1. question/research statement - K
   Through this project study and analysis, we focused on exploring two key questions:
   a. **Can our optimized predictive model accurately diagnose Alzheimer's disease in individuals?**
      The primary goal of this question was to leverage machine learning techniques to develop a predictive model that was capable of most accurately diagnosing Alzheimer's based on a range of predictors. These predictors include demographic information, lifestyle factors, medical history as well as the cognitive assessments. We aimed for this model to be able to provide a valuable tool for healthcare professionals, and help them to evaluate the probability of Alzheimer's disease early in its progression. This helped us spur onto the next question.
   b. **Can correlations between individuals' demographic information, lifestyle factors, medical history, and symptoms, and their clinical measurements and cognitive assessments be identified to inform recommendations for appropriate diagnostic tests?**
      Our goal was to investigate the relationships between patient attributes and Alzheimer's diagnosis, to streamline the diagnostic process. Through this analysis we were able to uncover some potential meaningful patterns between a broad spectrum of patient variables and Alzheimer's diagnosis. cInsights from these correlations may guide recommendations for prioritizing specific diagnostic tests based on patient profiles.
2. background/value -K
   Introduction:
   Alzheimer's disease is a progressive neurological disorder that severely impacts cognitive and functional abilities. Early and accurate diagnosis of the disease is essential for timely intervention, effective management, and better quality of life for patients.
   a. The practical impact of research question 1 could be reporting these findings to doctors or hospitals to help evaluate the probability of Alzheimer's using machine learning techniques
      Alzheimer's disease is a leading cause of disability and dependency in the elderly. However, its diagnosis remains challenging,and requires a combination of clinical, imaging, and laboratory assessments. Machine learning offers a promising avenue to improve diagnostic accuracy by analyzing large datasets with numerous variables, identifying patterns that are not only harder to understand through traditional methods, but might offer a way for early detection.
3. data - source/ EDA - B
   a. Our dataset comprises detailed health information for 2,149 patients. With 33 predictors spanning demographic details, lifestyle factors, medical history, clinical measurements, cognitive and functional assessments, symptoms, and Alzheimer's Disease diagnoses, our response variable is binary. Very lucky, there are no missing values, and our heat map revealed no multicollinearity issues among the variables. No abnormalities or outliers as well.

b. We draw violin plots for all continuous variables and side-by-side bar plots for all categorical or ordinal variables. We observed that variables like gender had similar proportions between diagnosed and undiagnosed groups, suggesting minimal predictive value. Individuals with higher education levels and those of Asian ethnicity showed a slightly higher likelihood of an Alzheimer's diagnosis. Interestingly, certain findings were counterintuitive: patients without a family history of Alzheimer's had a higher diagnosis rate (36%) compared to those with a family history (32%). Symptoms traditionally associated with Alzheimer's—such as confusion, disorientation, and forgetfulness—did not show significant differences in our data.
c. Continuous variables like age and BMI did not display significant differences either. However, scores from functional assessments and activities of daily living were notably higher in patients without Alzheimer's, indicating their potential as predictive features.
d. Given these insights, we found it challenging to correlate personal information directly with examination results to recommend specific diagnostic tests. Therefore, we shifted our focus to developing a machine learning model to improve diagnostic accuracy.

4. model
    a. variable selection:
        i. Given the complexities and surprising findings from our descriptive analysis—where many variables showed minimal predictive value for Alzheimer's Disease, and the imbalance in diagnostic groups (with significantly more patients not diagnosed with the disease than those who are)—we recognized the need to refine our model for better accuracy. Therefore, we employed best subset selection and ANOVA to identify the 10 most impactful features, filtering out less contributive predictors to enhance model performance.
    b. model selection:
        i. since we have binary predictors, we choose supervised learning -> classification methods (logistic regression, decision tree, XGBoost, and Random forest), since XGBoost has high accuracy, meaning ensemble methods work for this scenario, so we also explore the stacking method
        ii. Add Ridge and Lasso/ Elastic Net
        These models have low accuracies even after hyperparameter tuning. Random Forest still has the best performance

    c. model training, hyperparameters tuning
        i. Ran models after categorizing features into 4 different categories: demographic, lifestyle, medical, and cognitive
        ii. `Accuracy DataFrame:`

	Logistic Regression	Elastic Net	Random Forest	XGBoost
Demographic	0.646512	0.646512	0.646512	0.646512
Lifestyle	0.679070	0.679070	0.672093	0.676744
Medical	0.693023	0.693023	0.648837	0.665116

```
Cognitive              0.795349     0.795349       0.837209  0.837209
```

   d. model performance
      i. Overall Random Forest has the best accuracy score with a score of 0.952
      ii. Cognitive features performed the best and had similar accuracy with XGBoost and Random Forest. (add visualization image)
5. final model/output -G
   a. we can decide these variables are important to predict Alzheimer existence (DONE)
   b. compare different models (DONE)
   c. our best predictive model achieve (DONE)

Correctly diagnosing Alzheimer's disease has a significant impact on healthcare, as early detection enables timely intervention and improves patient outcomes. Our project utilized machine learning techniques to predict Alzheimer's diagnosis based on a dataset containing demographic variables, health metrics, lifestyle factors, and cognitive assessments. These predictors collectively provided a comprehensive view of factors contributing to AD risk, with the five cognitive and functional assessments emerging as the most significant contributors. Surprisingly, age is not a significant factor according to the results of the Student's t-test, which provides insufficient evidence that the mean age between the diagnosed group and the non-diagnosed group differs.

After testing various machine learning algorithms, Random Forest was the best predictive model, achieving the highest accuracy of 95.35% and demonstrating robust performance in capturing complex feature interactions. This model outperformed others such as Logistic Regression, Decision Tree, and XGBoost. While the approach has been validated on structured datasets in controlled settings, future work will focus on real-world validation using diverse populations over extended periods. Feedback from these trials will help refine the model and further enhance its predictive capabilities.

Additionally, we trained a model that excluded cognitive and functional assessment scores, focusing solely on demographic, medical history, lifestyle, and symptoms. This model achieved an accuracy of 71.16% using XGBoost. This result demonstrates the potential for early-stage Alzheimer's diagnosis in scenarios where detailed clinical measurements and assessments may not be immediately accessible.

In summary, our project highlights the potential of machine learning models for early Alzheimer's detection, leveraging cognitive and health data to identify at-risk individuals. With continued development and validation, such models could significantly increase diagnostic precision, improve patient care, and reduce the societal and economic burden of Alzheimer's disease.

6. what do we learn from the project
    a. From dividing tasks to integrating findings, we developed communication and coordination skills essential for collaborative problem-solving.
    b. Working on a healthcare-related project, it is important to concern the ethical implications of the model, such as the potential impact of false positives or negatives in medical diagnoses.
    c. We discovered that age was not a significant factor for Alzheimer's diagnosis, which challenged our preconceptions and taught us to rely on data-driven insights rather than assumptions.