LOG BOOK

Athulya Shanty (C00313623)

I started the notebook from scratch.

Filename	Dataset	Steps
1_Data_Cleaning.ipynb	Alzheimers Data	Load the CSV file into a DataFrame Check the number of rows and columns in the DataFrame Display basic information about the dataset Display summary statistics of the dataset Check for missing values Check if the data is imbalanced Check for duplicate rows in the dataset Save the cleaned data to a CSV file
	Student Performance Data	Load the CSV file into a DataFrame Check the number of rows and columns in the DataFrame Display basic information about the dataset Display summary statistics of the dataset Check for missing values Check for duplicate rows in the dataset Save the cleaned data to a CSV file
2_Exploratory_Data_Analysis.ipynb	Alzheimers Data	Load the CSV file into a DataFrame Plot the distribution of the Diagnosis column Plot the top 10 countries with the highest Alzheimer's cases Plot the gender distribution for diagnosed cases Plot the distribution of age groups by diagnosis Plot the distribution of the Family History of Alzheimer's Categorize BMI and plot the distribution by diagnosis Identify numerical & Categorical columns Outlier Detection
	Student Performance Data	Load the CSV file into a DataFrame Plotted Relation between hours studied and performance index Plot the relation between sleep hours and performance index correlation heatmap Outlier Detection
3_Data_Preprocessing.ipynb	Alzheimers Data	Load the CSV file into a DataFrame Identify numerical & Categorical columns Normalize the numerical columns using StandardScaler() Label Encoding Save the processed data to a CSV file Split the data into training (70%) and remaining (30%) Split the remaining data into test (50% of remaining) and validation (50% of remaining) Define the features and target variable for training, testing & validation Save the features and target for training, testing & validation as pickle file
	Student Performance Data	Load the CSV file into a DataFrame Identify numerical & Categorical columns Normalize the numerical columns using StandardScaler() Label Encoding Split the data into training (70%) and remaining (30%) Split the remaining data into test (50% of remaining) and validation (50% of remaining) Define the features and target variable for training, testing & validation Save the features and target for training, testing & validation as pickle file
4_Linear_Regression.ipynb	Student Performance Data	Load the features and target for training, testing & validation Create a linear regression model, Train the model using the training data, Predict & Evaluate RMSE and R^2 score using the test and validation data Plot actual vs predicted for test data Plot actual vs predicted for validation data
5_Logistic_Regression.ipynb	Alzheimers Data	Load the features and target for training, testing & validation Created a function *calculate_rmse_and_predict( )* to calculate the rmse value and predict the target values Created a function *train_and_evaluate_logistic_regression()* for logistic regression model Calculate the Test Accuracy Calculate RMSE for test & validation data Save the accuracy, RMSE on test & validation data as a pickle file
6_Decision_Trees.ipynb	Alzheimers Data	Load the features and target for training, testing & validation Created a function *calculate_rmse_and_predict( )* to calculate the rmse value and predict the target values Defined a range (1,15) to find optimal depth of decision tree by Plotting the mean cross-validated accuracy for each depth. Optimal depth was calculated to be 5 Train the Decision Tree Classifier with the optimal depth, Predict on the test data, Calculate the accuracy Calculate RMSE for test & validation data Plot the decision tree with a depth of 2 levels (opted 2 for better visualization) Update the accuracy, RMSE on test & validation data to the pickle file
7_Random_Forest.ipynb	Alzheimers Data	Load the features and target for training, testing & validation Created a function *calculate_rmse_and_predict( )* to calculate the rmse value and predict the target values Initialize the Random Forest Classifier, Train the classifier, Predict on the test data, Calculate the accuracy Calculate RMSE for test & validation data Update the accuracy, RMSE on test & validation data to the pickle file
8_KNN.ipynb	Alzheimers Data	Load the features and target for training, testing & validation Created a function *calculate_rmse_and_predict( )* to calculate the rmse value and predict the target values Performed K Fold Cross Validation to find the optimal k value and optimal k= 17 Created a function *train_knn_model()* for KNN model Implemented KNN with Euclidean distance & Calculated RMSE for test & validation data Implemented KNN with Manhattan distance & Calculated RMSE for test & validation data Implemented KNN with Minkowski distance & Calculated RMSE for test & validation data Update the accuracy, RMSE on test & validation data to the pickle file
9_SVM.ipynb	Alzheimers Data	Load the features and target for training, testing & validation Created a function *calculate_rmse_and_predict( )* to calculate the rmse value and predict the target values Created a function *train_and_evaluate_svm()* for SVM model Implemented SVM with linear kernel & Calculated RMSE for test & validation data Implemented SVM with RBF kernel & Calculated RMSE for test & validation data Update the accuracy, RMSE on test & validation data to the pickle file
10_Naive_Bayesian.ipynb	Alzheimers Data	Load the features and target for training, testing & validation Created a function *calculate_rmse_and_predict( )* to calculate the rmse value and predict the target values Applied the Gaussian Naive Bayes model & Calculated Accuracy Calculate RMSE for test & validation data Applied the Bernoulli Naive Bayes model & Calculated Accuracy Update the accuracy, RMSE on test & validation data to the pickle file
11_K_Means_Clustering.ipynb	Alzheimers Data	Import the dataset and remove Diagnosis column for clustering Applied Elbow method to find the optimal number of clusters and k=2 Applied K Means Clustering with k=2 Predict the clusters