Team 8 Notes :

Step followed for Data understanding**:**

The Alzheimer's dataset consists of **2,84,142** rows and **31** columns, which include year start, year-end, LocationAbbr, LocationDesc, Datasource, class, Topic, Question, Data_value, Data_value_Type, Data_value_Alt, Low cofidence_limit, High confidence_limit, Stratification1, Stratification2, TopicID, ClassID.

We initially loaded the dataset into the Jupyter Notebook file and began the data preparation. We started by checking for missing values using code to identify the null values in the column. And checked the info on each column

Steps followed for Data Preprocessing:

Removed Irrelevant columns:

The columns like Datasource, LocationAbbr, Geolocation, stratification category1 etc were dropped since they did not serve any purpose for the analysis

Renamed columns:

Stratification1 has been renamed as Age Group for better understanding.

Identified columns with missing values like Data_Value, Low_confdence_Limit, High_confidence_Limit and filled "unknown" in the missing values of Stratification2.

All the rows containing Data_Value_Type as 'Mean' have been removed since this will not be useful for analysis

Reviewing **Missing Data**:

Checked the total number of missing values, particularly in critical columns like Data_Value and Confidence_Limits.

Final Dataset:

Validated the dataset through the generation of summary statistics, ensuring it was ready for modeling

Creation of **Synthetic Column**:

We created a synthetic column named Alzheimers by using the survey data and the confidence level in their responses.

Logic used:

Set 1 : Negative Questions

If the response for a negative question(for example, unhealthy habits) has a Data_value value graeter than mean(39.74) the synthetic column marks as "Yes" for higher chances of Alzheimers

Set 2: Positive Questions

Foe positive questions- abnormal habits the column marks it as "yes" if the data_values is leeds than the mean(39.74)

Confidence Bounds

Confidence Limits define how sure the answer is

High confidence: When the high limit> low limit -more sure. Low confidence when the low limit> high limit less sure. This synthetic column enhances the estimation of Alzhheimer's risk by combining data values, question type and confidence.

EDA:

Line Graph: We have a year column, where we filtered based on the Alzheimer's cases and presented the line graph

Bar chart: based on the Alzheimer cases whether they have alzheimer then it is 'Yes' and if they don't have alzheimer 'No'

Grouped Bar Chart : based on the count of Alzheimers (yes/no) by Race/ethnicity, presenting as a sample of these Race/ethnicity on graph Asian or Pacific Islander, African American, not Hispanic, Latino, Indigenous American/Alaskan Indigenous, Caucasian, non-Latino.

Stacked Bar chart: Based on the gender count did the Distribution of Alzheimer’s cases where the condition is yes/no.
Stacked Bar Chart: We considered the count of Alzheimer's cases as Yes as the condition we presented based on the regions.
Correlation matrix: here we used a heat map; a feature correlation matrix is obtained which can then be used to identify the levels of feature values that correspond to the other measure.

Modeling

For modeling firstly we preprocessed the categorical variables including ‘Alzheimers’,’Age group and ‘stratification 2. These were changed into numeric values using Label Encoder to allow the machine learning model to handle them. For feature selection we used followed correlation based selection The features were chosen depending on their corelation with the target variable ‘Alzheimers’ Strongly correlated features were selected to enhance model performance while excluding irrelevant ones

Model evaluation and selection:
First of al, several machine learning models such as Random Forest, KNN, SVM, Logistic regression have been tested. Feature selection based on the correlation with target variable was performed before the model training as well as imputation of missing values to ensure that the dataset is clen and ready for analysis

Model performance:
Overfitting concerns:
Some of the models that we used, especially KNN and Random Forest, show signs of overfitting that may make them unreliable for unseen data. That is the model does well on the training data but fails in new real-world data

KNN
Validation Accuracy:98.68% Test Accuracy:98.72%
These results demonstrate that KNN went very well. However KNN usually suffers from overfitting in the presence if noisy data

Support Vector Machine (SVM):
The model shows consistency in results with a validation accuracy of 82.31%
Precision, recall and F1-scores of the model provide evidence that it generalizes well with no overfitting.
qThe usage of SVM is highly reliable in this case as the model is robust in handling high dimensional data and works great when the dataset size is low.
Because the other models overfitted or didn’t generalize well and hence cannot be very reliable on unseen data.
KNN and Random Forest did a great job but raised some flags on overfitting.
SVM was chosen since it balanced the trade-off between accuracy, generalization, and reliability.
This ability of the model to avoid overfitting and provide robust predictions across datasets made it the fittest for our problem.

Prediction:
We used SVM for prediction The model predicted the possibility of Alzheimer’s disease based on given features of demographic and cognitive health metrics.
For this prediction, based on the input features, the model predicts no signs of Alzheimer’s, hence the individual is cognitively healthy and safe.
These results pinpoint the model’s capability for reliable insights into early detection and informed decision making