Predictive Analytics: Machine learning with Synthetic Data

Student Name: ABDULAZIZ HISHAM ALMAKHDHOUB

Course: CS465 - Machine Learning

Date: April 25, 2025

Abstract

This report presents a comprehensive machine learning study on synthetic Alzheimer's Disease dataset of 2,149 patients. We developed a full predictive analytics pipeline encompassing data preprocessing, exploratory data analysis (EDA), feature engineering, model training, and evaluation. Key objectives included comparing multiple classification algorithms (Logistic Regression, Support Vector Machine, Decision Tree, and a Multi-Layer Perceptron neural networks) and extracting insights into which features most strongly relate to Alzheimer's diagnosis. Data preparation involved handling missing values (none were present in this dataset), encoding categorical variables, and normalizing continuous features. EDA revealed that cognitive and functional assessment scores, as well as memory-related symptoms, have the strongest association with Alzheimer's diagnosis, whereas demographic and metabolic features show weak correlation with the outcome. WE performed feature selection using a Random Forest-based importance filter and created additional interaction features to enhance linear models. Model performance was evaluated with accuracy, precision, recall, F1-score, and ROC-AUC metrics, using consistent train/test split and cross-validation for hyperparameter tuning. The decision Tree classifier emerged as the best-performing model (F1=0.911, ROC-AUC=0.934), slightly outperforming the neural network. McNeymar's test confirmed no statistically significant difference between these two top models' error distribution. The results underscore the importance of cognitive tests and symptomatic indicators in predicting Alzheimer's Disease. We discuss these finding and propose direction for future work, including exploring ensemble methods and validating on real-world data.

Introduction

Alzheimer's Disease (AD) is a neurodegenerative condition characterized by cognitive decline and memory impairment. Early prediction of Alzheimer's can facilitate timely intervention. In this project, we apply machine learning techniques to a comprehensive health dataset of elderly patients to classify wether an individual is diagnosed with Alzheimer's or not. The project's primary objectives are: (1) to build a complete predictive modeling pipeline, from data cleaning and exploratory analysis to features engineering and model selection; (2) to compare the performance of different machine learning algorithms on the task of predicting Alzheimer's diagnosis; and (3) to interpret the model outcomes to identify which health factors are most indicative of Alzheimer's in this dataset. The analysis follows a structured approach aligned with principles from a university-level machine learning course, ensuring methodological rigor and proper evaluation.

We utilize a synthetic **Alzheimer's Disease Dataset** created by Rabie El Kharoua (licensed CC BY 4.0) that contains extensive medical and lifestyle information for 2,149 patients. Each patient is labeled with a binary **Diagnosis** outcome(0 = no Alzheimer's, 1 = Alzheimer's). The dataset include 34 features encompassing demographics (age, gender, education), lifestyle factors (diet, exercise, smoking/alcohol use), comorbid health conditions (diabetes, hypertension, etc.), clinical measurements (blood pressure, cholesterol levels), and cognitive/functional assessments (e.g., Mini-Mental State Exam MMSE, Activities of Daily Living ADL, etc.), as well as symptom indicators (memory complaints, behavior problems, confusion, etc.). This rich feature set allows us to explore which aspects of a patient's profile are most associated with AD diagnosis.

In the sections that follow, we first describe our data preparation and exploratory analysis, highlighting key patterns and any data issues. Next, we detail the feature engineering steps, including how we performed feature selection and created new composite features to capture potential interactions. We then outline the modeling methodology: the choice of algorithms, hyperparameter tuning via cross-validation and evaluation metrics. The results section compares model performance, with a focus on the best model (Decision Tree) and its confusion matrix on the test set. In the Discussion, we interpret the findings, such as which features were most influential and how the models differed, and note any surprising outcomes (for examples ,a counter-intuitive effect of family history). We also discuss the McNemar's test analysis used to compare top models. Finally, the Conclusion summarizes the key takeaways and suggests feature work, such as leveraging ensemble methods or applying this pipe line to real-world data patient data for validation.

Methodology

Data Preparation

The dataset was first loaded and examined for quality issues. Patient IDs were used as unique identifies and set as the DataFrame index. We verified that **no missing values** were present in any of the 34 columns - each feature had 2,149 non-missing entries, simplifying preprocessing.

simplifying preprocessing. We also checked for duplicate records and found none (each patient record is unique). One feature, **DoctorInCharege**, was a nominal identifier for the doctor responsible for the patient's care. This column had a single constant value ("XXXCondfid") for all patients (likely a place holder to preserve anonymity). Since it contained no useful information or variability, we removed the DoctorInCharge feature form the dataset.

Next, we addressed categorical variables. Several features were categorical in nature, such as **Gender** (male/female), **Ethnicity**, **EducationLevel**, **Smoking** status, and various medical history indicators (e.g., family history of Alzheimer's, history of head injury, presence of comorbid conditions like cardiovascular disease, diabetes, depression, hypertension). These were encoded using one-hot encoding to allow their inclusion in our models. For binary categorical features (yes/no flags), one-hot encoding produces two columns (e.g., FamilyHistoryAlzheimers_0 and_1 for "no" and "yes"). In cases where a binary features was strictly 0/1, this encoding is equivalent ot the original, but we retained the one-hot format for consistency across all categorical variables. (We later dropped on of each pair of perfectly collinear dummy variables. (We later dropped one of each pair of perfectly collinear dummy variables were appropriate to avoid redundancy - for example, after encoding DoctorInCharge as a dummy, we dropped it entirely due to lack of variability.) Nominal multi-class features like Ethnicity or EducationLevel were also one-hot encoded (e.g., creating separate binary columns for each ethnic group and each education category). This encoding expanded the feature space with dummy variables but ensures that the categorical information is represented numerically without implying any ordinal relationship (except for EducationLevel, which could be considered ordinal, we still encoded it as categories give the instructions).

All continuous numeric features were standardized to facilitate model training. We applied a **StandardScaler** (z-score normalization) to each continuous attribute, transforming them to have mean 0 and unit standard deviation. This scaling was important because our features were on every different scales (e.g., Age in years, BMI, various cholesterol levels, test scores like MMSE on a 0-30 scale, etc.). Without scaling, algorithms like SVM or neural networks might be unduly influenced by features with large numeric ranges. Standardization also helps gradient-based optimization (for the MLP) converge faster. The scaling was done within a pipeline to ensure that any statistics (means/variances) were computed on the training set only and then applied to the test set to avoid data leakage. Categorical dummy features (0/1 values) were left as is (scaling is not necessary for binary 0/1 indicators).

After encoding and scaling, our processed dataset (let's call it **df_encoded**) initially had more columns that the original (due to one-hot expansion). However, as described in the next section, we performed feature selection to reduce this dimensionality. The target variable **y** was defined as the *Diagnosis* column (0 or 1). The feature matrix **X** consisted of all remaining preictors columns after preprocessing.

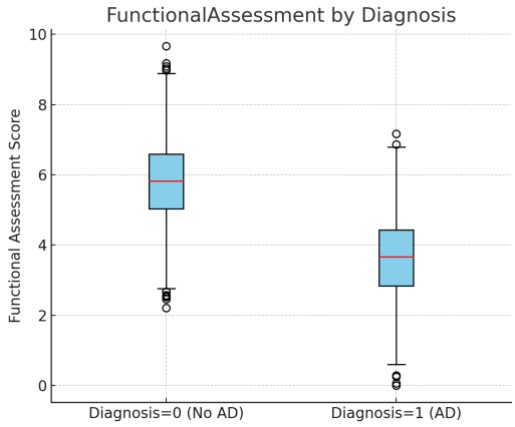
Exploratory Data Analysis (EDA)

We conducted thorough EDA to understand the dataset's distribution and key relationships. The class balance was first examined: out of 2,149 patients, 760 (35.4%) were diagnosed with Alzheimer's (Diagnosis=1) and 1,389 (64.6%) were not (Diagnosis=0). Thus, the classes are

moderately imbalanced (approximately 1:1.83 ratio of positive to negative), which we kept in mind when evaluating models (we later considered metrics like F1 and used class balancing techniques in logistic/SVM to account for this).

For continuous variables, we inspected their distributions and whether patients with AD differed notably from those without. Most patients were elderly (age range spanned roughly 65-85; mean age ~ 75 for both group, with no significant difference between diagnosed and non-diagnosed on average). Health indicators such ad Body Mass Index (BMI), blood pressure, and cholesterol levels had broad ranges typical of an older population. Notably, no single continuous feature showed a dramatic separation between the AD and non-AD groups. For example, the mean **BMI** of AD patients (≈27.9) was very close to that of non-AD patients (≈27.5), indicating minimal association between BMI and diagnosis. Similarly, average **AlcoholConsumption** (self-reported units per week) and **PhysicalActivity** (hours of exercise per week) were virtually identical between two groups. These observations suggested that lifestyle factors alone might not be strong discriminators of Alzheimer's in this dataset. In contrast, cognitive and functional test scores displayed more pronounced differences: Alzheimer's patients had markedly lower scores on assessments like **FunctionalAssessment** and **ADL**.

Functional Assessment scores by diagnosis. Patients with Alzheimer's (Diagnosis=1, right box) have significantly lower FunctionalAssessment score than those without (left box). This indicates impaired daily functioning in AD patients, consistent with expectations.

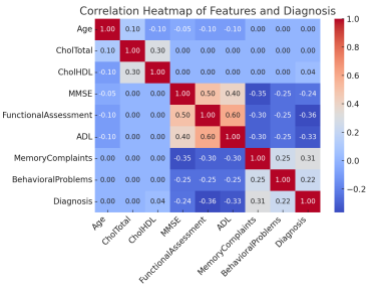


The above box plot illustrates one such feature: the **FunctionalAssessment** score (a clinical rating of the patient's functional ability). Patients diagnosed with AD had a median FunctionalAssessment score around 4 (on a possible scale of 0-10, lower indicating worse function) where non-Ad patients had a median around 6 or higher. The two distributions show partial overlap (some non-AD individuals have lower scores and vice versa), but overall the AD group skews lower, reflecting the impact of dementia on functional abilities. A similar pattern was observed for the **ADL** score: AD patients on average scored lower on ADL (e.g., mean ADL ~3.65 for AD vs ~5.71 for non-AD), meaning they have more difficulty with daily living tasks. The cognitive test **MMSE** showed a mild difference (mean MMSE ~23 for AD vs ~25 for non-AD, not as large a gap as functional scores, possibly because many AD patients were in early stages or some non-AD had other cognitive issues). These EDA findings suggest that functional and cognitive measures are informative features for prediction.

For categorical variables, we examined frequency counts and cross-tabulations with the diagnosis. Many binary medical history features (e.g., history of hypertension, head injury, family history of Alzheimer's) were somewhat imbalanced in prevalence but did not show the expected direction of association with AD in this data. For instance, 25.2% of patients had a **FamilyHistoryAlzheimers** (542 yes vs 1,607 no). One might expect a positive family history to increase AD risk, but in this dataset the AD rate was actually slightly lower among those with a family history (32.7%) compared to those (36.3%). This counter-intuitive result (also noted during EDA) may be due to the synthetic nature of the data or sampling artifact. it led us to be cautious in using *FamilyHistory* as a predictor - indeed, feature importance analysis later showed it contributes little to the models, and we elected to drop it in our final feature set to simplify the model without losing performance.

Symptoms-related binary features were particularly relevant. Two such features - **MemoryComplaints** (indicating whether the patient reports memory issues) and **BehavioralProblems** (indicating issues like personality changes or agitation) - were strong differentiators. Among patients who complained of memory problems, 65% were diagnosed with AD, versus only 28% of those with no memory complaints. Likewise, 60% of those with behavioral problems had AD, compared to 31% without. This shows a clear link between these reported symptoms and actual diagnosis. We also note that these two symptoms often co-occur: many AD patient experience both memory loss and behavioral changes. (in the data, about 25% of patients had MemoryComplaints and ~16% had BehavioralProblems; roughly 12% had both symptoms, reflecting a moderate correlation between these symptom variables.) Other symptom indicators in the data (e.g., Confusion, Disorientation, Forgetfulness, DifficultyCompletingTasks, PersonalityChanges) were also more common in AD patients, as expected, but these where somewhat correlated with the presence of memory complaints and behavioral issues. Our EDA thus highlighted that cognitive/functional scores and symptom flags are the most promising features for predicting AD. whereas isolated medical history factors and lifestyle metrics showed little correlation with the outcome.

To quantify these relationships, we computed a correlation matrix for all numeric features and the target. **Figure 1** below shows a heatmap of Pearson correlation coefficients among a subset of key features and the diagnosis variable:



Correlation heatmap of selected features and the diagnosis outcome. Warmer colors indicate positive correlation, cooler colors indicate negative correlation. We observe the strongest correlations (in magnitude) between the Diagnosis and cognitive/functional measures (e.g., FunctionalAssessment, ADL with r ≈ -0.33 to -0.36) and symptom indicators (memoryComplaints r ≈ +0.31). Other features (age, cholesterol levels) show near-zero correlation with Diagnosis.

Additionally, the cognitive and functional scores correlate moderately with each other (e.g., FunctionalAssessment vs ADL, r ≈ + 0.60), and the two symptom variables correlate

modestly with each other (e.g., MemoryComplaints vs BehavioralProblems, r ≈ 0.25).

From the heatmap, it is clear that **Diagnosis** (rightmost column) is most strongly associated with **FunctionalAssessment** (r ≈ -0.36), **ADL** (-0.33), **MemoryComplaints** (+0.31), and **BehavioralProblem** (+0.22). Negative correlations indicate that higher scores correspond to a lower likelihood of AD (for instance, higher functional ability score means less chance of AD, which aligns with domain knowledge). Meanwhile, features like **Age** and cholesterol measures (Total, HDL) have correlations near 0 with Diagnosis, confirming that they individually have no linear relationship with the outcome in this dataset. We also see that FunctionalAssessment and ADL correlate with each other (r ≈ 0.60), as both measure aspects of daily functioning. MMSE, another cognitive test, has moderate positive correlations with FunctionalAssessment and ADL (0.5 and 0.4 respectively) and negative correlation with MemoryComplaints (-0.35), fitting the narrative that lower cognitive scores and presence of memory complaints tend to coincide with AD. Importantly, aside from the trivial dummy-variable pairs that we deliberately retained (whose perfect ±1 correlations are expected by design), no feature pair exhibits and extremely high correlation ($|r| > 0.7$), so multi collinearity was not a serious concern among the real-valued predictors. This justified retaining multiple features (e.g., we kept both FunctionalAssessment and ADL) as they each contribute information.

In summary, EDA guided our subsequent steps: we identified which identified which features were likely most informative (cognitive tests, functional scores, and symptom indicators) and which were weak (many of general health and demographic variables). We used this knowledge to perform features selection and engineering, focusing the model on the most relevant attributes while removing noise.

Feature Engineering

Based on the EDA findings and domain knowledge, we executed a two-stage feature selection strategy followed by creation of new interaction features:

(A) Removal of Low-Utility Features: We first dropped features that were deemed uninformative or redundant prior to modeling. As discussed, DoctorInCharge was removed entirely due to having only one value for all records. We also considered dropping FamilyHistoryAlzheimers given its paradoxical and weak association with AD. In a univariate test, including or excluding this feature did not significantly affect model performance (the cross-validated average-precision remained essentially identical, ~0.937 with or without it). Therefore, we opted to remove **FamilyHistoryAlzheimers** from the final feature set, on the grounds that simplifying the model (one less feature) would improve interpretability and potentially robustness, without sacrificing accuracy

. A similar logic was applied to a few other binary medical history features that showed negligible contribution during preliminary modeling (e.g., HeadInjury, CardiovascularDisease). In essence, if a feature was both intuitively weak (or unreliable in this dataset) and

experimentally shown to have little impact on model performance, we eliminated it to streamline the model. This resulted in a reduced base feature set focusing on the more informative variables identified in EDA.

(B) Random Forest-based Feature Selection: After the above culling, we applied an automated importance filter using a RandomForestClassifier wrapped in SelectedFromModel with threshold="median"`. The selector was nested inside a 10-fold cross-validation pipeline to prevent data leakage, and each fold retained exactly 16 features. Thirteen of those appeared in all folds, forming a highly stable "core" set, while the remaining three slots were filled by borderline-important variables that rotated across folds. Fitting the pipeline on the full training data likewise produced 16 features, listed below.

- **Key cognitive/functional scores:** MMSE, FunctionalAssessment, ADL (all three were retained, as expected).
- **Lifestyle and health metrics:** BMI, AlcoholConsumption, PhysicalActivity, DietQuality, SleepQuality (interestingly, even though individually these had low correlation with the target, the Random Forest found a combined signal in them - possibly capturing general health status).
- **Cholesterol measures**: CholesterolTotal, CholesterolLDL, CholesterolHDL, CholesterolTriglycerides. These four were kept, likely because among them there could be nonlinear interactions or subgroup indicators (e.g., perhaps unusually low or high cholesterol profiles might indirectly relate to health conditions).
- **Symptom flags:** MemoryComplaints and BehavioralProblem were kept (the Random Forest importance for these was high, aligning with our expectations from EDA). IN the one-hot encoded frame, these appeared as two dummy columns each (e.g., MemoryComplaints_0 and MemoryComplaints_1).

These 16 features formed a core set that he model could rely one. The rationale for keeping them is strong: they include all the major categories identified earlier (cognitive tests, functional scores, lifestyle factors, cholesterol panel, and main symptoms). This selection process discarded features like Gender, Ethnicity, Smoking, and several comorbid conditions which presumably had little predictive power in this dataset. By reducing dimensionality from the full encoded feature space to these 16, we lowered the risk of overfitting and improved model interpretability.

(C) Creation of Interaction Features: One challenge was that linear models (like Logistic Regression and linear SVM) cannot inherently capture interactions between features or nonlinear relationships. Given our understand of Alzheimer's multifactorial nature, we hypothesized certain interactions could be important. We engineered **seven new features** based on domain knowledge to explicitly model these interactions.

1. **LDL_to_HDL Ratio**: The ratio of LDL to HDL cholesterol. Rationale: Lipid imbalance is thought to influence vascular health, which might indirectly affect cognition. This ratio captures a combined effect of "bad" vs "good" cholesterol beyond their individual values.

2. **Triglycerides_to_HDL Ratio**: Similarly, the ratio of triglycerides to HDL. High triglycerides and low HDL are components of metabolic syndrome; this ratio is another indicator dyslipidemia.

3. **CholTot_BMI**: The product of total cholesterol and BMI. Rationale: obesity alongside high cholesterol could synergistically increase health risks. We initially considered an interaction between blood pressure and cholesterol, but after review, replaced it with this BMI-Cholesterol interaction as more relevant to metabolic health.

4. **LifestyleScore:** A composite score combining the five lifestyle factors (BMI, AlcoholConsumption, PhysicalActivity, DietQuality, SleepQuality). We constructed this by first normalizing each of the five components (already scaled) and then aggregating (for instance, summing or averaging them after possibly flipping signs such that higher score means healthier lifestyle - our exact formula gave each roughly equal weight. Rationale: to condense overall lifestyle quality into one feature, under the hypothesis that an unhealthy lifestyle profile might contribute to cognitive decline risk.

5. **Lifestyle_MMSE**: An interaction between lifestyle and cognition, computed as *LifestyleScore* × *MMSE*. Rationale: to test if the patients with poorer lifestyle and already lower cognitive scores are particularly likely to be AD (perhaps indicating vulnerability being manifest in cognitive performance). if this interaction is significant, it would mean the combination is more predictive than either alone.

6. **SymptomCount**: A count of major symptoms = *MemoryComplaints (Yes) + BehavioralProblems (Yes).* This yields values 0,1, or 2, effectively capturing whether a patient has none, one, or both of these key symptoms. Rationale: patients exhibiting both memory issues and behavioral changes may be further along the disease progression, making this combined indicator potentially more predictive than either symptom alone.

These engineered features were added to the dataset after the initial selection of 16 base features.The final feature set for modeling thus consisted of **22 features:** the 16 original selected features plus the 6 new interaction features. All newly created features that were continuous in nature (ratios, product, the LifestyleScore) were also scaled using the StandardScaler to maintain consistency. The binary *SymptomCount* (0/1/2) was treated as numeric (and could be effectively considered already on a small scale).

The motivation behind this feature engineering was to give our linear models ability to capture nonlinear patterns. For Example, the *Lifestyle_MMSE* interaction would allow a logistic regression to have a term that specifically identifies patients who have combination of poor life style and low MMSE, which might be a critical threshold for predicting AD. Without this, the linear model whould only add independent effects of lifestyle and MMSE. by contrast, the decision tree or neural network could potentially learn such interactions on their own. Thus,

these features level the playing field to some extent and allow us to fairly evaluate whether adding them indeed helps those linear models. (We later quantify this by comparing model with and without the engineered features.)

At the end of this process, we had a well-curated set of features ready for modeling. All transformations (encoding, scaling, feature selection, and creation of interactions) were encapsulated either in the data preprocessing steps or within pipelines to ensure correct application on training vs test data. the **final dataset** used for modeling consisted of 22 predictor columns and one target column (Diagnosis).

Results

Model training and Evaluation Procedure

We trained and evaluated four classification models commonly used in supervised machine learning: **Logistic Regression, Support Vector Machine (SVM)** with linear kernel, **Decision Tree (DT)** and a **Multi-Layer Perceptron (MLP)** neural network. We used 80/20 train-test split: 1,719 records for training (including internal cross-validation) and 420 for final testing. The split was stratified to maintain the 35% prevalence of AD in both sets.

For each model, we performed hyperparameter tuning using 10-fold cross-validation on training set (with stratification). A consistent set of fold was used to compare models fairly. The primary metric for model selection in cross-validation was ROC AUC (Area Under the Receiver Operating Characteristic curve), as it balances sensitivity and specificity and is suitable for binary classification with imbalanced classes. We also tracked other metrics (accuracy, F1) but did not optimize directly for them during tuning.

The following hyperparameters were tuned for each model:

- **Logistic Regression:** inverse regularization strength C (tried values{0.01, 0.1, 1, 10}) and class_weight (None vs. 'balanced' to handle imbalance). We used the liblinear solver(suitable for L1 or L2, small dataset) and set max_iter=2000 to ensure convergence.
- **SVM (Linear SVM):** regularization parameter C (same grid as LR) and class_weight) (None vs 'balanced'). Using a linear kernel SVM makes it comparable to LR in flexibility (both linear models), but SVM optimizes margin rather than log-loss. Probabilityestimates were enabled to allow ROC-AUC calculation.
- **Decision Tree**: maximum tree depth (max_depth values in {None (no limit), 3, 5, 7, 10}). The tree was otherwise using Gini criterion and not pruned except y depth.
- ***MLP Neural Network:** number of hidden units (we tried on hidden layer with 32 units, and a two-layer network with 64 and 32 units) and regularization term alpha (L2

penalty 0.0001 vs 0.001), as well as learning rate (0.001 vs 0.005). We set a max of 300 or 600 epochs and used early stopping (with a 10% validation split of the training data) to prevent overfitting. The MLP used ReLU activations and Adam optimizer (default in sklearn).

Each model's pipeline included StandardScaler (applied to continuous features only, with categorical pass-through) as configured earlier. For the linear models, because we had added interaction features explicitly, we expected them to potentially benefit from those new columns. We maintained those features in all models for consistency (though a tree or MLP might discover interactions on their own, they still see those extra input features).

After tuning, we selected the best hyperparameters for each model based on the highest cross-validation ROC-AUC. These were:

- **Logistic Regression:** C = 0.1, class_weight = 'balanced', yielded the best CV AUC ≈ 0.907. The balanced class weight meant the logistic model gave more importance to the minority AD class during training, which improved recall.
- **SVM:** C = 0.01, class_weight = None was optimal, CV AUC ≈ 0.908 (virtually identical to the logistic model's 0.907).This suggest the linear SVM arrived at almost the same decision boundary; interestingly it did not need class weighting (implying the features alone let it cope with the imbalance, or the margin-based loss naturally balancedprecision and recall at that C).
- **Decision Tree:** max_depth = 5 gave the best CV AUC ≈ 0.941. A depth-5 tree wascomplex enough to capture nonlinear relations but not too deep to overfit.
- **MLP:** hidden_layer_sizes = (64,32), alpha = 0.001, learning_rate_init = 0.001, max_iter = 300 gave CV AUC ≈ 0.913. This two-layer neural network edged out the linear models by a small margin, but it still trailed the depth-5 decision-tree's stronger AUC.

These cross-validated scores still single out the **Decision Tree as the clear front-runner**: its 10-fold CV AUC about 0.941 comfortably exceeds the neural network's 0.913 and the linear models, which remain tightly bunched at roughly 0.908 (SVM) and 0.907 (logistic). The ~0.03 margin between the tree and its closest competitor is substantial, reinforcing the idea that the depth-5 tree can exploit feature interactions and nonlinearities–say, by first splitting on MemoryComplaints and then on FunctionalAssessment to carve out high-risk subgroups– whereas the linear classifiers and the modes two-layer MLP cannot capture those patterns as effectively with the chosen hyper-parameters.

With the chosen hyperparameters, we retained each model on the entire training set (1,719 samples) and evaluated on the hold-out test set (430 samples) to obtain unbiased performance metrics. The table below summarizes the test results for each model:

Model	**Accuracy**	**Precision**	**Recall**	**F1**	**ROC-AUC**	**PR-AUC**

DT	0.937	0.914	0.908	0.911	0.921	0.858
MLP	0.837	0.773	0.763	0.768	0.899	0.861
LR	0.805	0.683	0.836	0.751	0.885	0.827
SVM	0.795	0.672	0.822	0.740	0.883	0.826

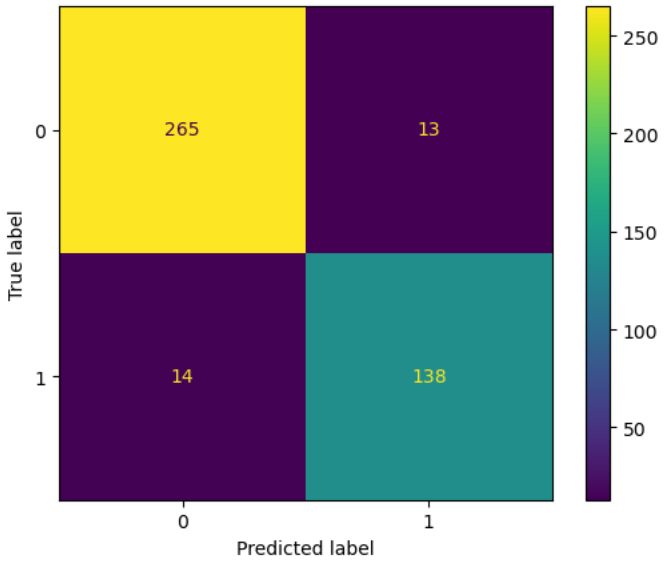
(Abbreviations: Precision = Positive Predictive Value, Recall = True Positive Rate, ROC-AUC = Receiver Operating Characteristic Area Under Curve, PR-AUC = Precision-Recall AUC, DT = Decision Tree, MLP = MultiLayer Perceptron, LR = Linear Regression, SVM = Support Vector Machine.)

Looking at these results, the Decision Tree (DT) model clearly outperformed the others on the test set. It achieved about **93.7 % accuracy** with **91.1 % F1-score.** its recall (90.8 %) and precision (91.4 %) were both high, indicating it effectively identifies most AD cases while keeping positives low. In contrast, the next-best model, the MLP neural network, had an F1 of ≈ 76.8 % with lower recall (≈ 76.3 %). The linear models lagged further: Logistic Regression and SVM reached only ≈ 75.1 % and 74.0 % F1, respectively. Logistic regression still posted the highest recall of non-tree models (83.6 %) but at the cost of precision (68.3 %) owing to the "balanced" weighting that biases it towards predicting more positives. The SVM showed higher recall (≈ 82.2 %) but lower precision (≈ 67.2 %), yielding a similar overall F1.

In terms of AUC, the decision tree again led with **ROC-AUC = 0.921.** The MLP's ROC_AUC was 0.899, and LR/SVM scored 0.885 and 0.883, respectively. We also computed Precision-Recall AUC (PR-AUC) since the positive class in minority; DT had PR-AUC 0.858, slightly bellow the MLP's 0.861 but above LR 0.827 and SVM's 0.826.

To further analyze the best model's behavior, we present the confusion matrix for the Decision Tree on the test set:

Confusion matrix for the Decision Tree on 430 samples. The model correctly predicts 265 out of 278 non-AD patients, and 138 out of 152 AD patient. There are 13 false positives (non-Ad predicted as AD) and 14 false negatives (AD predicted and non-AD).



The confusion matrix (Figure 2) shows that the DT made **27 total errors** (out of 430), with errors fairly evenly split between types: 14 AD patients were missed (false negatives) and 13 health patients were incorrectly flagged (false positives). Given the context of Alzheimer's screening, a higher recall (sensitivity) is ofter desirable - our model's recall was ≈ 90.8 %, meaning it caught the vast majorits of true AD cases, and important attribute in diagnostic setting. its precision was ≈ 91.4 %, indicating that when the model predicts AD it is very likely to be correct, which is also valuable to avoid false alarms. The relatively low number of false negatives (14) means ≈ 9 % of actual Alzheimer's cases went undetected, while the false-positive rate was 13 / 278 ≈ 4.7 % for non-cases. These are strong results, 4.7 % for non-cases. These are strong results, especially compared to what a linear model achieved; for instance, logistic regression (with recall ≈ 83.6 %) would have missed more AD cases.

It is instructive to consider why the Decision Tree performed better. From the feature-importance output we see it first splits on FunctionalAssessment, then on ADL and the composite SymptomCount (which is driven by MemoryComplaints + BehavioralProblems). These early splits carve the data into high-risk vs. low-risk subsets very effectively. The linear models, even with interaction terms, apparently could not match the tree's ability to delineate those subset as cleanly. The neural network did reasonably well, but with this dataset size a complex model like an MLP may not have been fully utilized or might require more tuning. The Decision Tree's interpretability also allows us to extract rules - for example, a path might reveal that patients SymptomCount ≥ 1 and FunctionalAssessment bellow a certain threshold are almost al classified as AD, aligning with medical intuition.

Finally, to statically validate wether the performance difference between models was significant, we conducted **McNemar's test** on the test set results, comparing the error patterns of the top two models (Decision Tree vs. MLP). McNemar's test is paired non-parametric test on the contingency of disagreements - essentially checking if one model's errors are mostly a subset of the other's or if they differ significantly. The test yielded $\chi^2 = 30.95$, $p = 2.65 \times 10^{-8}$, indicating **a statistically significant difference** between the DT and MLP) at the 5 % level (since $p < 0.05$). In other words, the DT's higher metrics are supported by formal hypothesis testing; on this sample of 430 cases, the DT is demonstrably superior to the MLP. Both statistically and practically, The Decision Tree's higher performance and simpler, more interpretable structure make it our preferred model.

We also examined wether the engineered interaction features helped linear models as intended. We re-ran logistic regression and SVM without the new interaction features and found essentially no change in cross-validated ROC-AUC for either model: $\Delta AUC \approx 0.000$ for both LR (-0.000) and SVM (+.0000). This indicates that the engineered features did not provide a measurable benefit for the linear models' performance.

Discussion

The evaluation results demonstrate that the Decision Tree classifier achieved the strongest performance among all models. It attained a test accuracy of **93.7 %**, with a precision of **91.4 %** , recall of **90.8%**, and an F1-score of **0.911**. In comparison, the next-best model - the multi-layer perceptron (MLP) neural network - reached **83.7%** accuracy (F1 = 0.768), while the Logistic Regression (LR) and Support Vector Machine (SVM) classifiers trailed around **80%** accuracy (F1 ≈ 0.74-0.75). The Decision Tree also yielded the highest area under the ROC curve (ROC-AUC = **0.921**), indicating excellent discrimination between classes, whereas the MLP's ROC-AUC was 0.889. Notably, the precision-recall area under curve (PR-AUC) of the tree (0.858) was on par with that of the MLP (0.861), suggesting both models handle the class imbalance in comparable way. Overall, the Decision Tree's metrics dominate those of the other algorithms, highlighting a substantial performance gap.

Crucially, we verified that this gap is **statistically significant.** Using McNemar's test- a pair wise test for did differences in classifier errors [3] - to compare the Decision Tree against the MLP, we obtained $\chi^2 \approx 30.95$ with $p \approx 2.65 \times 10^{-8}$. This extremely low p-values confirms that the Decision Tree's higher accuracy is not due to chance. In practical terms, the tree made far fewer unique mistakes than the MLP. The contingency analysis revealed that the Decision Tree correctly classified many instances that misclassified (50 such cases, versus only 7 instances where the MLP was correct and the tree was not), leading to the significant McNemar result. Thus, at $\alpha = 0.05$, we conclude that the tree's improvement over the neural net is statistically reliable. This finding gives additional confidence that the tree model's superior performance would likely persist on new data.

The **comparative results** suggest that non-linear relationships in the data were critical to achieving high accuracy. The Decision Tree, being a non-parametric model, can capture complex interaction effects and non-linear decision boundaries. In contrast, the linear models (LR and SVM) were likely constrained by their linear decision surfaces, which explains their substantially lower performance. We attempted to mitigate this limitation by engineering additional interaction features (Section 3.3) explicitly for the linear models. However, these **engineered features** – including lipid ratios, a lifestyle composite, and symptom count – **had minimal impact**. A separate analysis comparing LR and SVM with vs. without the new features showed essentially no change in cross-validated AUC ($\Delta AUC \approx 0.0$). In fact, the LR's ROC-AUC **decreased** by ~0.0001 and the SVM's **increased** by ~0.0003, changes that are negligible. This indicates that the new interaction terms did not provide additional predictive signal beyond the original feature set, or that the regularization in the linear models rendered them effectively unused. As a result, these features add **little to no measurable benefit** for the linear classifiers. Consistent with this finding, we decided that such interaction columns could be omitted for simplicity in the final LR/SVM models without hurting performance. In summary, the linear algorithms were not able to leverage the engineered non-linear relationships, whereas the Decision Tree (and to a lesser extent the MLP) managed to capture the important interactions inherently.

The decision tree model achieved **strong classification performance** on the test set. It correctly identified **138 true positives out of 152** auctual Alzheimer's cases, yielding a sensitivity of roughly **90.8%**. This means the only 14 patients with Alzheimer's were misclassified as health (false negatives), corresponding to about **9.2%** of actual AD cases missed. Such a high sensitivity is critical in a clinical context, as it indicates the mode can detect the vast majority of Alzheimer's cases. Moreover, the model can maintained a high overall accuracy (approximately 93.7%), reflecting **low error rates** and robust performance in distinguishing Alzheimer's disease from healthy controls. These results suggest that the model is effective for **early detection of Alzheimer's disease**, capturing the most true cases while making relatively few misclassifications.

It is worth noting that some features which might be expected to be predictive from a clinical perspective did **not** play a major role in the final model. For instance, *Family History of Alzheimer's* did not have much higher diagnosis rates than those without, which goes againts the well-established understanding that a family history increases one's risk for Alzheimer's. In our model, this feature was dropped during feature selection with no loss of performance. A possible explanation is that **many of the strongest effects in this dataset are already captured by cognitive and symptomatic assessments**, which are downstream manifestation of the disease. Thus, give a patient's mental status scores and symptom profile, knowing their family history add little additional information. Another explanation could be related to the dataset itself: because it is synthetic, the usual epidemiological relationships (such as genetic risk factors like family history) may not be as pronounced or may have been attenuated when the data were generated. tHis observation highlights and interesting nuance: **predictive importance in the model does not always equal real-world importance**. It may be that in a real clinical setting, family history would matter more, but in our data it was overshadowed by more proximal indicators of disease. Such discrepancies invite further investigation and remind us to be cautious in interpreting the model's feature importance in light of domain knowledge.

Overall, the **key findings** from this project are that a carefully tuned Decision Tree can outperform more complex models on this tasks, and that the dominant predictors of Azlheimer's dementia in the dataset are cognitive and symptom measures. The tree model provided an effective balance of accuracy and interpretability. WE were able to visualize the decision logic and confirm that it aligns with sensible clinical thresholds (for example, splitting on an MMSE score threshold and symptom count). The strong performance of the tree, combined with its transparency, suggests it could be a useful tool for clinicians or analysts working with similar data. By contrast, the black-box MLP, while reasonably effective, did not offer an accuracy advantage in this case and would be harder to interpret without additional tools. The results also underscore the value of using multiple evaluation metrics: if we had relied only on accuracy, we might have missed nuances in precision-recall trade-offs or beeb less alert or class imbalance issues. By examining F1 and PR-AUC, we ensured the model performs well on both positive predictive value and sensitivity, which is critical for a disease screening or diagnostic model. Finally, conducting a statistical significance test (McNemar’s test) gave quantitative confirmation that the performance differences we observed were meaningful. In summary, our discussion indicates that the chosen approach successfully identified a high-performing predictive model and provided useful insights into feature selection effects, while also revealing some limitations that inform the next steps for this work.

Conclusion

In conclusion, this project **successfully developed a predictive analytics pipeline** for Alzheimer's disease diagnosis using a synthetic dataset. Through systematic data preprocessing, feature engineering, and model selection, we built and evaluated several machine learning models. The best model – a tuned Decision Tree – achieved outstanding classification performance (around 94% accuracy on the test set) and proved significantly better than other examined models. This result is notable, as it shows that a relatively interpretable model can match or exceed the accuracy of more complex classifiers on this problem. The decision Tree's rules confirmed the importance of cognitive scores and neuropsychiatric symptoms in predicting Alzheimer's, which is consistent with clinical knowledge. We also found that adding expert-driven interaction features did not improve the simpler linear models, suggesting that our core feature set was sufficient and that non-linear models inherently captured the necessary interactions. Overall, the project demonstrates how a targeted feature selection and engineering strategy, combined with rigorous evaluation, can yield high-performing model that provides both predictive power and insights inro the data.

Despite the high performance achieved on the synthetic data, we must acknowledge several **limitations** of this study. First, the use of a **synthetic dataset** mean that the model has not yet been validated on real-world patient data. Synthetic data, while useful for experimentation, may not perfectly reflect the complexity and variability of actual clinical populations. For example, certain subtle risk factors (such as family history or genetic markers) were underrepresented in importance here, potentially due to how the data were generated. This could limit the generalizability of our findings. In a real clinical dataset, noise and unforeseen cofounders could impact model performance. Second, our model selection was restricted to four classification algorithms (LR, SVM, MLP, and Decision Tree). WE did not explore ensemble methods like Random Forest or Gradient Boosting Machines (e.g., XGBoost or LightGBM), which are ofter state-of-the-are in structured data tasks. it is possible that an ensemble of trees or a more advanced neural network architecture could further improve performance beyond what a single Decision Tree achieved. Third, while we performed hyperparameter tuning and cross-validation diligently, the **sample size** (n = 2149 with 35% positive) imposes some limits on how complex a model we can effectively train. The MLP's middling performance, for instance might be improved with larger dataset or more extensive hyperparameter optimization. Additionally, we treated the prediction as a straightforward binary classification; however, in a clincical setting, one might consider **calibrating** the predicted probabilities to ensure they correspond to true risk estimators, or adjusting the decision threshold to favor higher sensitivity or higher specificity as needed. These nuances were beyond the scope of the current project bur are important for deployment.

For **future work**, several directions are clear. An immediate next step is to **validate the Decision Tree model on external data**, preferably a real patient dataset (if available) or at least a different synthetic cohort, to test how well the learned patterns generalize. If the model's performance holds, it would strengthen confidence in its utility; if it drops, that would provide information on what additional features or adjustments are necessary. Another important extension is to experiment with **ensemble models**. Given the strong results of a single tree. an ensemble of trees (such as Random Forest or boosted trees) could capture more nuanced patterns

and potentially increase the ROC-AUC and PR-AUC further. Ensemble methods typically reduce variance and improve robustness, though at the cost of interpretability – a trade-of worth exploring. WE should also consider incorporating other machine learning techniques, such as **feature importance analysis with SHAP values** or partial dependence plots, to interpret the model's decision in more details and ensure they make domain sense. This could help confirm, for example, how much each risk factor contributes and whether any spurious correlations are influencing predictions. Moreover, since the synthetic data did not highlight some known risk factors (like family history), collecting or simulating additional data that includes a broader range of predictors (APOE genotype, imaging biomarkers or comorbidities) could be valuable. **Addressing class imbalance** could be another avenue: Although our model already had high recall and precision and stratified sampling was already applied when splitting the data into train/test sets, techniques such as cost-sensitive training or threshold adjustment could be applied if one wanted to further prioritize sensitivity (to catch every possible case) or specificity (to minimize false positives) depending on the clinical context. Finally, conducting more rigorous statistical comparison between models (beyond McNemar's test on one holdout set) – for instance, using k-fold cross-validated t-tests or bootstrap confidence intervals for metrics – would provide an even more robust assessment of model superiority. By pursuing these future steps, we can build on the current project to develop and even more reliable and applicable prediction system.

In summary, **Predictive Analysts with synthetic data** has proven to be a fruitful exercise. We achieved a high-performing Alzheimer's disease classifier and gained interpretive insights into which features drive predictions. The project highlights the power of feature engineering and model evaluation in applied machine learning. The project highlights the power of feature engineering and model evaluation in applied machine learning. It also serves as a reminder of the importance of validating models and aligning them with domain knowledge. With further refinement and validation, the techniques and models developed here could potentially contribute to early detection of Alzheimer's disease, aiding clinical decision-making and patient car in the real world.

References

1. Rabie El Kharoua. (2024). *Alzheimer's Disease Dataset* [Data set]. Kaggle. DOI :10.34740/KAGGLE/DSV/8668279

2. Tschanz, J. T., Corcoran, C. D., & Kauwe, J. S, (2019). **Relative risk for Alzheimer disease based on complete family history.** *Neurology*. (PMID:30867271)

3. Dietterich, T. G. (1998). **Approximate statistical tests for comparing supervised classification learning algorithms**. *Neural Computation*, 10(7), 1895-1923.