Documentation Structure for Machine Learning Projects

1. Project Title and Overview

- **Title**: Car Price Prediction (Regression) and Breast Cancer Image Classifier (Classification)
- **Objective**:
  - **Car Price Prediction**: Predict the price of cars based on given features using Linear Regression and KNN.
  - **Breast Cancer Image Classifier**: Classify breast cancer images as malignant or benign using Logistic Regression and KNN.
- **Technologies Used**: Python, Scikit-Learn, NumPy, Pandas, and Matplotlib.
- **Models Used**:
  - **Car Price Prediction**: Linear Regression and KNN
  - **Breast Cancer Image Classifier**: Logistic Regression and KNN

2. Dataset Description

2.1. Car Price Prediction Dataset

- **Source**: https://drive.google.com/file/d/1UIj8rOmDJn4UgqDTJezMNfzr7jh_behL/view
- **Dataset Size**: 8129 x 13.
- **Data Format**: CSV.
- **Target Variable**: The target variable is **Price** (the price of the car).
- **Features**: Name, year, km_driven, fuel, seller_type, transmission, owner, mileage,engine, max_power, torque, and seats.
- **Data Preprocessing**: Handling missing values, encoding categorical variables, feature scaling.

2.2. Breast Cancer Image Classifier Dataset

- **Source**: https://www.kaggle.com/datasets/hayder17/breast-cancer-detection.
- **Dataset Size**: 2708 images each image is 224 x 224.
- **Data Format**: JPG.
- **Labels**: Binary classification (Malignant = 1, Benign = 0).
- **Data Preprocessing**: Image resizing.

3. Algorithms Used

3.1. Car Price Prediction

- Linear Regression
- K-Nearest Neighbors (KNN) Regression

3.2. Breast Cancer Image Classifier

- Logistic Regression
- K-Nearest Neighbors (KNN) Classifier

4. Model Training and Testing

- **Data Split**: Train-test split (e.g., 80% training, 20% testing).
- **Model Training**: using *SCIKIT-Learn*

5. Evaluation Metrics

5.1. Car Price Prediction Metrics

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- R-Squared (R^2)

5.2. Breast Cancer Image Classifier Metrics

- Accuracy Score
- Precision
- Recall
- AUC-ROC Curve
- Confusion Matrix

6. Results and Insights

The **Linear Regression** model for car price prediction performed efficiently with quick training time and provided interpretable results but struggled with non-linear relationships, while **KNN Regressor** performed slightly better due to its ability to capture non-linear patterns, though it was slower during predictions and sensitive to feature scaling.

For the breast cancer classification task, **Logistic Regression** delivered good accuracy and precision on linearly separable data with fast training, but **KNN Classifier** achieved slightly better performance by capturing complex decision boundaries at the cost of slower inference. Challenges included noisy data, missing values, and overfitting, which were addressed by imputing missing data (mean/mode), applying Min-Max scaling to handle KNN's sensitivity to feature ranges. Overall, KNN outperformed the linear models for both regression and classification tasks, but at the expense of computational efficiency.

7. Conclusion

The goal of the project was to predict car prices using regression models and classify breast cancer images into malignant or benign categories. For car price prediction, **KNN Regressor** performed slightly better than Linear Regression due to its ability to capture non-linear patterns, while Linear Regression was faster and more interpretable. For breast cancer classification, **KNN Classifier** outperformed Logistic Regression by better capturing complex decision boundaries, although Logistic Regression was quicker to train. Both **Logistic Regression** and **KNN** were used for classification in the breast cancer project. Potential improvements include experimenting with different algorithms such as **Random Forests** or **Gradient Boosting** for both regression and classification tasks to improve performance and handle non-linear relationships more effectively.

8. Code: https://github.com/EgyptianFather/ML-Project.