Deep Dive into Behavioral Risk Factors that Might Lead to Cognitive Decline

Xinya Li, Zhiquan Shen, Keni Xue, Guanchen Xiao, Lavanya Shankar

Abstract

This study focuses on Healthcare Analytics with a specific emphasis on understanding the relationship between subjective cognitive decline and health-related risk behaviors in the United States. Cognitive decline can serve as an early indicator of dementia, making it crucial to investigate its determinants. Leveraging the Behavioral Risk Factor Surveillance System (BRFSS), a nationally representative survey conducted annually by the Centers for Disease Control and Prevention (CDC), we aim to quantify how daily habits influence subjective cognitive decline. The 2022 BRFSS dataset, gathered through cross-sectional phone surveys covering all 50 states, the District of Columbia, and three US territories, provides a rich source of information for this analysis. By examining factors such as exercise frequency, tobacco use, alcohol consumption, and other health-related behaviors, we seek to enhance our understanding of cognitive decline's drivers and implications for public health. This research endeavors to contribute valuable insights that can inform targeted interventions aimed at mitigating cognitive decline and improving overall public health outcomes.

1 Background

Dementia is a chronic or progressive condition characterized by the deterioration of cognitive function beyond what might be expected from the usual consequences of biological aging. It affects memory, thinking, orientation, comprehension, calculation, learning capacity, language, and judgment. Subjective Cognitive Decline (SCD) is increasingly recognized as a potentially early sign of dementia, particularly Alzheimer's disease. SCD refers to a self-experienced persistent decline in cognitive capacity in comparison with an individual's usual performance, yet without noticeable deficits on standard cognitive tests. Identifying SCD is crucial as it may be an indicator of early pathological changes in the brain, offering a vital window for intervention before more severe impairment or dementia onset.

Behavioral Risk Factor Surveillance System (BRFSS) is an annual, nationally representative survey in the United States. The data is a cross-sectional phone survey (including landline telephone and cellular telephone) initiated by the CDC (Centers for Disease Control and Prevention) and conducted by state health departments. The 2022 BFRSS collects data in all 50 states, the District of Columbia and three US territories (Guam, Puerto Rico, and the US Virgin Islands). It includes an optional cognitive decline module, with self-reported scale pertaining to challenges in daily life due to memory loss and growing confusion over the past twelve months.

The integration of machine and deep learning techniques in medical research, particularly in the study of diseases like dementia, has shown significant promise in uncovering complex patterns and relationships that are not easily detectable through traditional statistical methods. These technologies can process large volumes of data and identify intricate structures in the data by learning features and tasks directly from the data. In the context of SCD, machine and deep learning models can analyze the vast and varied data collected through health surveys like BRFSS to explore associations between SCD and various health-related risk behaviors. By employing models such as random forest, (multi-nomial) logistic regression, neural networks and boosting methods, we hypothesize that these health-related characteristics will predict the existence and severity level of SCD.

2 Literature Review

Latent Class Analysis The study by (Snead et al., 2022) leverages data from the BRFSS to explore latent classes of SCD among U.S. adults aged 45 and over. Using latent class analysis (LCA), the study identifies three distinct subgroups—Mild, Moderate, and Severe SCD—based on patterns in self-reported cognitive difficulties affecting daily functioning. Building on the their work, our study extends the use of LCA to not only replicate findings on latent classes of cognitive decline but also to explore how health-related behaviors predict membership in these latent classes. By integrating machine learning techniques, we aim to develop predictive models that more accurately identify individuals at risk of progressing from mild to more severe cognitive impairment.

Machine Learning on BRFSS Data In (H.N. et al., 2023), the authors applied machine learning algorithms to diabetes prediction using the BRFSS dataset. The findings suggest that while individual algorithms like SVM show superior performance in specific scenarios, a comparative approach considering various algorithms and datasets is crucial for enhancing prediction accuracy. In (Akkaya et al., 2022), the authors explored machine learning application to predict diabetes using the BRFSS dataset. The study compares the performance of three machine learning models: K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Logistic Regression. The research indicates that among the machine learning models tested, SVM outperforms others in terms of accuracy. In (Pochini et al., 2015), the authors focused on identifying asthma using machine learning techniques applied to the 2013 BRFSS dataset. Their study explored a broader set of methods, including decision trees, logistic regression, neural networks, and gradient boosting.

Research on Cognitive Decline The term subjective cognitive decline (SCD) was introduced in 2014 with an increasing awareness of brain health and Alzheimer’s disease among the general population(Jessen et al., 2020). Since identifying modifiable risk factors and promoting early intervention can help reduce or prevent the development of more severe impairments or dementia.In (Deary et al., 2009), study shows factors such as age, smoking, unhealthy lifestyle, social engagement, and genetics are key influences on cognitive decline. However, research (Williams and Kemper, 2010) indicates that modifiable factors such as not smoking, leading a healthy lifestyle, actively engaging in social activities, exercising regularly, and maintaining a proper diet can effectively intervene in cognitive decline.

3 Data Preparation and Analysis

3.1 Data

The 2022 BRFSS data comprised a total of 445,132 records and 328 features. A subset of this dataset was particularly relevant: over 60,000 adults aged 45 and older in the U.S. who had participated in the optional cognitive decline module of the survey. After filtering out respondents who were either not eligible for or did not complete the cognitive decline module, the dataset was narrowed down to 63,883 records. Within this refined dataset, 57,198 respondents reported no subjective cognitive decline, whereas 6,685 indicated experiencing such decline. Given the difference in the number of respondents reporting cognitive decline versus those who did not, class imbalance was considered and addressed during the modeling process to ensure accurate analysis and prediction. Of those reporting cognitive decline, the distribution between genders showed 3,013 males and 3,672 females.

3.2 Individual Feature Analysis

Here are the individual feature analysis results for key predictors of cognitive decline in individuals aged 45 and older.

Social Isolation

Social Isolation *(Survey Question: How often did you feel socially isolated from others?)* was analyzed through responses ranging from "Always" isolated to "Never". Figure 1. showed an increase in moderate and severe cognitive decline in those who were more frequently socially isolated. This suggested that greater isolation correlated with higher cognitive decline.

Difficulty Concentrating or Remembering (Removed)

The "Difficulty Concentrating or Remembering" feature evaluated responses to whether a physical, mental, or emotional condition caused serious difficulty in concentrating, remembering, or making decisions.Figure 2 indicated that those experiencing such difficulties had a higher incidence of moderate to severe cognitive decline. However, we realized that this feature was another way of asking if people had cognitive decline or not. To prevent redundancy, this feature was excluded from further analysis.

Difficulty Dressing or Bathing

Difficulty in dressing or bathing *(Survey Question: Did you have difficulty dressing or bathing?)* significantly correlated with cognitive decline, as shown

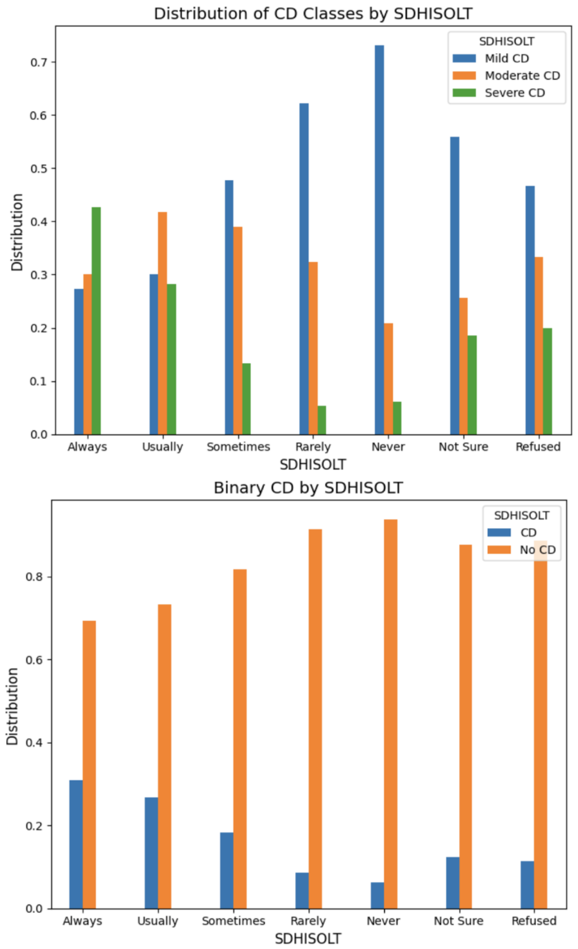

Figure 1: Distribution of cognitive decline classes by the level of social isolation.

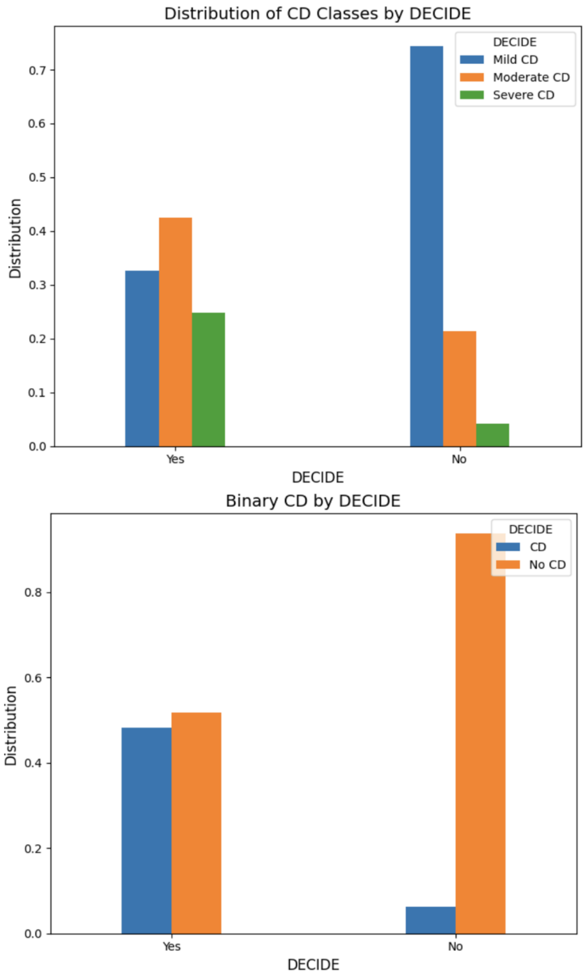

Figure 2: Distribution of cognitive decline classes by the level of difficulty concentrating or remembering.

in Figure 3. Those reporting "Yes" to difficulty in these activities showed a higher prevalence of moderate to severe cognitive decline.

Marijuana Usage in Past 30 Days

Marijuana Usage *(Survey Question: During the past 30 days, on how many days did you use marijuana or cannabis?)* was another representative feature. Figure 4 showed that the distribution of marijuana usage was different between cognitive decline classes, with the more severe cognitive decline group having greater variance in the distribution. This suggested a possible link between marijuana use and cognitive health.

Number of Years Smoked Cigarettes

The distribution of years smoked cigarettes in Figure 5 presented a higher density of all cognitive decline classes among those with more years of smoking. This highlighted smoking duration as a risk factor for cognitive decline. However, it was also counter-intuitive in the binary distribution that those who did not have cognitive decline seemed to have a higher mean year of smoking than those who had cognitive decline.

3.3 Feature Selection and Processing

Feature selection was carried out using filter models. Each individual feature was explored, with its quality evaluated by three crisp mathematical criteria, including Cramer's V correlation, ANOVA p-value, and the combined consideration of class-based entropy and Gini impurity gain.

3.3.1 Correlation and ANOVA Based

For categorical variables, the association between two nominal variables can be quantified using Cramer's V. This statistic measures the strength of their association on a scale from 0 to 1. Cramer's V is calculated by taking the square root of the chi-squared statistic divided by the sample size and normalized by the lesser of the number of rows or columns minus one. The formula for Cramer's V

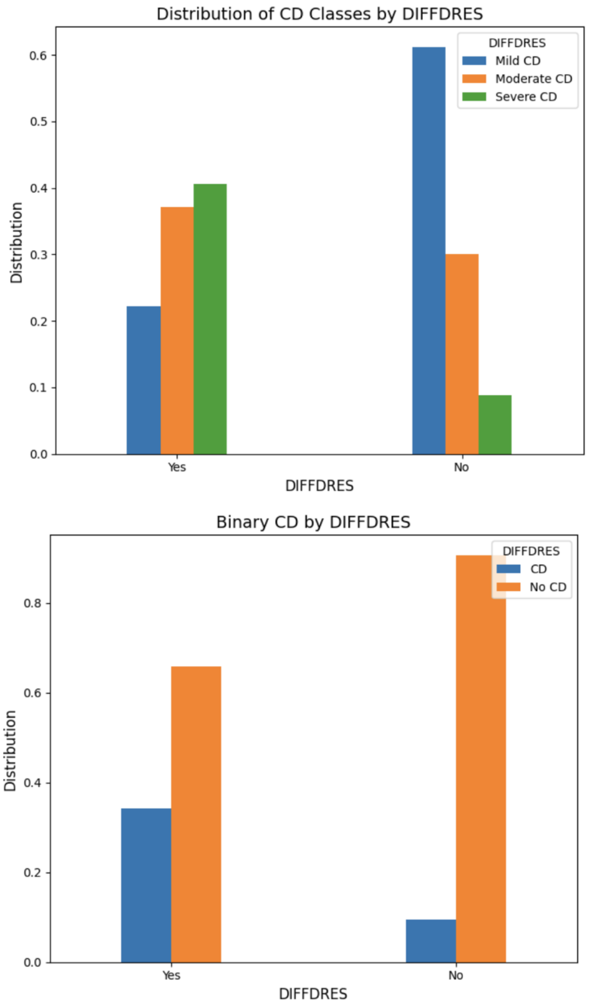

Figure 3: Distribution of cognitive decline by difficulty in dressing or bathing.

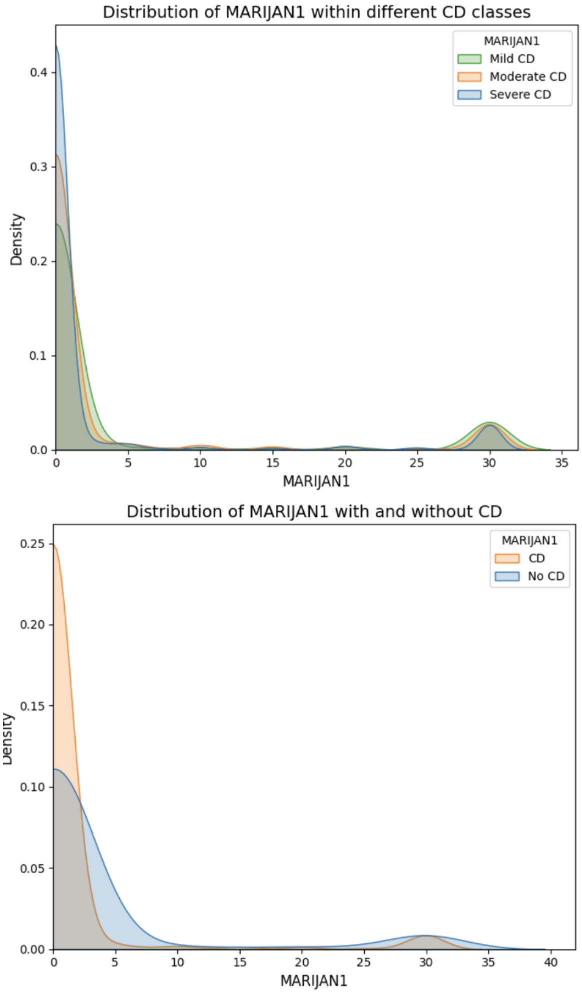

Figure 4: Distribution of cognitive decline classes among marijuana users.

is given by:

$$V = \sqrt{\frac{\phi^2}{\min(k-1, r-1)}} = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}}$$

where ϕ is the phi coefficient, χ^2 is derived from Pearson's chi-squared test, n is the total number of observations, k is the number of columns, and r is the number of rows.

For numerical variables, the Analysis of Variance (ANOVA) is utilized to analyze the differences among group means in a sample. The ANOVA tests the hypothesis that the means of various groups are equal, under the null hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$. The alternative hypothesis is that at least one group mean is different. The ANOVA summary table is structured as in Table 1.

For our first round of feature selection, categorical variables were retained based on a Cramér's V coefficient greater than 0.1, indicating a significant association in multicategory classification. Numerical variables were selected based on an ANOVA test with a p-value less than 0.1, suggesting statistically significant differences across groups. Further refinement of the feature set was conducted by examining each remaining variable individually, considering factors such as the number of missing values and redundancy among features. This meticulous approach effectively reduced the feature pool from over 300 to approximately 60.

3.3.2 Class-Based Entropy

We compute class-based entropy to measure the information gain resulting from a specific feature value. The entropy E of feature is the weighted average over r unique values of the feature:

$$E = \sum_{i=1}^{r} n_i E(v_i)/n,$$

Source	Sum Sq	df	Mean Sq	F Value
Between Groups	$SSB = \sum n_j(\bar{X}_j - \bar{X})^2$	$df_1 = k - 1$	$MSB = \frac{SSB}{k-1}$	$f = \frac{MSB}{MSE}$
Error	$SSE = \sum(X_i - \bar{X}_j)^2$	$df_2 = N - k$	$MSE = \frac{SSE}{N-k}$	
Total	$SST = SSB + SSE$	$df_3 = N - 1$		

Table 1: ANOVA Table for Feature Analysis

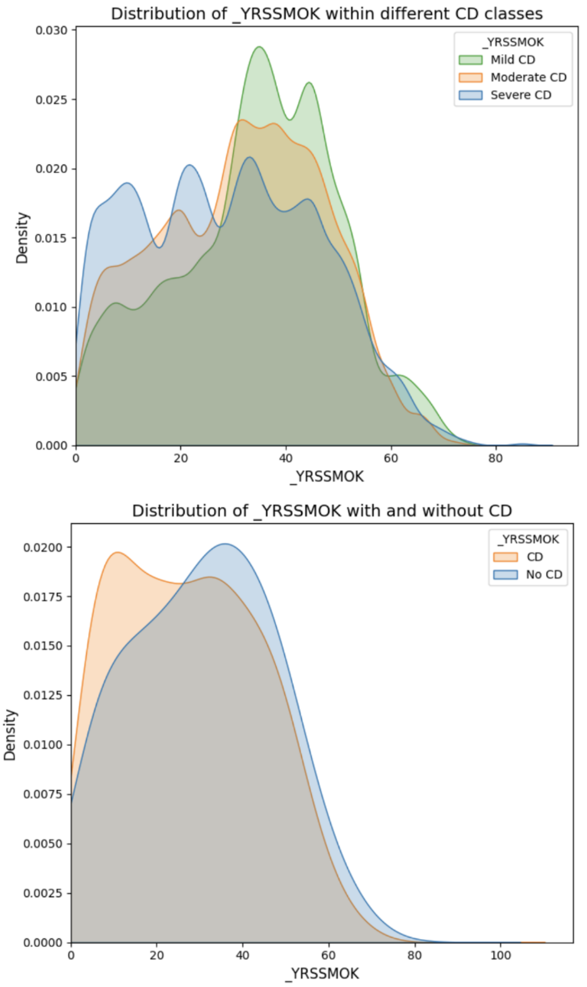

Figure 5: Distribution of cognitive decline by the number of years smoked.

where

$$E(v_i) = -\sum_{j=1}^{k} p_j \log_2(p_j)$$

and p_j is the fraction of data points belonging to the class j for feature v_i. Here, k is the number of target classes. The class-based entropy value lies between 0 and $\log_2 k$. Higher value of entropy imply higher mixing of different classes, so we sort entropy values in ascending order and aim to features with smaller entropy values.The results of top 10 class-based entropy is shown in Table 2.

3.3.3 Gini Impurity Gain

In decision tree algorithms such as random forest classifier, determining the best feature and value to split on at each node is crucial. One of the metrics used for this purpose is the Gini impurity gain, which measures the degree of impurity of features in a dataset.

$$I(D) = 1 - \sum_{i=1}^{C}(p_i)^2,$$

where C is the number of classes and p_i is the the probability of an instance belonging to class i in the dataset. Gini impurity gain is computed by comparing the impurity of the parent node before the split with the impurity of the child nodes after the split. The formula for Gini impurity gain is:

$$\begin{aligned} IG(D_p, x_i) =& I(D_p) \\ & - \frac{N_{\text{left}}}{N_p} I(D_{\text{left}}) \\ & - \frac{N_{\text{right}}}{N_p} I(D_{\text{right}}), \end{aligned} \quad (1)$$

where x_i is the feature to split, N_p is the number of samples in the parent node, N_{left}, N_{right} is the number of samples in the left/right child node. D_p is the subset of the parent node, D_{left}, D_{right} is the subset of right/left child node. By computing Gini impurity gain for each feature in our dataset, we gain a better understanding of the features that are most effective in splitting the data, thus helping us build a more accurate model. A feature with higher impurity gain implies higher effectiveness. The results of top 20 Gini impurity gain is shown in Table 3. This process improves the accuracy of our classification model by selecting the most informative features for splitting the dataset.

3.3.4 Data Processing

3.4 Latent Class Analysis

Latent class analysis (LCA) was used to determine unobserved subgroups of subjective cognitive decline (SCD) based on item response patterns, effectively extending our prediction problem from

Features	Entropy	Features	Entropy
Very satisfied	0.5429	PHYSHLTH	0.5762
DRNKWK2	0.5615	DIFFALON	0.5986
INCOMG1	0.5651	Health-Good or Better	0.5991
POORHLTH	0.5743	Health-Fair or Poor	0.5993
MENTHLTH	0.5761	mental not good:0day	0.6031

Table 2: Class-Based Entropy

Features	Gini Impurity Gain	Features	Gini Impurity Gain
POORHLTH	0.0083	bad mental 0 day	0.0055
MENTHLTH	0.0081	bad health 14+day	0.0051
SDHSTRE1	0.0067	EMPLOY1	0.0048
PHYSHLTH	0.0066	SDHISOLT	0.0048
DIFFALON	0.0064	bad health 0 day	0.0041
Health-Good or Better	0.0063	DIFFDRES	0.0039
Health-Fair or Poor	0.0063	EMTSUPRT	0.0038
bad mental 14+day	0.0061	SDHFOOD1	0.0031
ADDEPEV3	0.0057	INCOME3	0.0029
DIFFWALK	0.0055	Dissatisfied	0.0029

Table 3: Gini Impurity Gain

binary to multi-class. Four of the six SCD related questions are used, including 3 5-category ordinal and 1 binary. The four ordinal-response (*Always, Usually, Sometimes, Rarely, and Never*) questions were: *1) During the past 12 months, as a result of confusion or memory loss, how often have you given up day-to-day household activities or chores you used to do, such as cooking, cleaning, taking medications, driving, or paying bills?; 2) As a result of confusion or memory loss, how often do you need assistance with these day-to-day activities?; 3) During the past 12 months, how often has confusion or memory loss interfered with your ability to work, volunteer, or engage in social activities outside the home?* The binary-response question was *Have you or anyone else discussed your confusion or memory loss with a health care professional?*.

3.4.1 Method Review

Model Specification LCA models the probability of each observed response pattern by assuming the population is divided into a finite number of latent classes. Each class is characterized by a specific response pattern probability.

Parameter Estimation The model parameters (class probabilities and item-response probabilities) are estimated using maximum likelihood estimation. Let $\mathbf{Y}$ be the matrix of observed categorical responses for N individuals on J indicator variables, where each element y_{ij} indicates the response of individual i on indicator variable j. We want to maximize with respect to $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$

$$L(\boldsymbol{\pi}, \boldsymbol{\theta}) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \times P(\mathbf{Y}_i | \boldsymbol{\theta}_k),$$

where $\boldsymbol{\pi}$ is the vector of class membership probabilities with π_k being the probability of belonging to latent class k and $\boldsymbol{\theta}_k$ represents the vector of parameters for latent class k.

3.4.2 Selection of Best Fitting Latent Class

Various models with different numbers of classes are compared using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Lower values of these criteria generally indicate a better model fit, balanced against model simplicity. Four latent mixture models were run iteratively increasing the number of latent classes from one to tour. Table 4 provides the fit statistics. Based on AIC/BIC, we select the three latent class model. The distinction between latent subgroups is intuitive between the three classes, we can hence interpret the three classes as Mild, Moderate, and Severe.

# Classes	Log-Likelihood	AIC	BIC	Entropy	LMR-LRT
1	-115777	231594	231775	1.82	
2	-51626	103335	103706	0.81	$p < 0.001$
3	-48822	97769	98331	0.77	$p < 0.001$
4	-51626	103419	10417	0.81	$p = 1.000$

Table 4: LCA fit statistics

3.4.3 Analysis

The selected model separates response patterns into three discrete latent classes. Conditional and unconditional probability estimates from the 3-class model appear in Table 5. To proceed with the multi-class SCD prediction, labels are assigned to samples by picking the sub-class with the highest posterior probability.

3.5 Clustering Analysis

3.5.1 Spectral Clustering Overview

Spectral clustering treats the data points as nodes in a graph and aims to partition the graph into cluster by analyzing the eigenvalues and eigenvectors of the graph Laplacian matrix derived from the data. The graph Laplacian matrix L is defined as

$$L := D - A,$$

where A is the adjacency matrix with entries representing the absence or presence of an edge between the nodes, and D is a diagonal degree matrix, with the diagonal representing the number of edges attached to that vertex. By computing the eigenvector corresponding to the second-smallest eigenvalue of L, we essentially find a partition of the graph with few edges across the partitions.

3.5.2 Similarity Measure

Given our data predominantly comprises of categorical features, our similarity measure is defined as the simple matching measure with inverse occurrence frequency:

$$S(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{d} S(x_i, y_i),$$

where

$$S(x_i, y_i) = \{ \begin{array}{ll} 1/p_k(x_i)^2 & \text{if } x_i = y_i \\ 0 & \text{otherwise} \end{array}$$

Here $p_k(x_i)$ is the proportion of records in which kth attribute takes on the value of x in the dataset.

3.5.3 Results and Analysis

Spectral clustering was performed multiple times on random subsets of observations. Each time a graph was constructed to represent the data, with nodes representing data points and edges representing high similarity. By observing the similarity matrix, we select a threshold of 300 to define the presence of an edge between the nodes.

The values of the eigenvector (corresponding to the second-smallest eigenvalue) are shown in Figure 6 and the distributions of classes within clusters are shown in Figure 7. The results from the spectral clustering analysis provide revelations into the data structure. The ordered entries of the eigenvector show a clear bi-partition, indicating that the data can be effectively split into two distinct clusters, where the split is defined by the entries crossing the $y = -0.005$ line.

However, when examining the distribution of the target variable within these two clusters, the analysis reveals that both clusters have a remarkably similar composition of the target classes (binary and multi-class). This observation suggests that while the spectral clustering algorithm has successfully identified structural partitions based on data similarity, these partitions do not correspond to significant differences in the target variable. This similarity in class distribution across the two clusters imply that the features used for clustering, though effective in grouping similar data points, do not align well with the variations in the target variable. Clustering analysis underscores the necessity of further predictive modelling on the target variable.

4 Predictive Modelling and Results

4.1 Binary Prediction

4.1.1 Imbalanced Dataset

Our dataset has very imbalanced target classes Figure 8. In dealing with imbalanced datasets, downsampling and Synthetic Minority Over-sampling Technique (SMOTE) are two techniques we used. Imbalanced datasets occur when one class is sig-

	Subjective Cognitive Decline		
	Severe	**Moderate**	**Mild**
	13.5%	31.0%	55.5%
During the past 12 months . . .			
As a result of confusion or memory loss, how often have you given up day-to-day household activities or chores you used to do, such as cooking, cleaning, taking medications, driving, or paying bills?			
Always	27.6%	0.0%	0.0%
Usually	26.4%	3.4%	14.4%
Sometimes	39.3%	44.7%	4.7%
Rarely	3.1%	33.2%	12.1%
Never	3.7%	0.0%	81.5%
As a result of confusion or memory loss, how often do you need assistance with these day-to-day activities?			
Always	25.6%	1.4%	0.4%
Usually	23.9%	1.4%	0.2%
Sometimes	40.5%	34.8%	3.1%
Rarely	3.5%	45.2%	9.8%
Never	6.5%	18.1%	86.6%
How often has confusion or memory loss interfered with your ability to work, volunteer, or engage in social activities outside the home?			
Always	35.6%	2.8%	0.7%
Usually	25.8%	6.1%	0.2%
Sometimes	28.7%	38.0%	4.2%
Rarely	2.6%	36.4%	13.5%
Never	7.3%	16.7%	81.4%
Have you or anyone else discussed your confusion or memory loss with a health care professional?			
Yes	74.7%	58.6%	34.8%
No	25.2%	41.5%	65.0%

Table 5: Latent class conditional response probabilities

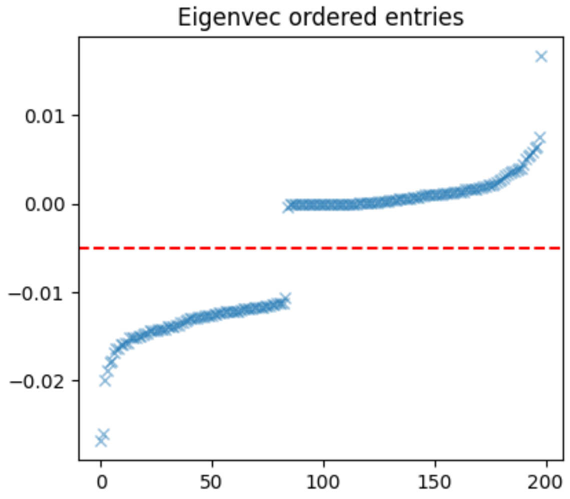

Figure 6: Ordered entries of the (second-smallest) eigenvector

nificantly lower in quantity compared to the other, which may lead to several challenges including the overshadowing of minority class patterns by the majority class, leading to difficulty in learning patterns in minority class. Moreover, models trained on imbalanced datasets may exhibit bias towards the majority class, resulting in poor generalization.

We first employ down-sampling method Figure 9 to deal with imbalanced dataset. This technique randomly removes samples from the majority class to balance the class distribution Figure 10. This helps in reducing the bias towards the majority class and allows the model to better learn patterns from the minority class. However, since our dataset high class imbalanced ratio, the train and test size becomes 1/5 compare to original. So we switch to the over-sampling method – SMOTE.

SMOTE works by synthesizing new minority class instances, thereby addressing the class imbalance problem. It creates synthetic samples that are

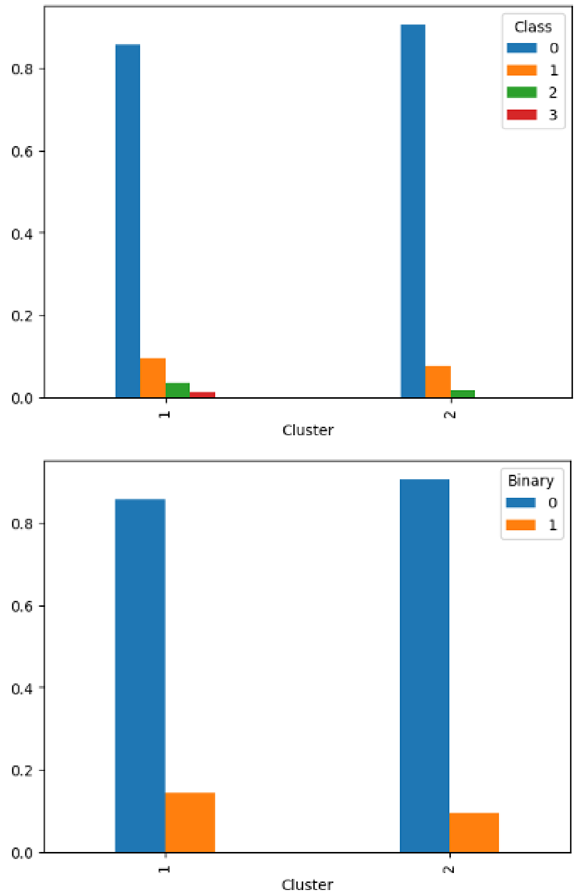

Figure 7: Distribution of the target variable: multi-class (above) vs binary (below)

similar to existing minority class samples. Now Data is balanced Figure 11. However, after applying SMOTE, we got similar accuracy and confusion matrix. This is because the class imbalance ratio is extremely high in our dataset. SMOTE generates synthetic samples that do not accurately represent the underlying distribution of the minority class, thus does not lead to improved performance on confusion matrix. Additionally, in our case, SMOTE lead to over-fitting because it generates a large number of synthetic instances. Because of the failure of SMOTE, we accepted the sample size reduction as a drawback to start modeling.

4.1.2 Modeling and Result

Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the mode of the classes as a single result. The basic idea behind Random Forest is to build a large number of decision trees, each trained on a random subset of the training data, and then combine their predictions to obtain a more accurate and robust model.

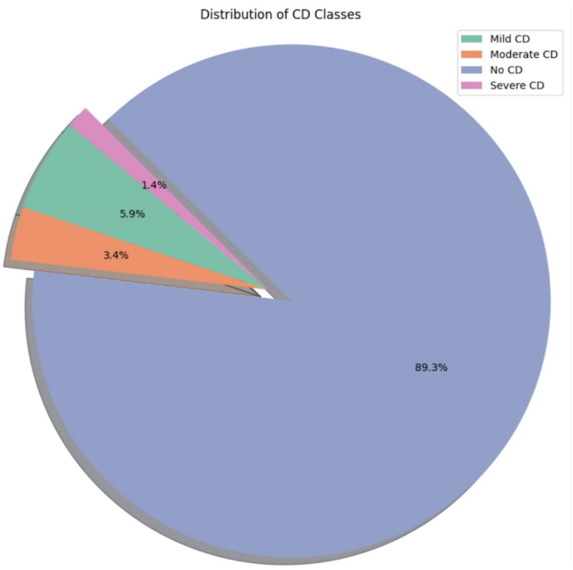

Figure 8: Unbalanced Data

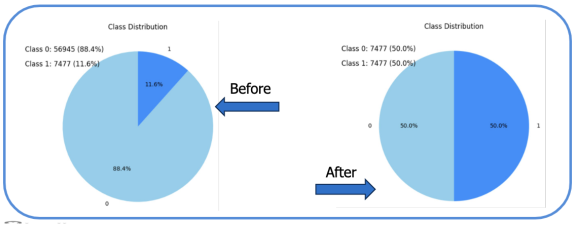

Figure 9: Down Sampling

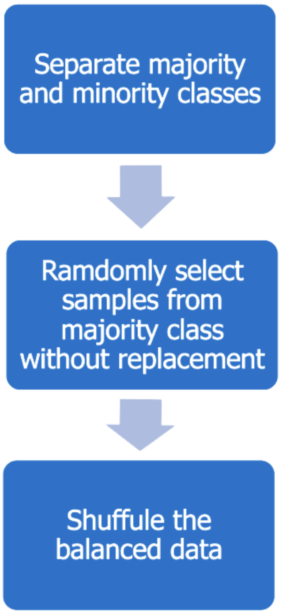
Separate majority and minority classes
Ramdomly select samples from majority class without replacement
Shuffule the balanced data

Figure 10: Down Sampling Flow

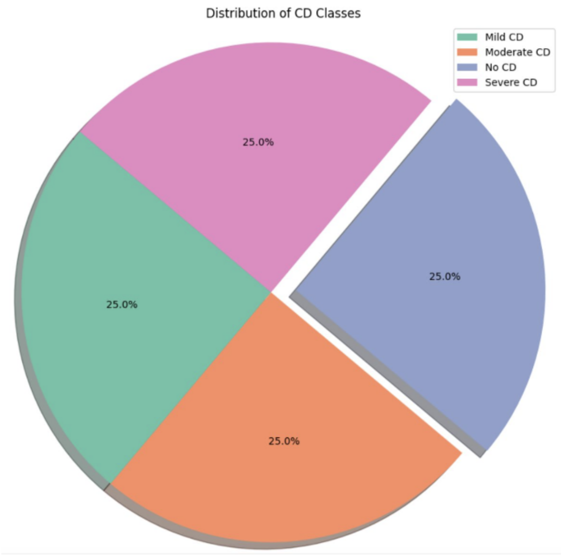

Figure 11: Balanced Data

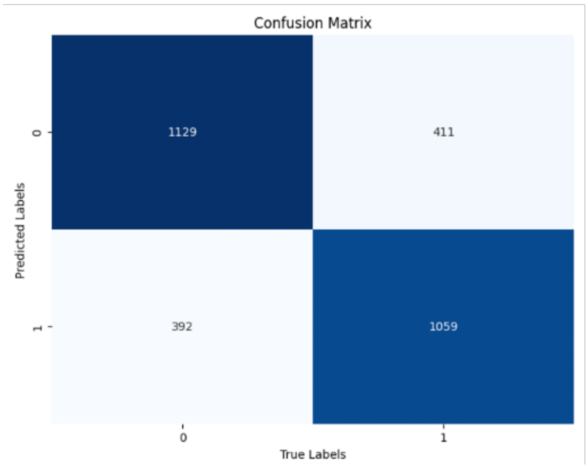

Figure 12: Logistic Regression Confusion Matrix

For Random Forest, we fine-tuned the n_estimators hyper-parameter, which represents the number of trees in the forest. The best model was obtained is n_estimators = 70. The accuracy of the Random Forest model on the test set is 73.15%. The confusion matrix (Figure 12) with more detailed representation tells that the model performance is good. The True Positive and True Negative rate are all very high.

Logistic Regression, on the other hand, is a linear classification algorithm that is used when the target variable is categorical. It models the probability that an instance belongs to a particular class using the logistic function. Logistic Regression starts with computing a linear combination of the input features, then passes the result into a sigmoid function.

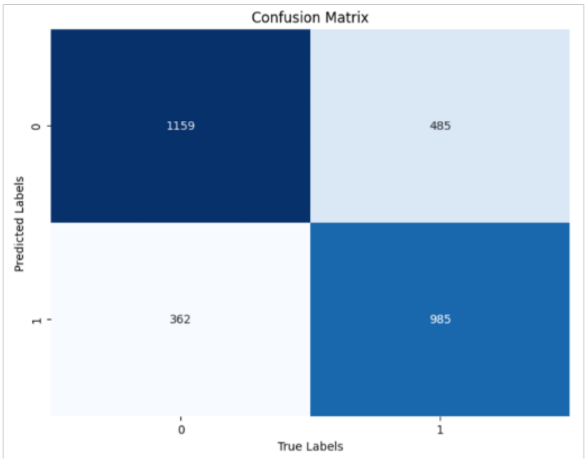

Figure 13: Random Forest Confusion Matrix

$$sigmoid(\sigma) = \frac{1}{1 + e^{-\sigma}}$$

Logistic Regression predicts the positive class with a threshold of 0.5.During training, the model parameters learned from the training data using optimization techniques such as gradient descent to find the values of the parameters that maximize the likelihood of the data. After that, regularization techniques such as L1 (Lasso) or L2 (Ridge) were applied to avoid over-fitting. For Logistic Regression, we fine-tuned the penalty and C(inverse of regularization strength) hyper-parameters. The best model was obtained with penalty='l2' and C=0.001. The accuracy of the Logistic Regression model on the test set was 71.68%. Although the accuracy of Logistic Regression model was slightly less than that of Random Forest, the confusion matrix (Figure 13) told that this model also had a good performance.

4.1.3 Multi-Layer Perceptron (MLP)

Overview MLP was a class of feed-forward artificial neural networks, consisting of at least three layers of nodes: an input layer, one or more hidden layers, and an output layer. Each node in one layer connected with a certain weight to every node in the following layer, making the network capable of learning non-linear models. MLPs used a backpropagation training algorithm, primarily with gradient descent optimization, to update the weights and minimize the error between predicted and actual outputs.

Architecture In our study, the MLP architecture (Chi) comprised five linear layers, with dropout layers inserted between each to reduce overfitting

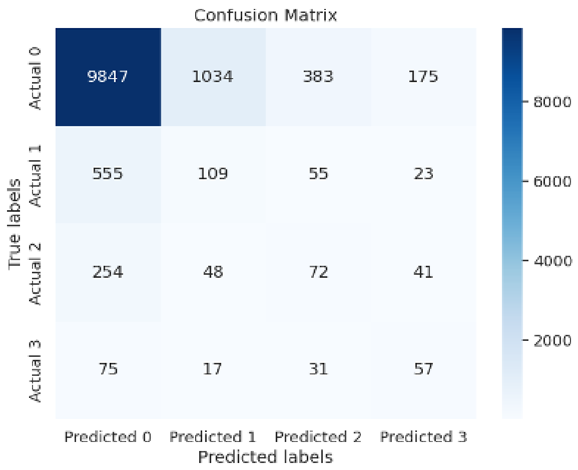

Figure 14: Confusion matrix of the MLP model

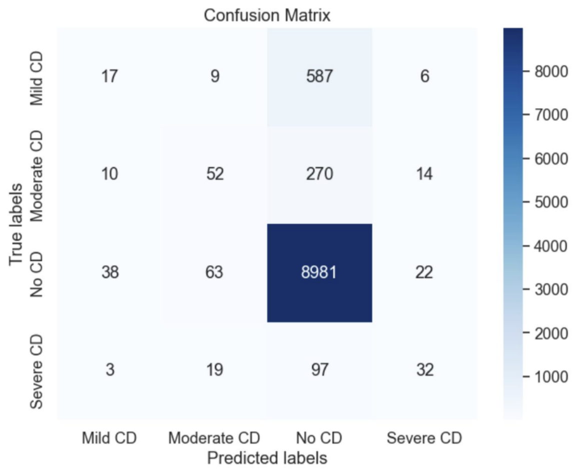

Figure 15: XGboost Confusion Matrix

by randomly omitting certain units during the training phase. ReLU activations were employed for non-linear transformations between layers.

Loss Function and Training The model was trained using a weight-adjusted cross-entropy loss function to handle class imbalance in the dataset. The network was trained over 100 epochs, with parameters adjusted iteratively to minimize the loss function. This ensured that the model learned sufficiently from the training data, while dropout layers ensured that the trained model generalized well beyond the training set.

Results and Performance The final model achieved an accuracy of 78.9%, indicating a relative robust performance across the multiple classes in the dataset. The confusion matrix (Figure 14) revealed detailed insights into the model's predictive capabilities: while it predicted no SCD with high accuracy, the predictive performance for differentiating severities of SCD showed room for improvement, particularly in reducing the number of false negatives and false positives for these classes.

4.1.4 XGBoost

XGBoost emerged as the preferred choice due to its robust performance and suitability for handling the complexities of our dataset. XGBoost offered inherent capabilities for dealing with imbalanced class distributions (Ping Zhang), a common challenge in healthcare analytics. Its gradient boosting framework allowed for adaptive weighting of minority classes, mitigating bias and improving model performance on underrepresented classes. Unlike some other algorithms, such as Random Forest, XGBoost's regularization techniques and iterative refinement process helped prevent biases towards classes with larger sample sizes. This was particularly crucial when dealing with categorical variables with differing numbers of classes, ensuring fair treatment across all categories.

In our experimentation phase, where we compared the performance of various models including logistic regression, decision trees, Random Forest, and XGBoost, the latter consistently outperformed its counterparts in terms of predictive accuracy. The highest accuracy achieved by our XGBoost model reached 88%, demonstrating its efficacy in capturing complex relationships within our dataset. The Confusion Matrix is given in Figure 15

4.1.5 Catboost

CatBoost, (Liudmila Prokhorenkova) a powerful open-source boosting library developed by Yandex, stood out for its ability to handle complex data with a large number of independent features. It was specifically designed to handle both categorical and numerical features, making it particularly useful for datasets with mixed data types. This capability allowed CatBoost to efficiently tackle a wide range of real-world problems, including regression and classification tasks. Moreover, CatBoost offered a significant advantage in handling missing values in the input data without requiring imputation, simplifying the preprocessing step and saving considerable time and effort. The library internally applied a cross-validation method to choose the best hyperparameters for the model, streamlining the process of hyperparameter tuning and ensuring robust model performance. In practical applications, CatBoost often demonstrated supe-

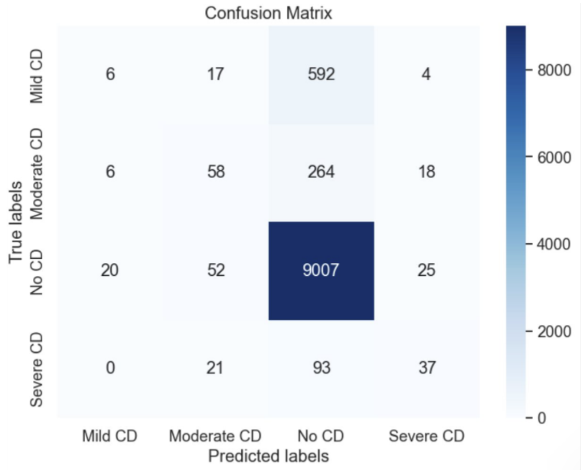

Figure 16: Catboost Confusion Matrix

rior performance compared to other algorithms, as evidenced by a notable accuracy improvement of approximately 89.11%. The visualization of the confusion matrix (Figure 16) provided valuable insights into the model's performance, aiding in identifying areas of strength and areas for optimization. Overall, CatBoost offered a comprehensive solution for handling challenging datasets, making it a compelling choice for various machine learning tasks, and enabling researchers and practitioners to achieve robust and reliable results with minimal complexity in model development and tuning.

5 Conclusion

Cognitive decline could impact individuals' health, well-being, and even interfere severely with daily living, as seen in Alzheimer's disease. However, identifying Subjective Cognitive Decline (SCD) and intervening earlier could effectively prevent dementia. In our study, we predicted the existence and severity level of SCD using BRFSS data. Through our predictive modeling work, we uncovered relationships between subjective cognitive decline and health/behavioral factors, achieving reasonable classification accuracy. However, we identified intrinsic bias within this dataset, possibly due to the survey's methodology or the specific groups surveyed. For example, our confusion matrix appeared more favorable in binary prediction than in multi-class prediction, potentially indicating noise in the dataset stemming from individuals' subjective views of degrees of cognitive decline. In binary-class prediction, Random Forest showed better performance, achieving a higher accuracy of 73.5% and outperforming other models on the confusion matrix. In multi-class prediction, CatBoost demonstrated superior performance with 89.11% accuracy and better results on its confusion matrix. Additionally, risk factors linked to SCD included social isolation, physical, mental, or emotional conditions, difficulty in performing daily activities such as dressing or bathing, and smoking duration.

References

Berke Akkaya, Ersin Sener, and Cem Gursu. 2022. A comparative study of heart disease prediction using machine learning techniques. In *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pages 1–8.

Zheru Chi. Mlp classifiers: Overtraining and solutions.

Ian J Deary, Janie Corley, Alan J Gow, Sarah E Harris, Lorna M Houlihan, Riccardo E Marioni, Lars Penke, Snorri B Rafnsson, and John M Starr. 2009. Age-associated cognitive decline. *British medical bulletin*, 92(1):135–152.

Lakshmi H.N., A. Srinivasa Reddy, and Kritika Naidu. 2023. Analysis of diabetic prediction using machine learning algorithms on brfss dataset. In *2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1024–1028.

Frank Jessen, Rebecca E Amariglio, Rachel F Buckley, Wiesje M van der Flier, Ying Han, José Luis Molinuevo, Laura Rabin, Dorene M Rentz, Octavio Rodriguez-Gomez, Andrew J Saykin, et al. 2020. The characterisation of subjective cognitive decline. *The Lancet Neurology*, 19(3):271–278.

Aleksandr Vorobev Anna Veronika Dorogush Andrey Gulin Liudmila Prokhorenkova, Gleb Gusev. Catboost: unbiased boosting with categorical features.

Youlin Shang Ping Zhang, Yiqiao Jia. Research and application of xgboost in imbalanced data.

Alma Pochini, Ben M. Williams, Hasanboy M. Isomitdinov, and Gongzhu Hu. 2015. A data mining analysis of asthma risk factors. In *2015 3rd International Conference on Applied Computing and Information Technology/2nd International Conference on Computational Science and Intelligence*, pages 349–354.

Robert Snead, Levent Dumenci, and Rebecca M Jones. 2022. A latent class analysis of cognitive decline in us adults, brfss 2015-2020. *BMC Public Health*, 22(1):1560.

Kristine N Williams and Susan Kemper. 2010. Interventions to reduce cognitive decline in aging. *Journal of psychosocial nursing and mental health services*, 48(5):42–51.