Master: Intelligent Processing Systems
Module: Data Mining
Dr. Anouar Riadsolh
February 24,2025

Alzmine "Analysis of Alzheimer's risk factors and symptoms to uncover hidden trends"

BOUDALI SALMA

Mohammed V University, Rabat

E-mail: Salma_boudali@um5.ac.ma

Abstract: Dementia, especially Alzheimer's disease (AD), represents a major global health challenge due to its progressive nature and impact on cognitive functions. Early detection of AD is essential for timely intervention and better patient outcomes. This study explores the application of data mining techniques, specifically Apriori, KNN, and K-means clustering, to detect early signs of AD from medical datasets. Apriori was used to identify recurring patterns and associations between symptoms and risk factors, while KNN was used for classification tasks to predict the likelihood of AD onset based on patient data. K-means clustering was used to group patients with similar characteristics, allowing for the identification of distinct subgroups at different stages of cognitive decline. Our results demonstrate that the combination of these algorithms improves diagnostic accuracy and provides valuable insights into the underlying patterns associated with the progression of AD. This research highlights the potential for integrating machine learning approaches into clinical settings for more effective and personalized management of dementia.

Keywords: Alzheimer's disease; early detection; Apriori algorithm; K-Nearest Neighbors (KNN); K-means clustering; data mining;

1 Introduction

Alzheimer's disease (AD) is a chronic neurodegenerative illness that is the most common form of dementia, affecting millions of individuals worldwide and exerting a tremendous burden on the healthcare system and families. AD is characterized by cognitive decline, memory loss, and behavioral changes and is frequently insidious in onset, with subtle early manifestations easily missed or misdiagnosed. Early diagnosis of Alzheimer's disease is critical for timely intervention as it allows therapeutic interventions to be applied that can slow the disease process, improve quality of life, and provide patients and caregivers with additional time to plan ahead. Diagnosis of AD in its early stages remains challenging due to the complexity of its clinical presentation and a lack of absolute biomarkers.

With the advances in data mining and machine learning technologies over the past few years, new opportunities have been developed for improving the accuracy and effectiveness of Alzheimer's diagnosis. Such computerized processes enable large

sets of data to be processed and patterns and relationships that are imperceptible through traditional diagnostic tools to be identified. Of these approaches, the Apriori algorithm has been applied with success to determine frequent itemsets and associations among risk factors and symptoms, which have provided valuable insights into the etiology of the disease. In addition, classification algorithms such as K-Nearest Neighbors (KNN) provide useful tools for prediction of Alzheimer's onset based on patient-specific data. Moreover, clustering techniques like K-means clustering offer the amenity of partitioning patient populations into clusters that facilitate the description of subtypes of cognitive loss and individualized treatment strategies.

It is the objective of this research to explore integrating the aforementioned data mining techniques—Apriori, KNN, and K-means clustering—aspects in detecting Alzheimer's disease. By exploiting the merits of both strategies, we intend to set up a unified framework for early diagnosis and better understanding of the disease's course. The findings of this study can provide access to more accurate diagnostic tests and contribute to the creation of targeted interventions that will subsequently advance the treatment and care for individuals affected by Alzheimer's disease.

2 Alzheimer's Disease Dataset

2.1 About dataset

The dataset contains extensive health information for 2,149 patients, each uniquely identified with IDs ranging from 4751 to 6900. The dataset includes demographic details, lifestyle factors, medical history, clinical measurements, cognitive and functional assessments, symptoms, and a diagnosis of Alzheimer's Disease. The data is ideal for researchers and data scientists looking to explore factors associated with Alzheimer's, develop predictive models, and conduct statistical analyses.

2.2 List of Columns

- **PatientID**: A unique identifier assigned to each patient (4751 to 6900).
- **Age**: The age of the patients ranges from 60 to 90 years.
- **Gender**: Gender of the patients, where 0 represents Male and 1 represents Female.
- **Ethnicity**: The ethnicity of the patients.
- **EducationLevel**: The education level of the patients.
- **BMI**: Body Mass Index of the patients.
- **Smoking**: Smoking status.
- **AlcoholConsumption**: Weekly alcohol consumption in units, ranging from 0 to 20.
- **PhysicalActivity**: Weekly physical activity in hours, ranging from 0 to 10.
- **DietQuality**: Diet quality score, ranging from 0 to 10.

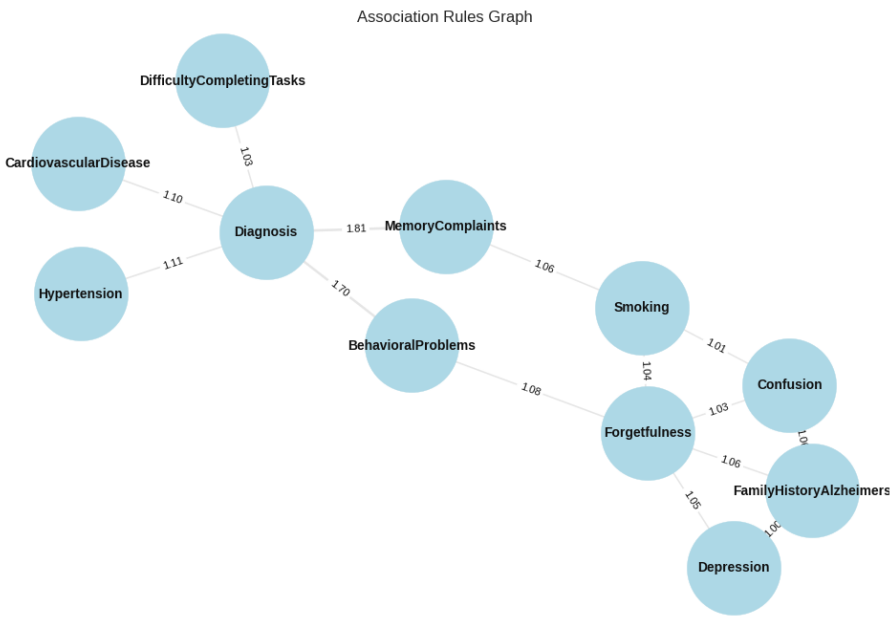
- **SleepQuality**: Sleep quality score, ranging from 4 to 10.
- **FamilyHistoryAlzheimers**: Family history of Alzheimer's Disease.
- **CardiovascularDisease**: Presence of cardiovascular disease.
- **Diabetes**: Presence of diabetes.
- **Depression**: Presence of depression.
- **HeadInjury**: History of head injury.
- **Hypertension**: Presence of hypertension.
- **SystolicBP**: Systolic blood pressure, ranging from 90 to 180 mmHg.
- **DiastolicBP**: Diastolic blood pressure, ranging from 60 to 120 mmHg.
- **CholesterolTotal**: Total cholesterol levels, ranging from 150 to 300 mg/dL.
- **CholesterolLDL**: Low-density lipoprotein cholesterol levels, ranging from 50 to 200 mg/dL.
- **CholesterolHDL**: High-density lipoprotein cholesterol levels, ranging from 20 to 100 mg/dL.
- **CholesterolTriglycerides**: Triglycerides levels, ranging from 50 to 400 mg/dL.
- **MMSE**: Mini-Mental State Examination score, ranging from 0 to 30. Lower scores indicate cognitive impairment.
- **FunctionalAssessment**: Functional assessment score, Lower scores indicate greater impairment.
- **MemoryComplaints**: Presence of memory complaints.
- **BehavioralProblems**: Presence of behavioral problems.
- **ADL**: Activities of Daily Living score, Lower scores indicate greater impairment.
- **Confusion**: Presence of confusion.
- **Disorientation**: Presence of disorientation.
- **PersonalityChanges**: Presence of personality changes.
- **DifficultyCompletingTasks**: Presence of difficulty completing tasks.
- **Forgetfulness**: Presence of forgetfulness.
- **Diagnosis**: Diagnosis status for Alzheimer's Disease, where 0 indicates No and 1 indicates Yes.
- **DoctorInCharge**: This column contains confidential information about the doctor in charge, with "XXXConfid" as the value for all patients.

3 Association Rules & Apriori Algorithm

We employed the Apriori algorithm and association rule mining to find out frequent patterns and co-relationships among risk factors and Alzheimer's disease symptoms. Data included binary variables representing varied health conditions, behavior symptoms, and diagnosis findings. By using the Apriori algorithm for minimum support as 0.05, we established frequent itemsets, i.e., a set of factors occurring simultaneously with high frequency in data. These itemsets were found to have strong associations, for instance, smoking and Alzheimer's family history, depression and memory complaints, and confusion with forgetfulness.

We then generated association rules using the confidence measure with a threshold of 0.1 and evaluated the strength of the rules using the lift measure. Rules with a lift greater than 1 and confidence greater than 0.1 were retained for further analysis. These rules established robust associations between specific antecedents (e.g., smoking, family history of Alzheimer's, depression) and consequents (e.g., forgetfulness, Alzheimer's diagnosis). For instance, the presence of "FamilyHistoryAlzheimers" was strongly correlated with an increased likelihood of an Alzheimer's diagnosis, highlighting the contribution of genetic susceptibility to disease prediction.

We represented the complex relationships uncovered by the association rules by constructing a directed graph using the NetworkX library. In the following graph, single symptoms or risk factors are represented by nodes and edges are the level of association (provided by the lift value). The network generated provided a clear representation of the relationships between the risk factors and the symptoms and revealed clusters of highly connected variables. As observed, "Diagnosis" was one such central node with a large number of strong connections with a number of antecedents such as memory complaint, confusion, and behavioral difficulty.



4 K-Nearest Neighbors (KNN) algorithm

4.1 Methodology

To develop a predictive model for Alzheimer's disease, we employed the K-Nearest Neighbors (KNN) algorithm, a widely used supervised learning classification technique. We preprocessed the data by splitting it into a training set (80%) and a test set (20%) using stratified sampling to have class-balanced representation. Feature scaling was performed using the StandardScaler to normalize feature values such that all variables received equal weightage during distance measurements. This preprocessing step is critical in the case of KNN because the algorithm relies on distance metrics to predict.

Hyperparameter tuning was conducted using GridSearchCV with 5-fold cross-validation to optimize the performance of the KNN model. The parameter grid explored various configurations, including:

- **n_neighbors**: Number of neighbors considered (3, 5, 7, 9, 11, 13, 15).
- **weights**: Weighting scheme for predictions (uniform or distance).
- **metric**: Distance metric used for similarity calculations (euclidean, manhattan, minkowski).
- **p**: Power parameter for Minkowski distance (1 for Manhattan, 2 for Euclidean).
- **algorithm**: Algorithm used to compute nearest neighbors (auto, ball_tree, kd_tree, brute).
- **leaf_size**: Leaf size passed to tree-based algorithms (10, 20, 30, 40).

The goal of this exhaustive search was to identify the combination of hyperparameters that maximized classification accuracy while minimizing overfitting.

4.2 Results

After evaluating multiple combinations of hyperparameters, the optimal configuration was identified as follows:

- **Algorithm**: auto
- **Leaf Size**: 10
- **Metric**: manhattan
- **Number of Neighbors (n_neighbors)**: 15
- **Power Parameter (p)**: 1 (corresponding to Manhattan distance)
- **Weights**: distance (weighted by inverse distance)

This configuration achieved a cross-validated accuracy of 76.44% on the training data, indicating strong generalization capabilities across different subsets of the dataset. When evaluated on the independent test dataset, the model demonstrated a test accuracy of 73.26% , showcasing its ability to accurately predict Alzheimer's disease diagnosis based on the provided features.

5 K-means clustring

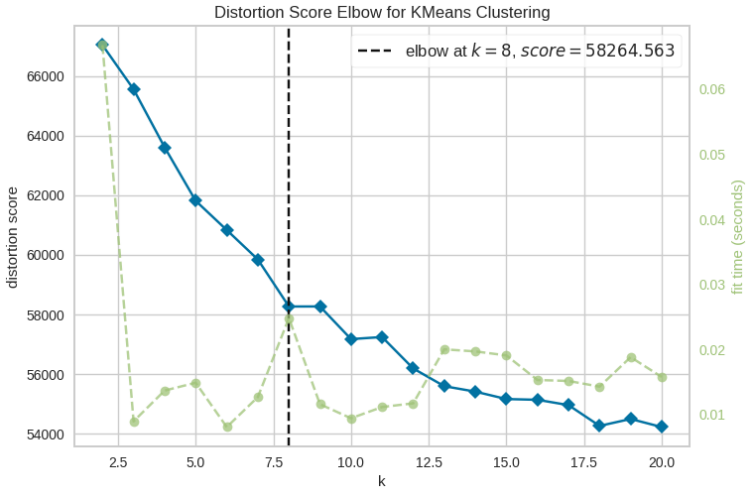
5.1 Methodology

To identify distinct subpopulations of patients based on their symptom profiles and health characteristics, we employed the K-Means clustering algorithm, a type of unsupervised learning technique that is usually used to group data points into groups with similar characteristics. The dataset was processed by removing the target variable (Diagnosis) to ensure that the clustering only relies on the input features. Feature scaling was performed using the StandardScaler to ensure the feature values are normalized such that every feature equally contributes in distance calculations with the K-Means algorithm.

Determination of the ideal cluster (k) is critical for effective clustering. We employed the use of the Elbow Method, where the within-cluster sum of squares (WCSS) or the distortion score is computed for varying numbers of clusters to determine the optimal number of clusters. The goal is to find the "elbow point," at which additional clusters do not produce additional significant drops in the distortion score, and thus diminishing returns in cluster quality. To run this analysis, we utilized the KElbowVisualizer in the Yellowbrick library and evaluated k values 1 through 20.

5.2 Results

The K-Means clustering analysis, guided by the Elbow Method, revealed that the optimal number of clusters for our data was k=8, as indicated by the distinct "elbow point" in the distortion score plot below. This segmentation revealed eight unique subpopulations of patients with varying levels of cognitive impairment and risk factors, which provided us with informative insights into the heterogeneity of Alzheimer's disease.



6 Conclusion

This study explored the application of data mining algorithms Apriori, K-Nearest Neighbors (KNN), and K-Means clustering to discover early predictors of Alzheimer's disease and determine patterns in medical data. The Apriori algorithm was successful in discovering frequent itemsets and association rules, which indicated strong associations between risk factors and symptoms. Strong associations were observed between the co-occurrence of smoking and Alzheimer's family history, depression and memory complaints, and confusion with forgetfulness. These results highlight the importance of studying numerous interconnected factors in Alzheimer's diagnosis and set the stage for the description of the disease etiology.

The K-Nearest Neighbors (KNN) classifier worked adequately when classifying Alzheimer's disease diagnosis with 73.26% test accuracy after hyperparameter tuning using GridSearchCV. The best configuration used 15 neighbors, Manhattan distance, and inversely weighted predictions, showing the capability of KNN to predict patients accurately from their medical and behavioral features. While the achieved accuracy is encouraging, more can be done through feature engineering, exploring more advanced algorithms, or utilizing larger datasets.

Finally, K-Means clustering revealed the heterogeneity of Alzheimer's disease by dividing patients into seven distinct subpopulations based on symptom profiles and health characteristics. The Elbow Method plot validated k=8 as the optimal number of clusters, which identified clusters from initial symptoms like memory complaints to severe symptoms like disorientation and personality changes. These findings emphasize the need for personalized treatment strategies based on certain patient groups.