Project Title: Breast Cancer Prediction

Description:
This project aims to develop a machine learning model to predict the likelihood of breast cancer diagnosis within a 90-day period using demographic, clinical, and environmental factors. The dataset includes features such as patient demographics, medical history, environmental factors like air quality, and socioeconomic indicators. The goal is to build a predictive model that can assist healthcare professionals in identifying patients at higher risk of breast cancer, enabling early intervention and personalized treatment strategies.

Key Features:
- Exploratory data analysis (EDA) to understand the dataset and identify patterns.
- Data preprocessing techniques including handling missing values, encoding categorical variables, and feature scaling.
- Implementation of machine learning algorithms such as XGBoost, LightGBM, and Random Forest for classification.
- Evaluation metrics including AUC score, F1 score, and log loss to assess model performance.
- Cross-validation techniques to ensure robustness and generalization of the models.

Results:
- Achieved an average AUC score of 0.80 across 10 folds using XGBoost.
- Explored feature importance to identify key factors influencing breast cancer diagnosis.
- Provided insights into the predictive capabilities of the developed models and their potential clinical applications.

Technologies Used:
- Python
- scikit-learn
- XGBoost
- LightGBM
- pandas
- matplotlib
- seaborn

Future Work:
- Further refinement of the models through hyperparameter tuning and feature engineering.
- Integration of additional data sources to enhance predictive performance.
- Deployment of the model as a web application or API for real-time predictions.