	0.8552
Stacking Classifier	0.8452	0.8655	0.8284	0.8465	0.9036
Bagging Classifier	0.9099	0.9735	0.8482	0.9065	0.9238

Best Model: XGBoost

1. Accuracy: 0.8827
2. Precision: 0.9209
3. Recall: 0.8449
4. F1 Score: 0.8812
5. ROC AUC: 0.9261

CONCLUSION

The XGBoost model emerged as the best performer in the osteoporosis risk prediction task, achieving the highest ROC AUC score of 0.9261. This indicates that XGBoost is highly effective in distinguishing between individuals at risk and not at risk for osteoporosis.