**Osteoporosis Risk Prediction**

**Project Report**

**Submitted To: Sir Rasikh  
Submitted By: Waqar Abbas  
Roll No: SU92-BSDSM-F23-010**

**Section: BSDSM-3A**

# **Abstract**

This project focuses on predicting the risk of osteoporosis using machine learning techniques. By analyzing various demographic, lifestyle, and medical factors, we aim to develop a predictive model that can assist healthcare professionals in identifying individuals at high risk of developing osteoporosis.

# **Introduction**

Osteoporosis is a chronic disease characterized by low bone mass and deterioration of bone tissue, leading to increased fracture risk. Early identification of individuals at risk can lead to timely interventions and preventive measures. This project utilizes a dataset containing various features related to osteoporosis, including age, gender, lifestyle choices, and medical history.

# **Dataset Description**

The dataset comprises 1958 records and 16 features, including:

* **Age**: Age of the individual.
* **Gender**: Gender (Male/Female).
* **Hormonal Changes**: Hormonal status (Normal/Postmenopausal).
* **Family History**: Family history of osteoporosis (Yes/No).
* **Race/Ethnicity**: Racial or ethnic background.
* **Body Weight**: Body weight category (Normal/Underweight).
* **Calcium Intake**: Level of calcium intake (Low/Adequate).
* **Vitamin D Intake**: Level of vitamin D intake (Sufficient/Insufficient).
* **Physical Activity**: Level of physical activity (Active/Sedentary).
* **Smoking**: Smoking status (Yes/No).
* **Alcohol Consumption**: Alcohol consumption level (Moderate/None).
* **Medical Conditions**: Relevant medical conditions.
* **Medications**: Medications taken.
* **Prior Fractures**: History of prior fractures.
* **Osteoporosis**: Target variable indicating the presence of osteoporosis (1 for Yes, 0 for No).

# **Technologies Used**

The project utilized various technologies and libraries for data manipulation, visualization, machine learning, and deep learning. The following libraries were employed:

## **Data Manipulation and Visualization:**

* + **numpy**: For numerical operations and array manipulations.
  + **pandas**: For data manipulation and analysis.
  + **matplotlib**: For creating static, interactive, and animated visualizations.
  + **seaborn**: For statistical data visualization.

## **Data Preprocessing:**

* + **sklearn.preprocessing**: For preprocessing tasks, including label encoding and standardization.
  + **sklearn.model\_selection**: For splitting the dataset and performing cross-validation.
  + **imblearn.over\_sampling**: For handling class imbalance using SMOTE.

## **Machine Learning Models:**

* + **sklearn.ensemble**: For ensemble methods like Random Forest, AdaBoost, and Gradient Boosting.
  + **sklearn.linear\_model**: For Logistic Regression.
  + **sklearn.svm**: For Support Vector Classifier (SVC).
  + **lightgbm**: For the LightGBM classifier.
  + **xgboost**: For the XGBoost classifier.
  + **catboost**: For the CatBoost classifier.

## **Deep Learning:**

* + **tensorflow.keras**: For building and training neural network models.
  + **tensorflow.keras.layers**: For defining layers in the neural network.
  + **tensorflow.keras.optimizers**: For optimization algorithms.
  + **tensorflow.keras.callbacks**: For implementing callbacks like early stopping and learning rate scheduling.
  + **tensorflow.keras.regularizers**: For applying regularization techniques.

## **Miscellaneous:**

* + **joblib**: For saving and loading models and preprocessing objects.
  + **warnings**: For managing warnings during execution.

# **Data Preprocessing**

1. **Loading the Data**: The dataset was loaded into a pandas DataFrame for analysis.
2. **Handling Missing Values**: Identified columns with missing values and filled them with 'None'.
3. **Dropping Irrelevant Columns**: The 'Id' column was dropped as it does not contribute to the prediction.
4. **Encoding Categorical Variables**: Categorical features were encoded using Label Encoding to convert them into numerical format.

# **Exploratory Data Analysis (EDA)**

* **Distribution of Osteoporosis**: A pie chart was created to visualize the proportion of individuals with and without osteoporosis.
* **Age Distribution**: Histograms were plotted to show the distribution of ages for individuals with and without osteoporosis.
* **Count Plots**: Various count plots illustrated the relationship between osteoporosis and categorical features such as Gender, Hormonal Changes, Family History, Race/Ethnicity, Body Weight, Calcium Intake, Vitamin D Intake, Physical Activity, Smoking, Alcohol Consumption, Medical Conditions, and Medications.

# **Model Development**

Multiple machine learning models were evaluated, including:

* Logistic Regression
* Random Forest Classifier
* Support Vector Classifier (SVC)
* AdaBoost Classifier
* Gradient Boosting Classifier
* XGBoost Classifier
* LightGBM Classifier
* CatBoost Classifier

# **Model Training and Evaluation**

1. **Data Splitting**: The dataset was split into training, validation, and test sets.
2. **Standardization**: Features were standardized using **StandardScaler**.
3. **Model Training**: Each model was trained on the training set, and hyperparameter tuning was performed using Grid Search for models like Random Forest and XGBoost.
4. **Model Evaluation**: Models were evaluated using accuracy, precision, recall, and F1 score.

# **Results**

The results of the models are summarized as follows:

* The best-performing model was identified based on cross-validated accuracy.
* The LightGBM classifier achieved an accuracy of approximately 89%, with precision, recall, and F1 scores also indicating strong performance.
* The ROC curve was plotted for each model to visualize performance, with AUC scores calculated.

# **Deep Learning Approach**

An advanced neural network architecture was developed using TensorFlow/Keras:

* The model consisted of multiple dense layers with batch normalization, dropout, and LeakyReLU activation functions.
* A custom learning rate scheduler and early stopping were implemented to enhance training efficiency.
* The model was trained and evaluated, achieving high accuracy and other performance metrics.

# **Conclusion**

The osteoporosis risk prediction model developed in this project demonstrates the potential of machine learning and deep learning techniques in healthcare. By identifying individuals at high risk for osteoporosis, healthcare providers can implement preventive strategies to reduce the incidence of fractures. Future work may involve integrating additional features and exploring more advanced deep learning architectures to further enhance predictive accuracy.