# Alzheimer's Risk Prediction

Andrew Jowe
919586453
ajowe@ucdavis.edu

Kaushal Marimuthu
920979842
kmarimuthu@ucdavis.edu

Brien Pike
922092730
bdpike@ucdavis.edu

Devyn Simmons
919526646
dksimmons@ucdavis.edu

Lejia Xu
918195599
aijxu@ucdavis.edu

**Abstract**

This analysis fits and compares four classification models—logistic regression, random forest, support vector machine (SVM), and artificial neural network (ANN)—to predict Alzheimer's disease risk using a publicly available clinical dataset of 2,149 patients and 32 variables. The variables are a mixture of demographic, cognitive, and lifestyle features. Data preprocessing combined one-hot encoding for categorical variables and z-score normalization for continuous predictors. Hyperparameters were optimized via grid search with stratified k-fold cross-validation. The models were evaluated on an 80/20 train/test split using accuracy, precision, recall, and F1 score. Logistic regression provided a robust baseline (test F1 = 0.787), while random forest achieved the highest generalization performance (test F1 = 0.928; accuracy = 94.9%), identifying ethnicity, smoking status, gender, education level, BMI, and age as the most important factors for predicting Alzheimer's risk. Both SVM (test F1 = 0.758) and ANN (test F1 = 0.713) exhibited overfitting, suggesting the need for further regularization or expanded data. Overall, random forest demonstrates strong potential for early Alzheimer's risk stratification, and future work should explore larger cohorts and deep learning architectures to improve predictive reliability and clinical applicability.

## Contents

## 1 Introduction

Alzheimer's disease is a progressive neurodegenerative disorder that affects millions of people around the world. Because early diagnosis is essential for improving patient outcomes, our project goal is to develop a predictive model that can evaluate the likelihood of a patient developing Alzheimer's by using a wide array of medical and lifestyle information. By leveraging various machine learning techniques we aim to enhance risk assessment accuracy and provide valuable insights for early intervention. This work is crucial because clinical symptoms often surface only after irreversible cognitive decline has begun. In essence, our primary objective is the early prediction of Alzheimer's risk using patient data and our central research questions are:

- What patient characteristics and clinical indicators are most strongly correlated with the risk of developing Alzheimer's disease?
- How accurately can different machine learning models predict Alzheimer's onset based on the available features?
- Which machine learning method delivers the most effective and reliable predictions for this dataset?

These questions are critical since early detection can enable timely therapeutic interventions, potentially slowing disease progression and enhancing the quality of life for patients.

For our analysis, we will be using the Alzheimer's Disease Dataset from Kaggle. This dataset comprises clinical records that include a variety of features such as demographic details, cognitive test results, biomarkers, and additional clinical measurements typically employed in assessing Alzheimer's risk. Each row corresponds to an individual patient, and each column represents a distinct attribute relevant to their neurological health. The target variable is binary, indicating whether a patient is at high risk (1) or not at risk (0) for developing Alzheimer's disease. Early detection through this model could play a pivotal role in enabling prompt medical intervention, potentially slowing the trajectory of the disease and ultimately saving lives.

## 2 Exploratory Data Analysis

### Data Preparation, Preprocessing and Visualization

For our exploratory data analysis, we looked at a dataset of 2,149 patients and 32 predictors that includes a wide range of health, lifestyle, and cognitive information. The goal was to explore how different features relate to Alzheimer's diagnosis. The dataset was already clean, with no missing values, which made it easy to work with. It includes both numerical and categorical variables like age, gender, education level, cholesterol levels, sleep quality, and cognitive symptoms like forgetfulness and memory complaints. We used one-hot encoding to convert the categorical variables for analysis.

From the count plot (Figure 1), we saw that most patients in the dataset were not diagnosed with Alzheimer's, meaning the data has a class imbalance. In our model, we need to ensure that the training and test splits have a fair balance of both classes.

One of the most important findings came from the correlation heatmap and grouped statistics: patients diagnosed with Alzheimer's tended to have lower MMSE scores and lower functional assessment scores, which makes sense since those are tied to cognitive ability. They also showed higher rates of memory complaints, behavioral problems, and difficulty with daily tasks. Interestingly, physical health features like blood pressure and cholesterol didn't vary much between groups.

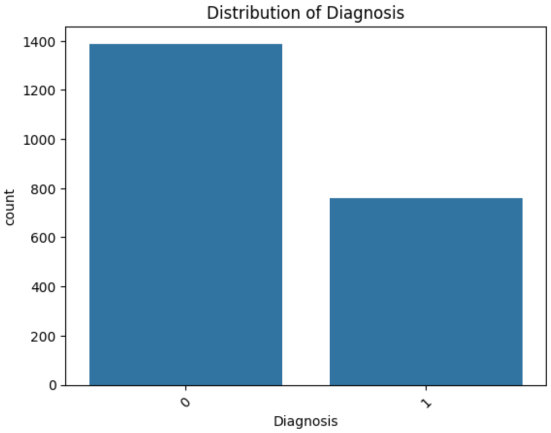

Figure 1: Distribution of Alzheimer's diagnosis showing class imbalance.

Histograms and boxplots helped us visualize how features like BMI, alcohol use, and MMSE scores were distributed. We found no issues with outliers. We also looked at gender and found that it was pretty evenly distributed, which means that there wasn't a clear difference in diagnosis rates between men and women. When we looked at age, we noticed that people diagnosed with Alzheimer's were generally older, though there was still a lot of overlap in ages between groups.

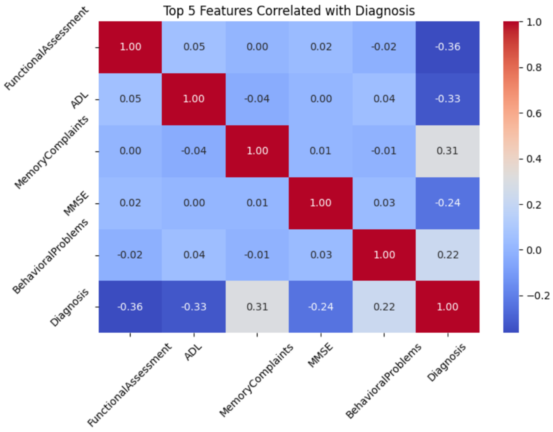

Figure 2: Correlation heatmap showing relationships between key numerical features.

Overall, the EDA gave us a clear picture of which features are more strongly connected to Alzheimer's and will help guide our next steps in modeling and prediction.

# 3 Methodology

### Model Fitting and Cross-Validation

All models used the same preprocessing. Categorical features were one-hot encoded, and numerical features were standardized using z-score normalization. We used a `ColumnTransformer` to combine these steps, which were embedded in a `Pipeline` along with each model estimator We used a standard experimental setup across all our models in order to ensure consistency. We split the data into training and testing at an 80/20 ratio. For each model, we performed hyperparameter tuning with `GridSearchCV` in order to identify the optimal parameters. We then trained the model on the training set using these settings. We evaluated performance on both training and test, and compared them in order to assess how the model generalizes and whether it overfits. Our performance metrics were then captured for display.

### 3.1 Logistic Regression

We used logistic regression as a baseline due to its simplicity and strong performance in binary tasks. `GridSearchCV` was used to find the best solvers (e.g., `liblinear`, `lbfgs`) and regularization strengths ($\alpha$), with repeated stratified 5-fold cross-validation. The best model, `lbfgs` with $\alpha = 0.5$, achieved an F1-score of 0.752. We created a model with these parameters, fit it to the training data, and evaluated it on both the test and training set.

### 3.2 Random Forest

We used a Random Forest Classifier in order to capture the interactions between variables in hopes to increase our predictability. Random forests are essentially a set of decision trees, they generate multiple during training and output the class that is the mode of the predictions of each tree. This method is ideal for high-dimensional, structured data like ours, as it can help prevent overfitting while maintaining accuracy.

To optimize this model, we performed grid search using 5-fold stratified cross-validation. Our hyperparameter grid included:

- `n_estimators`: Number of trees in the forest (25 to 175)
- `max_depth`: Maximum tree depth (None or 5 to 15)
- `min_samples_split`: Minimum number of samples required to split an internal node (2 to 9)
- `min_samples_leaf`: Minimum number of samples required at a leaf node (1 to 13, odd numbers)
- `max_features`: Number of features to consider at each split (sqrt, log2)

This search found the best model to be `n_estimators = 100`, `max_depth = 10`, `min_samples_split = 2`, `min_samples_leaf = 1`, and `max_features = log2`, with an F1 score of 0.932.

We also performed feature selection, and found:

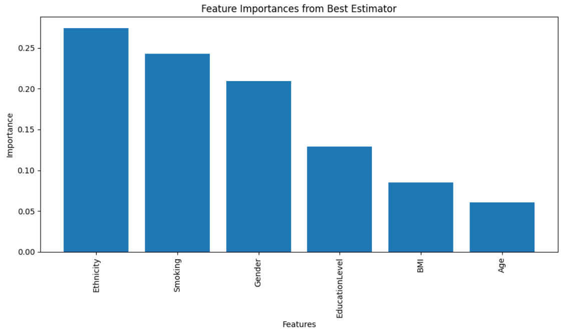

Figure 3: Distribution of importance among features.

We used these parameters and features in our model, fit it to the training data, and compared the train and test results.

### 3.3 Support Vector Machine

We implemented a support vector machine (SVM) using the kernel trick to model non-linear decision boundaries. Four kernels—linear, RBF, polynomial, and sigmoid—were compared using repeated stratified 10-fold cross-validation. The RBF kernel achieved the best performance and was selected for the final model. `GridSearchCV` was used to find the best $\gamma$ and regularization strengths ($\alpha$), with repeated stratified 5-fold cross-validation. This gave us $\gamma = 0.01$ and $\alpha = 1$ with an F1 score of 0.772. We also applied an imbalance factor of 2 for class 1, essentially making class 1 results twice as important. We refit the SVM on these parameters, and evaluated it on both the training and test set.

### 3.4 Artificial Neural Network

We trained an Artificial Neural Network (ANN) using `MLPClassifier` from `scikit-learn` to model non-linear relationships in the heart disease dataset. We applied a comprehensive grid search using 5-fold cross-validation to tune hyperparameters. The grid included:

- `hidden_layer_sizes`: Single and multi-layer configurations such as (100,), (150, 75), and (100, 50, 25)
- `activation`: `relu`, `tanh`
- `alpha`: Regularization strengths from $10^{-5}$ to $10^{-1}$
- `learning_rate_init`: Initial learning rates from 0.0001 to 0.1
- `solver`: Fixed as `adam`
- `max_iter`: 5000

The best model configuration used `relu` activation, one hidden layer of size (100,), $\alpha = 0.001$, and a learning rate of 0.01. It achieved a cross-validated F1 score of 0.759.

We then refit this data to the training data, and compared the results from training and testing.

## 4 Results

**Comparison of Various Methods**

We compared the performance of four machine learning models—Logistic Regression, Random Forest, Support Vector Machine (SVM), and Artificial Neural Network (ANN)—in predicting Alzheimer's disease. Model performance was evaluated using key classification metrics on both training and testing datasets.

### 4.1 Logistic Regression

| Metric | Train Set | Test Set |
|---|---|---|
| Accuracy | 0.8284 | 0.8349 |
| Precision | 0.7233 | 0.7238 |
| Recall | 0.8339 | 0.8618 |
| F1 Score | 0.7746 | 0.7868 |

Table 1: Logistic Regression Evaluation Metrics for Train and Test Sets

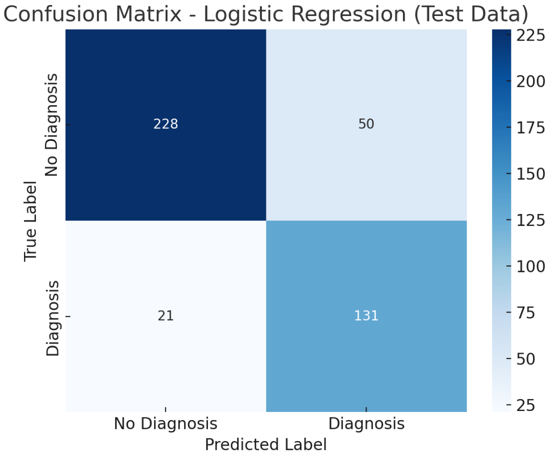

Figure 4: Logistic Regression Confusion Matrix.

Logistic Regression showed balanced performance with strong generalization. The recall of 86.2% suggests decent sensitivity to identifying Alzheimer's cases. As a sanity check, we compared the train set accuracy and test set accuracy to be roughly the same, signaling no overfitting.

### 4.2 Random Forest Performance

| Metric | Train Set | Test Set |
|---|---|---|
| Accuracy | 0.9825 | 0.9488 |
| Precision | 0.9865 | 0.9276 |
| Recall | 0.9638 | 0.9276 |
| F1 Score | 0.9750 | 0.9276 |

Table 2: Artificial Neural Network (ANN) Evaluation Metrics for Train and Test Sets

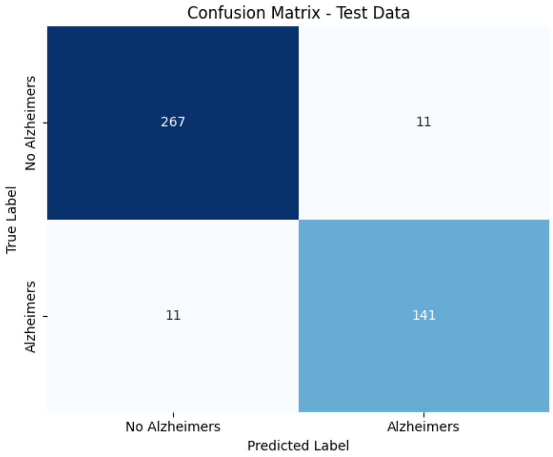

Figure 5: Random Forest Confusion Matrix.

Random Forest achieved the highest performance across all metrics, particularly excelling in precision (98.65%) and generalization. Its robust test accuracy (94.9%) .

### 4.3 Support Vector Machine

| Metric | Train Set | Test Set |
|---|---|---|
| Accuracy | 0.9453 | 0.8116 |
| Precision | 0.8813 | 0.6940 |
| Recall | 0.9770 | 0.8355 |
| F1 Score | 0.9267 | 0.7582 |

Table 3: SVM Evaluation Metrics for Train and Test Sets

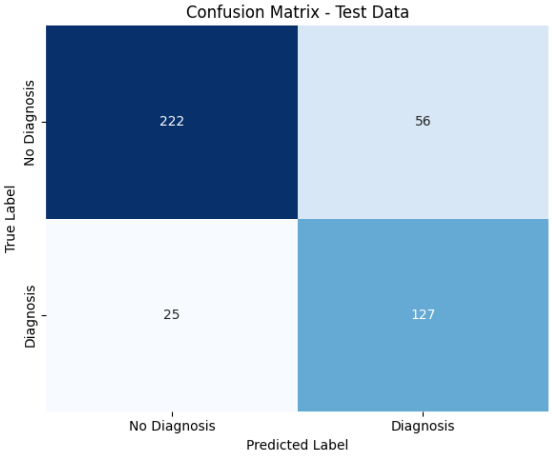

Figure 6: SVM Confusion Matrix.

SVM performed well on the training data but showed signs of overfitting, with a notable drop in test set precision (69.4%). Further tuning may improve its ability to generalize, but due to computation limitations we could not attempt a larger parameter grid.

ANN severely overfit the training data, with nearly perfect training performance but reduced test set accuracy (80.0%). This suggests the model is too flexible for the data size or configuration. Further tuning may improve its ability to generalize, but due to computation limitations we could not attempt a larger parameter grid.

### 4.4 Artificial Neural Network

| Metric | Train Set | Test Set |
|---|---|---|
| Accuracy | 0.9965 | 0.8000 |
| Precision | 0.9967 | 0.7230 |
| Recall | 0.9934 | 0.7039 |
| F1 Score | 0.9951 | 0.7133 |

Table 4: ANN Evaluation Metrics for Train and Test Sets

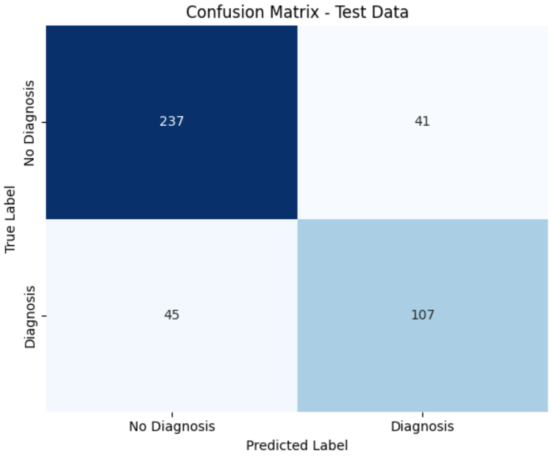

Figure 7: ANN Confusion Matrix.

## 5 Conclusion

This analysis aimed to address three primary research questions: which patient characteristics are most associated with Alzheimer's risk, how accurately machine learning models can predict Alzheimer's onset, and which method yields the most reliable predictions. We evaluated four models—Logistic Regression, Random Forest, SVM, and ANN—using the preprocessing pipeline on a dataset that included demographic, cognitive, and lifestyle features. The Random Forest classifier outperformed the other models and achieved the best results in every single test metric which supports its robustness in clinical prediction tasks. Additionally, the Random Forest model incorporated feature selection and found that ethnicity, smoking, gender, education level, BMI, and age were the most important predictors in dianosing Alzheimer's risk. Logistic Regression provided a strong, interpretable baseline but the performance from SVM and ANN were both disappoing. Both models severely overfit the training data and would likely benefit from more hyperparameter optimization. Performance was an issue in implementing a thorough grid search for optimal hyperparameters for both models, especially ANN, and access to more compute power would likely provide significant improvements to the performance of both models. Overall, the strong results of the Random Forest classifier underscore its potential to support early detection and improve clinical decision-making.

# 6 Appendix

- **GitHub Code Repository:** https://github.com/kaushal2m2/sta141c-final-proj
- **Kaggle Dataset:** https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset