LOGBOOK

Athulya Shanty (C00313623)

Steps	Reason/Explanatiom	Time Taken (mins)	Difficulty Level	File Name	Dataset
Checked the number of rows, displayed information and summary statistics of both the datasets.	Inorder to understand the datasets better	5	Easy	1_Data_Cleaning.ipynb	Both Alzheimers Data & Student Performance Data
Checked for missing values	Inorder to prevent biased analysis and for improving the Quality of the model.	5	Easy	1_Data_Cleaning.ipynb	Both Alzheimers Data & Student Performance Data
Checked whether the data is balanced or Imbalanced	Inorder to prevent biased predictions, I checked whether the data is balanced by counting occurances of each target variable and then calculating the imbalanced ratio. I choose the threshold as 1.5	10	Easy	1_Data_Cleaning.ipynb	Both Alzheimers Data & Student Performance Data
Check for duplicate rows in the dataset	Inoder to avoid bias in the training and prevent overfitting,I think it's a good practice to check whether the data/entire row is duplicated in the dataset.	5	Easy	1_Data_Cleaning.ipynb	Both Alzheimers Data & Student Performance Data
Plot the distribution of the Diagnosis column	Inference: The dataset shows a moderate imbalance in Alzheimer's diagnosis, with more "No" cases than "Yes" cases, but the distribution is not extreme enough to severely impact model performance.	10	Easy	2_Exploratory_Data_Analysis.ipynb	Alzheimers Data
Plot the top 10 countries with the highest Alzheimer's cases	Inference: The United States has the highest number of Alzheimer's cases among the top 10 countries, followed by China, Brazil, and India.	10	Easy	2_Exploratory_Data_Analysis.ipynb	Alzheimers Data
Plot the gender distribution for diagnosed cases	Inference: The diagnosed cases are nearly evenly split between females (50.1%) and males (49.9%), indicating minimal gender disparity.	10	Easy	2_Exploratory_Data_Analysis.ipynb	Alzheimers Data
Plot the distribution of age groups by diagnosis	I created distinct age brackets for better understanding of the data. Inference: The data shows a higher concentration of diagnosed cases among older age groups (50+). Younger brackets have comparatively fewer diagnoses, indicating an age-related trend in Alzheimer's incidence.	15	Easy	2_Exploratory_Data_Analysis.ipynb	Alzheimers Data
Plot the distribution of the Family History of Alzheimer's	At first, I counted how many people have or do not have a family history of Alzheimer's (family_history_counts) and creates a single pie chart to show that distribution. Then I created two pie charts side by side to compare family history for those diagnosed with Alzheimer's versus those not diagnosed.	20	Easy	2_Exploratory_Data_Analysis.ipynb	Alzheimers Data
Categorize BMI and plot the distribution by diagnosis	I categorized BMI into different groups, grouped the data by BMI category and diagnosis status, and plotted a grouped bar chart to visualize the distribution of BMI categories by diagnosis	15	Easy	2_Exploratory_Data_Analysis.ipynb	Alzheimers Data
Identify numerical & Categorical columns	Different types of data required different type of processinf techniques. For example, we need to label encode categorical variables.	5	Easy	2_Exploratory_Data_Analysis.ipynb	Alzheimers Data
Plotted Relation between hours studied and performance index	Inference: The line plot shows a positive correlation between hours studied and average performance index, indicating that increased study hours are associated with better performance	10	Easy	2_Exploratory_Data_Analysis.ipynb	Student Performance Data

Plot the relation between sleep hours and performance index	At first, I grouped the data by sleep hours and then calculated the average performance index of each category. Then plotted a bar graph to visualize the relationship between sleep hours and performance. Also, the bar plot suggests that sleep hours have a relatively stable impact on performance index	10	Easy	2_Exploratory_Data_Analysis.ipynb	Student Performance Data
correlation heatmap	The correlation heatmap shows that previous scores have the strongest positive correlation with the performance index, while hours studied also have a moderate positive correlation, indicating their importance in predicting student performance.	10	Easy	2_Exploratory_Data_Analysis.ipynb	Student Performance Data
Outlier Detection	It improves model accuracy and reliability by identifying and handling abnormalities/anomalies in the data. There were no outliers detected.	5	Easy	2_Exploratory_Data_Analysis.ipynb	Both Alzheimers Data & Student Performance Data
Normalize the numerical columns using StandardScaler()	Normalization ensures that features have a mean of 0 and a standard deviation of 1, which can improve the model performance especially for KNN like models.	15	Easy	3_Data_Preprocessing.ipynb	Both Alzheimers Data & Student Performance Data
Label Encoding	Converts Categorical varibles to numerical format, allowing ml models to process and interpret categorical data while maintaining the ordinal reltionships if present	5	Easy	3_Data_Preprocessing.ipynb	Both Alzheimers Data & Student Performance Data
Save the processed data to a CSV file	Ease of use	2	Easy	3_Data_Preprocessing.ipynb	Both Alzheimers Data & Student Performance Data
Split the data into training (70%) and remaining (30%)	Data = Training (70%) + Remaining (30%)	5	Easy	3_Data_Preprocessing.ipynb	Both Alzheimers Data & Student Performance Data
Split the remaining data into test (50% of remaining) and validation (50% of remaining)	Remaining data = test (50%) + validation (50%)	5	Easy	3_Data_Preprocessing.ipynb	Both Alzheimers Data & Student Performance Data
Define the features and target variable for training, testing & validation	get the variables ready for training, testing and validation	5	Easy	3_Data_Preprocessing.ipynb	Both Alzheimers Data & Student Performance Data
Save the features and target for training, testing & validation as pickle file	Saved the required variables as pickle file to preserve the structure and datatype	10	Easy	3_Data_Preprocessing.ipynb	Both Alzheimers Data & Student Performance Data
Load the features and target for training, testing & validation	Loading Pickle data	5	Easy	4_Linear_Regression.ipynb	Student Performance Data
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	R^2 is close to 1 and the model has low RMSE (Root mean square error) for both test data and validation data, suggests strong predictive accuracy and minimal error.	20	Intermediate	4_Linear_Regression.ipynb	Student Performance Data
Plot actual vs predicted for test data	Shows strong linear relation along the diagonal, indicates good model	10	Easy	4_Linear_Regression.ipynb	Student Performance Data
Plot actual vs predicted for validation data	Shows strong linear relation along the diagonal, indicates good model	10	Easy	4_Linear_Regression.ipynb	Student Performance Data

Load the features and target for training, testing & validation	Loading Pickle data	5	Easy	5_Logistic_Regression.ipynb	Alzheimers Data
Created a function calculate_rmse_and_predict() to calculate the rmse value and predict the target values	Function to calculate rmse and predict the target values (Ease of Use - At first, I used a single file for all models)	10	Easy	5_Logistic_Regression.ipynb	Alzheimers Data
Created a function train_and_evaluate_logistic_regression() for logistic regression model, Calculate the Test Accuracy, and Calculate RMSE for test & validation data	Model training , finding accuracy and RMSE for both test and validation data.Shows good classification performance	25	Intermediate	5_Logistic_Regression.ipynb	Alzheimers Data
Save the accuracy, RMSE on test & validation data as a pickle file	Ease of use (for comparison)	5	Easy	5_Logistic_Regression.ipynb	Alzheimers Data
Load the features and target for training, testing & validation	Loading Pickle data	5	Easy	6_Decision_Trees.ipynb	Alzheimers Data
Created a function calculate_rmse_and_predict() to calculate the rmse value and predict the target values	Function to calculate rmse and predict the target values (Ease of Use - At first, I used a single file for all models)	10	Easy	6_Decision_Trees.ipynb	Alzheimers Data
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	Here, I did hyperparameter tuning, specifically depth optimization using cross validation to find the optimal depth that gives best accuarcy while avoiding overfitting. Optimal depth was found to be 5.	20	Intermediate	6_Decision_Trees.ipynb	Alzheimers Data
Train the Decision Tree Classifier with the optimal depth, Predict on the test data, Calculate the accuracy, and Calculate RMSE for test & validation data	Model training , finding accuracy and RMSE for both test and validation data.Shows good classification performance	15	Intermediate	6_Decision_Trees.ipynb	Alzheimers Data
Plot the decision tree with a depth of 2 levels (opted 2 for better visualization)	I tried plotting decision tree with optimal depth, since I got a small diagram and have many nodes, I changes the depth to 2 for better visualization.	10	Intermediate	6_Decision_Trees.ipynb	Alzheimers Data
Update the accuracy, RMSE on test & validation data to the pickle file	Ease of use (for comparison)	5	Easy	6_Decision_Trees.ipynb	Alzheimers Data
Load the features and target for training, testing & validation	Loading Pickle data	5	Easy	7_Random_Forest.ipynb	Alzheimers Data

Created a function calculate_rmse_and_predict() to calculate the rmse value and predict the target values	Function to calculate rmse and predict the target values (Ease of Use - At first, I used a single file for all models)	10	Easy	7_Random_Forest.ipynb	Alzheimers Data
Initialize the Random Forest Classifier, Train the classifier, Predict on the test data, Calculate the accuracy and Calculate RMSE for test & validation data	Model training , finding accuracy and RMSE for both test and validation data.Shows good classification performance	10	Intermediate	7_Random_Forest.ipynb	Alzheimers Data
Update the accuracy, RMSE on test & validation data to the pickle file	Ease of use (for comparison)	5	Easy	7_Random_Forest.ipynb	Alzheimers Data
Load the features and target for training, testing & validation	Loading Pickle data	5	Easy	8_KNN.ipynb	Alzheimers Data
Created a function calculate_rmse_and_predict() to calculate the rmse value and predict the target values	Function to calculate rmse and predict the target values (Ease of Use - At first, I used a single file for all models)	10	Easy	8_KNN.ipynb	Alzheimers Data
Performed K Fold Cross Validation to find the optimal k value	K fold Cross validation to find the optimal value, which was found to be 17.	15	Intermediate	8_KNN.ipynb	Alzheimers Data
Created a function train_knn_model() for KNN model	For Ease of use (Function)	15	Intermediate	8_KNN.ipynb	Alzheimers Data
Implemented KNN with Euclidean distance & Calculated RMSE for test & validation data	With Euclidean Distance	5	Easy	8_KNN.ipynb	Alzheimers Data
Implemented KNN with Manhattan distance & Calculated RMSE for test & validation data	With Manhattan Distance	5	Easy	8_KNN.ipynb	Alzheimers Data
Implemented KNN with Minkowski distance & Calculated RMSE for test & validation data	With Minkowski Distance	5	Easy	8_KNN.ipynb	Alzheimers Data
Update the accuracy, RMSE on test & validation data to the pickle file	Ease of use (for comparison)	5	Easy	8_KNN.ipynb	Alzheimers Data
Load the features and target for training, testing & validation	Loading Pickle data	5	Easy	9_SVM.ipynb	Alzheimers Data

Created a function calculate_rmse_and_predict() to calculate the rmse value and predict the target values	Function to calculate rmse and predict the target values (Ease of Use - At first, I used a single file for all models)	5	Easy	9_SVM.ipynb	Alzheimers Data
Created a function train_and_evaluate_svm() for SVM model	For Ease of use (Function)	25	Intermediate	9_SVM.ipynb	Alzheimers Data
Implemented SVM with linear kernel & Calculated RMSE for test & validation data	Implemented using Linear Kernel	5	Easy	9_SVM.ipynb	Alzheimers Data
Implemented SVM with RBF kernel & Calculated RMSE for test & validation data	Implemented using rbf Kernel	5	Easy	9_SVM.ipynb	Alzheimers Data
Update the accuracy, RMSE on test & validation data to the pickle file	Ease of use (for comparison)	5	Easy	9_SVM.ipynb	Alzheimers Data
Load the features and target for training, testing & validation	Loading Pickle data	5	Easy	10_Naive_Bayesian.ipynb	Alzheimers Data
Created a function calculate_rmse_and_predict() to calculate the rmse value and predict the target values	Function to calculate rmse and predict the target values (Ease of Use - At first, I used a single file for all models)	5	Intermediate	10_Naive_Bayesian.ipynb	Alzheimers Data
Applied the Gaussian Naive Bayes model & Calculated Accuracy, Calculate RMSE for test & validation data	With Gaussian NB	15	Easy	10_Naive_Bayesian.ipynb	Alzheimers Data
Applied the Bernoulli Naive Bayes model & Calculated Accuracy, Calculate RMSE for test & validation data	With Bernoulli NB	15	Easy	10_Naive_Bayesian.ipynb	Alzheimers Data
Update the accuracy, RMSE on test & validation data to the pickle file	Ease of use (for comparison)	5	Easy	10_Naive_Bayesian.ipynb	Alzheimers Data
Import the dataset and remove Diagnosis column for clustering	Loading Pickle data and Removed the target variable to use unsupervised learning algorithm	5	Easy	11_K_Means_Clustering.ipynb	Alzheimers Data
Applied Elbow method to find the optimal number of clusters	Elbow Method to find optimal k , which was found to be 2	15	Intermediate	11_K_Means_Clustering.ipynb	Alzheimers Data
Applied K Means Clustering with k=2	K means clustering with k=2	10	Intermediate	11_K_Means_Clustering.ipynb	Alzheimers Data

Predict the clusters	Predicted the output	10	Intermediate	11_K_Means_Clustering.ipynb	Alzheimers Data
Feature Engineering	I tried to Implement Feature Engineering using Random Forest Classifier to improve the accuracy.But the accuracy remained almost same. I tried using these new variables to get accuracy for each algorithm and then it got very complicated to manage and I dropped them.	40	Difficult	3_Data_Preprocessing.ipynb	Alzheimers Data