

ZION TECH HUB: CAPSTONE PROJECT

Project 1: Breast Cancer Detection Model

Objective

Develop a machine learning model to predict breast cancer malignancy based on diagnostic features, with potential to assist in early detection and treatment planning.

Potential Benefits

- Enable earlier detection of malignant breast tumors, improving patient outcomes
- Reduce unnecessary biopsies for benign cases
- Assist radiologists in diagnostic decision-making
- Improve hospital resource allocation by identifying high-risk cases
- Provide insights into key factors contributing to malignancy

Machine Learning Workflow

1. Data Preprocessing:

- Handle missing values.
- Encode categorical variables (e.g., one-hot encoding).
- Split the dataset into training and testing sets.
- Perform feature engineering.

2. Model Selection:

- Evaluate at least five classification models (e.g., Random Forest, SVM, Decision Tree, etc.).

3. Model Training:

- Train the selected models using the training set.
- Tune hyperparameters for improved performance.

4. Model Evaluation:

- Assess using classification metrics (accuracy, recall, F1-score, confusion matrix).
- Use cross-validation to ensure generalizability.

5. Model Deployment:

- Deploy the best-performing model into a production environment for clinical use.
- Continuously monitor and update the model as needed.

Dataset

Click the link below to access the dataset and its description.
Kaggle Breast Cancer Data

Project 2: Market Sales Prediction Model

Objective

Develop a machine learning regression model to predict market sales using historical data. The goal is to generate actionable insights for business planning, inventory management, and marketing strategy.

Business Problem

- Forecasting market sales can help businesses better allocate resources, adjust inventory levels, and design effective marketing campaigns.
- Understanding the key factors driving sales will also empower decision-makers to plan for future growth.

Machine Learning Workflow

1. Data Preprocessing:

- Handle any missing values within the dataset.
- Encode categorical variables (if present) using techniques such as one-hot encoding.
- Split the data into training and testing sets.
- Conduct feature engineering to enhance model performance.

2. **Model Selection:** • Since this is a regression task, consider at least five regression models. For example: Linear Regression Decision Tree Regressor, Random Forest Regressor, Support Vector Regression (SVR), Gradient Boosting Regressor

3. Model Training:

- Train the different regression models using the training set.
- Apply hyperparameter tuning to optimize each model's performance.

4. **Model Evaluation:** • Evaluate the regression models using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared.

- Utilize cross-validation techniques to assess the models' generalization ability.

5. **Model Deployment:** • Integrate the best-performing regression model into a production system for ongoing sales forecasting.

- Continuously monitor model performance and update as new data becomes available.

Dataset

Click the link below to access the dataset and its description.
Kaggle Market Sales Data