Report: Machine Learning Coding Exercise

In this report, we will discuss the implementation and analysis of machine learning models on three healthcare-related datasets: Dataset1, Dataset2, and Dataset3.

Dataset1: Predicting Diseases from Symptoms

- **Classification Task:**
    - **Step 1:** Loading Data, Data Pre-processing, EDA
        - Loaded the dataset from **Training.csv** and **Testing.csv**.
        - Pre-processed the data by handling missing values and encoding categorical variables.
        - Conducted exploratory data analysis (EDA) to understand the distribution of data.
    - **Step 2:** Feature Engineering, Creating Train, and Test Datasets
        - Engineered features and created the feature matrix (X) and target variable (y).
        - Split the data into training and testing sets.
    - **Step 3:** Apply at least 2 algorithms for classification
        - Utilized Random Forest Classifier and Logistic Regression algorithms for classification.
        - Trained and tested both algorithms on the dataset.
    - **Step 4:** Generate at least 2 Evaluation Metrics on each algorithm
        - Evaluated the performance of both algorithms using metrics like accuracy.
        - Computed precision, recall, and F1-score to gain deeper insights into model performance.
    - **Step 5:** Comparing the results
        - Compared the performance of Random Forest Classifier and Logistic Regression using evaluation metrics.
        - Analyzed the strengths and weaknesses of each algorithm.
    - **Step 6:** Fine Tune the best algorithm
        - Fine-tuned the Logistic Regression algorithm using GridSearchCV to find the best hyperparameters.
        - Evaluated the fine-tuned model's performance on the test set.

Dataset2: Predicting Heart Stroke

- **Classification Task:**
    - **Step 1:** Loading Data, Data Pre-processing, EDA
        - Loaded the dataset from **healthcare-dataset-stroke-data.csv**.
        - Handled missing values and encoded categorical variables.
        - Conducted exploratory data analysis (EDA) to understand the distribution of data.
    - **Step 2:** Feature Engineering, Creating Train, and Test Datasets
        - Engineered features and created the feature matrix (X) and target variable (y).
        - Split the data into training and testing sets.
    - **Step 3:** Apply at least 2 algorithms for classification
        - Utilized Random Forest Classifier and Logistic Regression algorithms for classification.
        - Trained and tested both algorithms on the dataset.
    - **Step 4:** Generate at least 2 Evaluation Metrics on each algorithm
        - Evaluated the performance of both algorithms using metrics like accuracy.
        - Computed precision, recall, and F1-score to gain deeper insights into model performance.
    - **Step 5:** Comparing the results
        - Compared the performance of Random Forest Classifier and Logistic Regression using evaluation metrics.
        - Analyzed the strengths and weaknesses of each algorithm.
    - **Step 6:** Fine Tune the best algorithm
        - Fine-tuned the Logistic Regression algorithm using GridSearchCV to find the best hyperparameters.
        - Evaluated the fine-tuned model's performance on the test set.
- **Regression Task:**
    - **Step 1:** Loading Data, Data Pre-processing, EDA
        - Loaded the dataset from **healthcare-dataset-stroke-data.csv**.

        - Handled missing values and encoded categorical variables.
        - Conducted exploratory data analysis (EDA) to understand the distribution of data.
    - **Step 2:** Feature Engineering, Creating Train, and Test Datasets
        - Engineered features and created the feature matrix (X) and target variable (y).
        - Split the data into training and testing sets.
    - **Step 3:** Apply at least 2 algorithms for regression
        - Utilized Linear Regression and Ridge Regression algorithms for regression tasks.
        - Trained and tested both algorithms on the dataset.
    - **Step 4:** Generate at least 2 Evaluation Metrics on each algorithm
        - Evaluated the performance of both regression algorithms using metrics like mean squared error and R2 score.
        - Assessed the goodness of fit and predictive capabilities of the models.
    - **Step 5:** Comparing the results
        - Compared the performance of Linear Regression and Ridge Regression using evaluation metrics.
        - Analyzed the effectiveness of each algorithm in predicting medical insurance costs.
    - **Step 6:** Fine Tune the best algorithm
        - Fine-tuned the Ridge Regression algorithm using RandomizedSearchCV to optimize model performance.
        - Assessed the impact of hyperparameter tuning on the regression model's performance.

Dataset3: Predicting Medical Insurance Costs

- **Regression Task:**
    - **Step 1:** Loading Data, Data Pre-processing, EDA
        - Loaded the dataset from **insurance.csv**.
        - Handled missing values and encoded categorical variables.

    - Conducted exploratory data analysis (EDA) to understand the distribution of data.
- **Step 2:** Feature Engineering, Creating Train, and Test Datasets
    - Engineered features and created the feature matrix (X) and target variable (y).
    - Split the data into training and testing sets.
- **Step 3:** Apply at least 2 algorithms for regression
    - Utilized Linear Regression and Ridge Regression algorithms for regression tasks.
    - Trained and tested both algorithms on the dataset.
- **Step 4:** Generate at least 2 Evaluation Metrics on each algorithm
    - Evaluated the performance of both regression algorithms using metrics like mean squared error and R2 score.
    - Assessed the goodness of fit and predictive capabilities of the models.
- **Step 5:** Comparing the results
    - Compared the performance of Linear Regression and Ridge Regression using evaluation metrics.
    - Analyzed the effectiveness of each algorithm in predicting medical insurance costs.
- **Step 6:** Fine Tune the best algorithm
    - Fine-tuned the Ridge Regression algorithm using RandomizedSearchCV to optimize model performance.
    - Assessed the impact of hyperparameter tuning on the regression model's performance.

Through these tasks, we aimed to explore various machine learning techniques for healthcare-related predictions, ranging from disease diagnosis to medical cost estimation. The models developed in this exercise provide valuable insights for healthcare professionals and policymakers in making informed decisions.