Construction and Validation of an Early Screening Model for Alzheimer's Disease based on Big Data

Leyuan Wang *

School of Public Health and Management, Guangzhou University of Chinese Medicine, Guangzhou, 510000, China

* Corresponding Author Email: Wangleyuan@gzzyydx17.wecom.work

Abstract. Alzheimer's disease (AD) is a serious neurodegenerative disease. The incidence of AD is increasing with global aging. This study utilized the Alzheimer's_disease_data dataset from the Kaggle website, which contains multi - index information on 2149 patients. After data preprocessing, Methods such as the t-test, analysis of variance, and chi-square test were used to analyze the relationships between numerical variables and categorical variables. Key indicators such as "smoking" and "family history of Alzheimer's disease" were screened out. Based on these results, four different early screening models, namely logistic regression, naive Bayes, K-nearest neighbors, and polynomial regression, were meticulously constructed. Subsequently, a series of indicators such as accuracy, sensitivity, specificity, and the area under the receiver operating characteristic curve (AUC) were compared to validate and analyze each model. The confusion matrices were also analyzed to gain a deeper understanding of the performance of each model. This study provides a direction for the optimization of early screening models for Alzheimer's disease (AD), and is expected to provide more effective tools for clinical diagnosis and intervention.

Keywords: Alzheimer's disease, early screening model, sleep quality, risk factors.

1. Introduction

Alzheimer's disease (AD) is a serious neurodegenerative disorder, characterized by disorientation, mood swings, inability to manage self - care, and behavioral problems. With the acceleration of global aging, the incidence of AD is constantly increasing and has become a global public health problem. According to the "World Alzheimer's Disease Report 2016", 47 million people worldwide are afflicted with dementia and this number may reach 131 million by 2050. It not only seriously reduces the quality of life of patients, but also imposes a heavy burden on families, healthcare systems, and society as a whole [1, 2].

Related studies have shown that associated brain lesions may begin 20 or more years before the diagnosis of Alzheimer's disease [3]. The brain's compensatory mechanism allows patients to appear normal in the initial stage of the lesions, without affecting their daily lives. When the initial changes occur, the brain compensates for them and the patients are able to continue to live and work normally. However, as the degree of the lesion deepens, the damage to the neurons in the brain will continue to worsen. The brain's compensatory mechanism will no longer be sufficient to offset the damage. Patients' cognitive abilities and language communication skills will gradually decline, and they will experience a decline in their self - care abilities, ultimately affecting their quality of life [4].

Early detection of AD is crucial. Early detection offers the possibility of early intervention, which can slow disease progression, improve patients' quality of life, and reduce the overall costs associated with long-term care. However, traditional diagnostic methods for AD usually rely on invasive procedures, such as cerebrospinal fluid analysis, or expensive imaging techniques, such as positron emission tomography (PET). Not only are these methods inaccessible to many individuals, but they may not be suitable for early screening based on large populations [5].

There are three traditional clinical diagnostic methods for AD. The first approach is Non-invasive diagnostic methods. These diagnostic methods mainly rely on clinical data from hospitals, including patient history, clinical observations, and cognitive assessments. Among them, cognitive assessment is particularly informative, and the patient's cognitive state is generally assessed by neurocognitive

scales, usually using the Mini-Mental State Examination (MMSE) [6]. The second method is Biomarker diagnostic method. Biomarkers can provide information on many mechanisms related to the occurrence of AD, such as the relationship between amyloid Abeta-42 protein and atrophy of the inner olfactory cortex and hippocampus, acetylcholine catabolism, and neuroprogenitor fibril tangles, which can help to gain insight into the underlying pathological processes of AD [7, 8]. The third technique is Modern imaging technology diagnostic methods utilize various imaging modalities. Imaging data from modalities such as computed tomography (CT), positron emission tomography (PET), diffusion tensor imaging (DTI), functional magnetic resonance imaging (fMRI) can reveal the changes in brain structure and function caused by this disease.

In addition, genetic data that can identify AD susceptibility genes has promoted a comprehensive understanding of this disease. This genetic information, while not strictly falling into the category of traditional clinical diagnostic methods, complements the existing diagnostic approaches and provides valuable insights into the underlying causes of AD [9, 10].

In recent years, the emergence of big data has changed traditional research methods in the medical field, bringing new hope to patients and their families. In the research on Alzheimer's disease (AD), big data provides an extensive and rich information source. By utilizing the power of algorithms and integrating diverse big data sources such as clinical scales, biological samples, and imaging data, it has become possible to construct an early screening model for AD. Such models have the potential to overcome the limitations of traditional diagnostic methods and provide a non - invasive, cost - effective, and widely applicable approach for early detection of AD. The focus of this study lies in the development and validation of an early - stage AD model, with the ultimate goal of improving the early identification of AD, buying more time for AD patients to fight their disease, and thus improving AD care and treatment outcomes.

In this study, four models, namely Logistic Regression, Naive Bayes, K-Nearest Neighbor and Polynomial Regression, will be employed to construct an early screening model for Alzheimer's disease. These models will be validated and compared based on performance indicators and confusion matrices to optimize and select the model that is most suitable for the early screening of Alzheimer's disease.

2. Methods

2.1. Data Source and Description

The "alzheimers_disease_data" dataset originates from the Kaggle website and is provided in CSV format. This collection of data comprises details for 2,149 individuals. It encompasses a range of attributes, including demographic information, lifestyle aspects, medical backgrounds, clinical metrics, cognitive and functional assessments, symptomatology, and Alzheimer's disease diagnostic outcomes.

2.2. Indicator Selection and Characteristics

Table 1 presents the indicators related to Alzheimer's disease in four research directions, namely demographics, lifestyle, disease history, and medical examinations. It includes the value range of each indicator and the specific reasons for selecting these indicators for inclusion in the study. Through this information, one can understand how various factors affect Alzheimer's disease from different aspects, providing a foundation and basis for the subsequent research and analysis of the early screening model for this disease.

Table 1. Attribute Information

Indicator Name	Range	Reason for Selection
Age	6090 years	65+ AD risk
Gender	0/1	Gender modulates hormones/lifestyle
EducationLevel	03	Cognitive reserve affects the risk of disease
BMI	1540 kg/m²	Abnormal BMI is related to metabolic disorders
Smoking	0/1	Damages the vascular system
AlcoholConsumption	020 g/day	Alcohol consumption harms the nervous system
PhysicalActivity	010 hours/week	Exercise boosts blood/cognition
DietQuality	score (010)	Antioxidantrich diet protects brain
SleepQuality	score (010)	Linked to AD pathology
FamilyHistoryAlzheimers	0/1	Genes raise AD risk
CardiovascularDisease	0/1	Reduced brain blood flow raises risk
Diabetes	0/1	Metabolic/nerve damage links to AD
Depression	0/1	Neuroendocrine issues accelerate brain structural changes
HeadInjury	0/1	Inflammation boosts AD risk
Hypertension	0/1	Cerebral vessel damage impairs perfusion
MMSE	score(030)	Low score = early cognitive decline
FunctionalAssessment	score (010)	Early AD sign
ADL	score (010)	Early AD: daily function decline

2.3. Method Introduction

2.3.1. Data preprocessing techniques

Due to the large volume of data and the lack of uniformity in data types, to facilitate the advancement of research, the following methods are adopted for data preprocessing: Standardization of numerical indicators (Z - score). This is applicable to continuous variables such as Age, BMI, Alcohol Consumption, and Physical Activity. It can eliminate the influence of dimensions, making variables with different units more comparable in the model. Additionally, it improves the efficiency of model training.

Encoding of categorical variables. In this study, one - hot encoding and binarization are adopted for encoding categorical variables. One - hot encoding is used for unordered categorical variables like Gender, Ethnicity, and Education Level. It transforms categorical variables into binary dummy variables, preventing the model from misjudging the sequential relationship between categories. Binarization, on the other hand, is suitable for binary - classification variables (0/1) such as Smoking and Family History of Alzheimer's.

2.3.2. Variable analysis method

To investigate the differences of numerical variables among different groups, two methods, t-test and ANOVA, were adopted. The t-test was used to analyze the mean differences of individual numerical variables such as age, education level, and BMI among different groups. The significance of the differences was determined by calculating the t-statistic and P-value. Analysis of variance (ANOVA) further explores the potential relationships between numerical variables and categorical variables. It conducts combined analyses of numerical variables with various categorical variables, and also determines the significance of the combinations based on the P - value.

For categorical variables, chi-square test is employed to analyze the correlations among various combinations of different categorical variables. By calculating the chi-square value and the corresponding P-value, it is determined whether there is a significant connection between the variables, and thereby identifying the variable combinations that may have an impact on the pathogenesis of Alzheimer's disease.

2.3.3. Model construction and validation

Based on the analysis of numerical and categorical variables, variables with significant associations were selected as key indicators for the early screening model of Alzheimer's disease. Four models were initially constructed: The Logistic Regression is:

$$P = (\mathrm{Y}=1|\mathrm{X}) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\beta n x n)}} \quad (1)$$

The Naive Bayes is:

$$\mathrm{K} = \arg\, max_{j=1}^{\mathrm{K}} P(Y=j)\Pi_{i=1}^{n} P(X_i = x_i|Y=j) \quad (2)$$

K-Nearest Neighbors: Given a new data point x, its class is determined by the majority vote of its K nearest neighbors in the training set. Calculate the distances between the new data point and all the data points in the training set, select the K data points with the closest distances, count the occurrence frequencies of each class among these K points, and the class with the highest occurrence frequency is the predicted class of the new data point x.

The Polynomial Regression is:

$$\mathrm{y} = \beta_o + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n + \epsilon \quad (3)$$

These four models were then verified and compared. Evaluation was conducted using performance indicators like accuracy and area under the receiver operating characteristic (AUC) value. Additionally, confusion matrices for each model were analyzed. The confusion matrix shows the relationship between the actual and predicted results. By examining the correct and incorrect predictions for different categories, this paper can better understand each model's performance in classification. This comprehensive evaluation of model performance and confusion matrix results provides a basis for subsequent model optimization and selection.

3. Results and Discussion

3.1. Analysis of Numerical Variables

T-tests were conducted for multiple numerical variables such as age, education level, and BMI. The results shown in Figure 1 indicate that the P-values of all variables are greater than 0.05. This suggests that under the current research conditions, there were no significant differences in the mean values of these numerical variables across different groups. This series of data shows that, under the current research conditions, the differences in the means of these numerical variables among different groups (such as the group of Alzheimer's disease patients and the group of non-patients) are not significant. Therefore, it can be inferred that in the initial analysis stage, the distribution of a single variable in different populations was relatively similar and could not be directly used as an effective indicator for distinguishing whether someone has Alzheimer's disease.

In order to further explore the potential relationships among variables, this study conducted an ANOVA analysis. As can be seen from the Figure 2, the P-values of most numerical variables combined with categorical variables were greater than 0.05, failing to reach a significant level. However, for the combinations of "functional assessment" and "diabetes" (with a P-value of approximately 0.0435) as well as "diet quality" and "hypertension" (with a P-value of approximately 0.0432), the P-values were less than 0.05, showing significant differences. This suggests that the state of diabetes may have a significant impact on the functional assessment indicators, and there is a significant association between hypertension status and diet quality. These findings have potential application value in the early screening process of Alzheimer's disease (AD) and are worthy of further in-depth research.

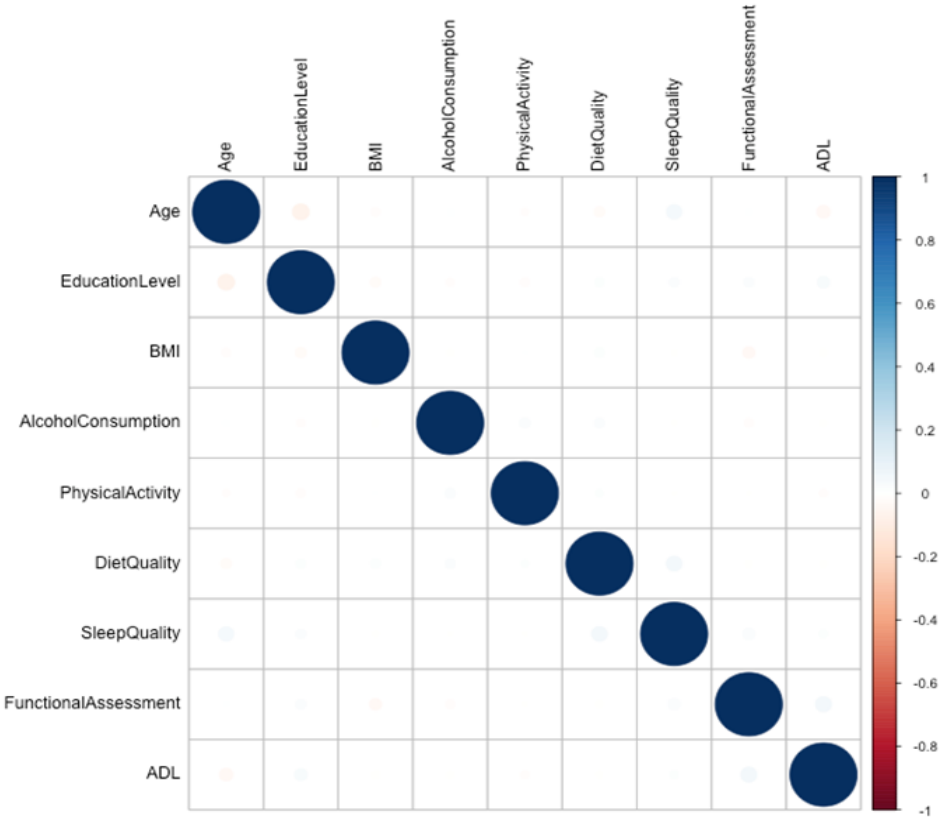

Figure 1. Numeric_vars_correlation

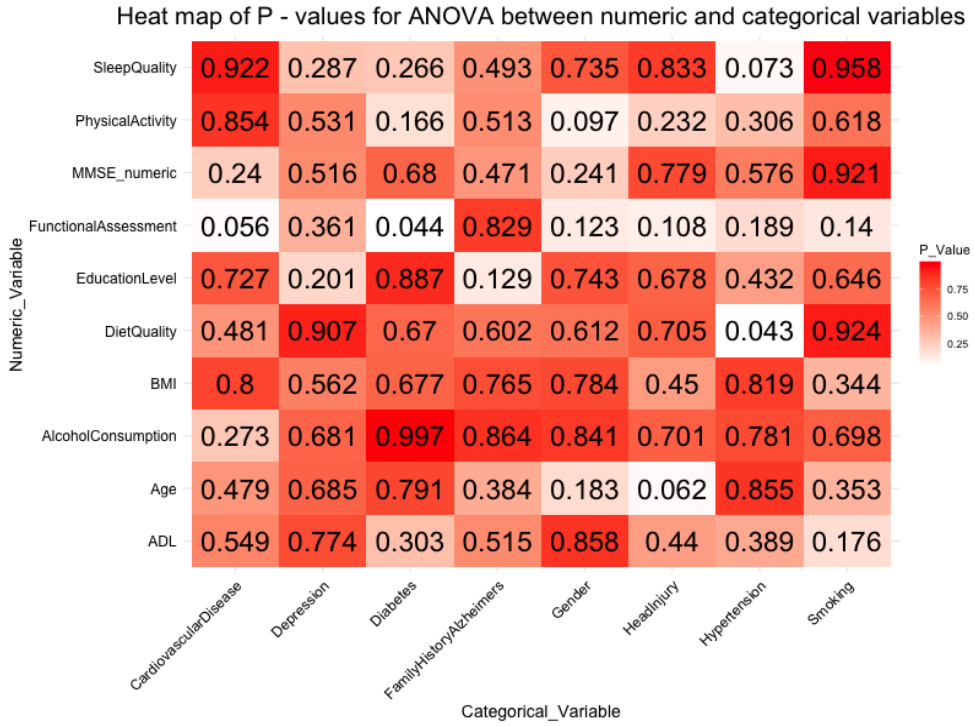

Figure 2. Heat map of P-values for ANOVA between numeric and categorical variables

3.2. Analysis of Categorical Variables

In the analysis of categorical variables, this study employed the chi-square test (Figure 3). The P-values of most variable combinations were greater than 0.05, indicating no significant association among the variables. However, for the combination of "smoking" and "family history of Alzheimer's disease", the P-value was approximately 0.0386, which was less than 0.05, suggesting a significant

correlation. This indicates that there is a certain connection between smoking behavior and the family history of Alzheimer's disease. Among individuals with a family history, the distribution of smoking behavior may differ from that among those without a family history. Smoking might be a potential risk factor influencing the occurrence of Alzheimer's disease, and it interacts with genetic factors in a complex manner.

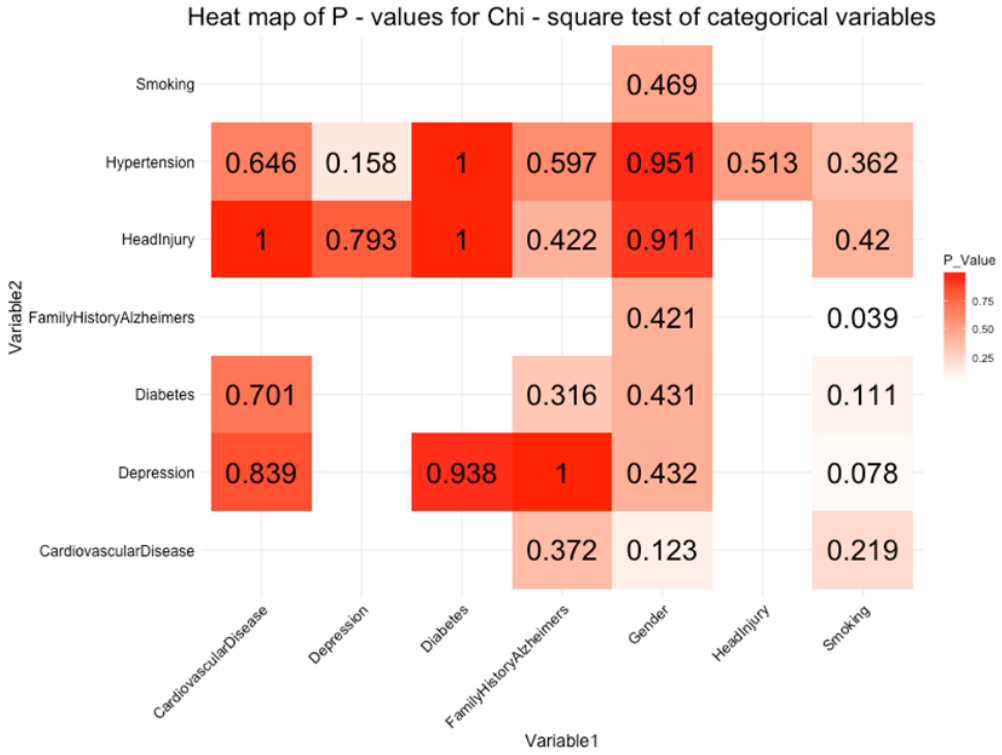

Figure 3. Heat map of P-values for Chi-square test of categorical variables

3.3. Model Results

Based on the above analysis results, this study selected the variables with significant correlations, such as "smoking", "family history of Alzheimer's disease", "functional assessment under diabetes state", and "diet quality under hypertension state", as key indicators and incorporated them into the early screening model for Alzheimer's disease. During the model construction process, the research team adopted various methods and fully considered the interactions among variables. Through advanced techniques such as stepwise regression, numerous variables were screened to determine the variable combinations that made greater contributions to the early screening of Alzheimer's disease. Thus, the specific form and parameters of the model were clarified, and the preliminary construction of four models, namely Logistic Regression, Naive Bayes, K-Nearest Neighbors, and Polynomial Regression, was completed.

From the various indicators of model performance as shown in the Table 2, the accuracy rate of the logistic regression model is 0.6537, and its specificity reaches 1, indicating that it has a strong ability to identify negative cases. However, the AUC value is only 0.4808, suggesting that there is still room for improvement in the ability to distinguish between positive and negative cases. The accuracy of the Naive Bayes model is 0.6553, the sensitivity is 0.0090, the specificity is as high as 0.9976, and the AUC is 0.5033. This shows that the model performs well in terms of specificity, but the sensitivity is relatively low, and there is a possibility of missing some positive cases The accuracy of the K-Nearest Neighbors model is 0.5720, which is the lowest among these models. Its sensitivity is 0.2377, the specificity is 0.7506, and the AUC is 0.4941, reflecting that both the classification accuracy and the discrimination ability of this model need to be strengthened. The accuracy rate of the Polynomial Regression model is 0.6491, with a specificity of 0.9929 and an AUC of 0.5274. It performs reasonably well in specificity but also requires further optimization in overall performance.

Table 2. Model results

Model	Accuracy	Sensitivity	Specificity	AUC
Logistic Regression	0.6537	0	1	0.4808
Naive Bayes	0.6553	0.0090	0.9976	0.5033
K-Nearest Neighbors	0.5730	0.2377	0.7506	0.4941
Polynomial Regression	0.6491	0	0.9929	0.5273

While establishing the model, this study also conducted a horizontal comparative analysis of four models. Firstly, a comparative analysis was carried out among the various indicators of the models. In terms of accuracy, the Naive Bayes model has the highest accuracy, the K-Nearest Neighbors model has the lowest, and the remaining models are similar. However, accuracy alone cannot comprehensively evaluate the performance of the models. In terms of sensitivity, K-Nearest Neighbor was relatively the highest, while the other models were generally lower. In terms of specificity, all models are generally high, and they can all identify non-patient samples well. In terms of AUC values, Polynomial Regression and Naive Bayes had relatively higher values, indicating that their ability to distinguish between patient and non-patient samples was relatively stronger.

When considering the confusion matrices of each model (as shown in the figure), the accuracies of both the Logistic Regression model and the Polynomial Regression model are 0.65. The distribution of the matrices of these two models reflects the matching between the actual values and the predicted values. The accuracy of the Naive Bayes model is 0.66, and from its confusion matrix, the advantages and disadvantages of this model in the classification process can be seen. The accuracy rate of the K-Nearest Neighbor model is 0.57, and the matrix shows that there is a large deviation in the category judgment.

By synthesizing the evaluation indicators and the results of the confusion matrices of these models, it can be concluded that different models have their own advantages and disadvantages in the early screening of Alzheimer's disease (AD), which provides important references for the subsequent optimization and selection of models.

Figure 3 shows the result of Confusion Matrix of Logistic Regression Model. This matrix is used to evaluate the performance of the logistic regression model in the early screening of Alzheimer's disease. The number of cases where the prediction is 0 and the actual value is also 0 is 0; the number of cases where the prediction is 0 but the actual value is 1 is 223; the number of cases where the prediction is 1 and the actual value is 0 is 421; the accuracy rate of the model is 0.65. However, from the data of the matrix, it can be seen that the model makes more errors in predicting samples with actual value of 1.

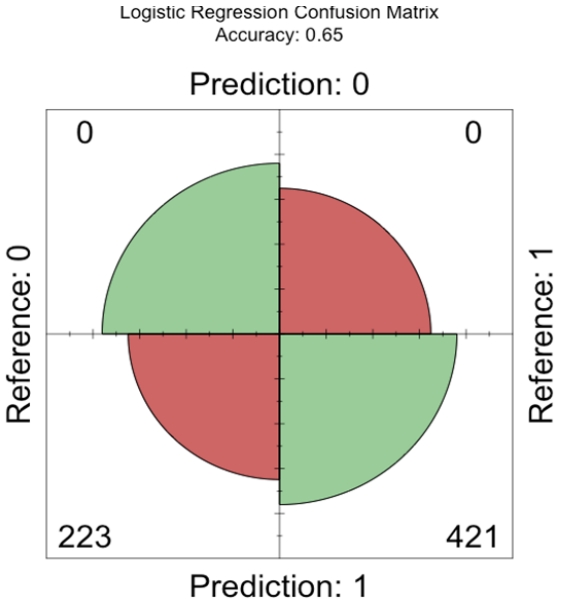

Figure 3. Logistic Regression Confusion Matrix

From Figure 4, it can be seen the results of the confusion matrix of the Polynomial Regression model. It shows the evaluation of the Polynomial Regression model. The number of cases where the prediction is 0 and the actual value is also 0 is 3; the number of cases where the prediction is 0 but the actual value is 1 is 223; and the number of cases where the prediction is 1 but the actual value is 0 is 418. It has the same accuracy of 0.65 as the Logistic Regression model, but there are differences in the specific prediction results.

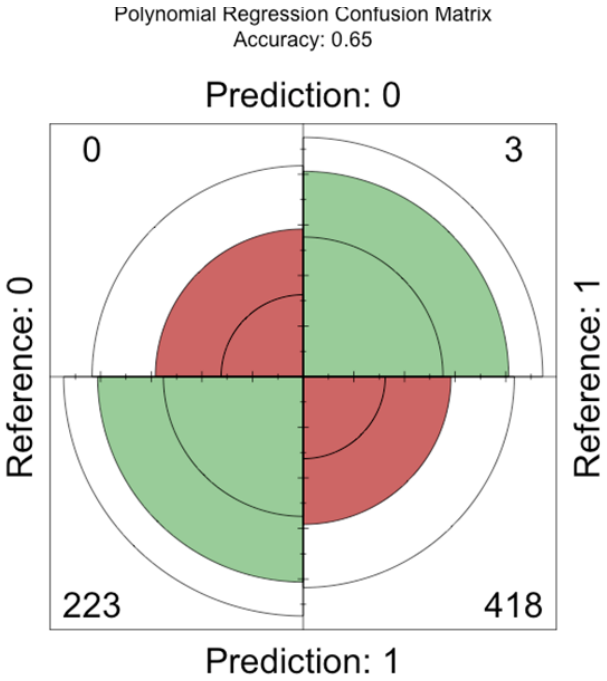

Figure 4. Polynomial Regression Confusion Matrix

Figure 5 presents the findings about the confusion matrix of the nearest neighbor model, which is used to evaluate the disease prediction effect of the K - nearest neighbor model. For predictions of 0, the actual 0 cases are 53 and the actual 1 cases are 170. For predictions of 1, the actual 0 cases are 105 and the actual 1 cases are 316. The accuracy rate of this model is 0.57, which is lower than that of the previous two models, indicating that its overall prediction accuracy is relatively poor.

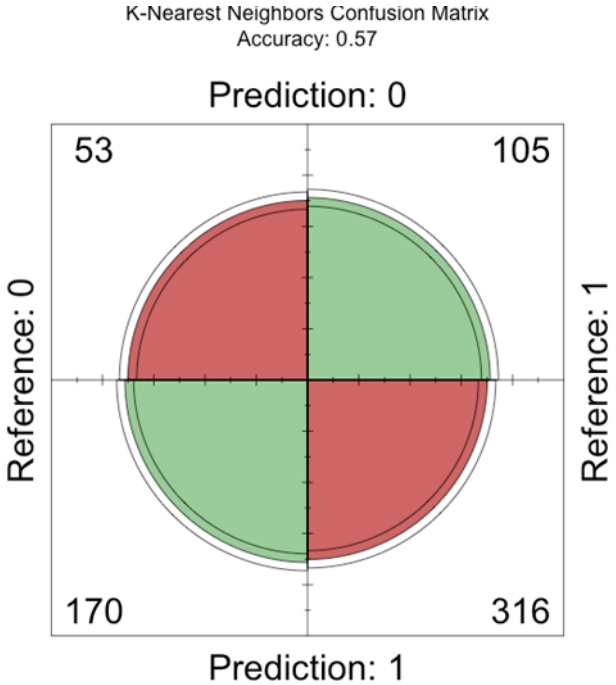

Figure 5. K-Nearest Neighbors Confusion Matrix

Figure 6 shows the outcomes of the confusion matrix of the Naive Bayes model. The confusion matrix reflects the performance of the Naive Bayes model in the early screening. For predictions of 0, there are 2 cases where the actual is also 0 and 221 cases where the actual is 1. For predictions of 1, there is 1 case where the actual is 0 and 420 cases where the actual is 1. The accuracy of the model is 0.66, which is the highest among the four models. It has an advantage in identifying samples of patients who actually have Alzheimer's disease.

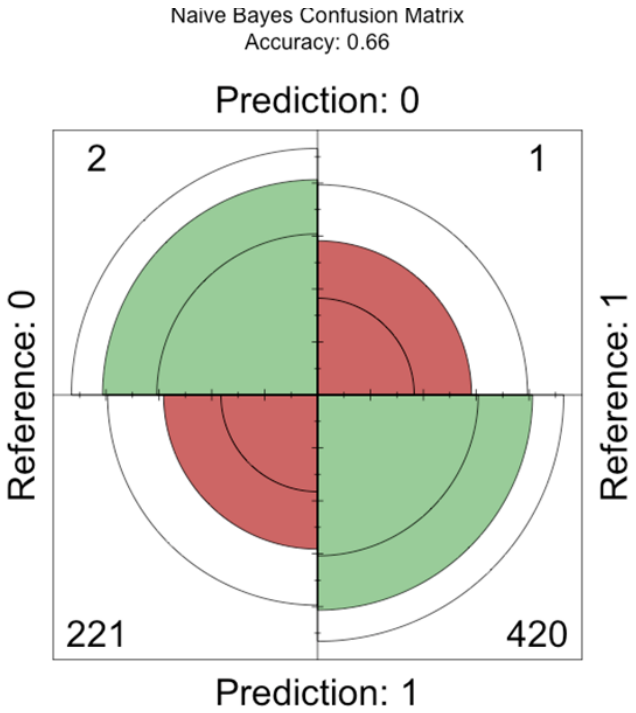

Figure 6. Naive Bayes Confusion Matrix

Overall, in the early screening of Alzheimer's disease, the Logistic Regression model and the Polynomial Regression model show similar performance in multiple indicators, and perform poorly in key indicators such as sensitivity and AUC. The K-Nearest Neighbors model has a low accuracy rate and its overall performance is also unsatisfactory. The Naive Bayes model has relatively outstanding comprehensive performance in multiple indicators. For instance, it has certain advantages in terms of accuracy and AUC values, and its specificity is also relatively high. Although the Naive Bayes model is relatively good, the overall performance of all models has not reached the ideal state. There is room for improvement in the ability to identify diseased samples. Different models exhibit diverse performance characteristics in the early screening task of Alzheimer's disease. This paper can try methods such as ensemble learning to combine the advantages of multiple models to improve the overall performance of the model, further optimize the model structure and parameters, and better apply it to the early screening of Alzheimer's disease to improve the accuracy and reliability of the screening model and gain valuable time for clinical diagnosis and intervention.

3.4. Disucssion

Through a systematic analysis of multiple numerical and categorical variables, this study identified several potential risk factors related to Alzheimer's disease and successfully constructed and validated an early screening model. To a certain extent, the model has the ability to distinguish between early Alzheimer's disease (AD) patients and non-patients, but it still needs further optimization and improvement research can expand the sample size, incorporate more potential influencing factors and details, etc., to further enhance the accuracy and universality of the model and provide more effective tools for the early diagnosis and intervention of Alzheimer's disease.

4. Conclusion

In this study, an early screening model for Alzheimer's disease was constructed through big data analysis. Methods such as the t - test and analysis of variance were used to study numerical variables like age and BMI, as well as categorical variables such as gender and smoking status. It was found that there were significant associations among variable combinations, including "functional assessment" and "diabetes", "diet quality" and "hypertension", and "smoking" and "family history of Alzheimer's disease". Based on these findings, four early screening models were constructed by selecting key variables. The performance of these models was evaluated using indicators such as accuracy, sensitivity, specificity, and AUC values, as well as the confusion matrix. The results showed that the Naive Bayes model had relatively better overall performance, but the overall performance of all models had not yet reached the ideal state. Possible improvement methods include increasing the sample size, optimizing feature selection, and adjusting model parameters. Visualization methods effectively presented the relationships between variables and the performance of the model. These findings are helpful for understanding the risk factors of Alzheimer's disease, providing clearer decision - making bases for clinicians and researchers, and providing guidance for developing more accurate early screening models.

References

[1] Anh T, Bao T P, Young J K, et al. Alzheimer's diagnosis using deep learning in segmenting and classifying 3D brain MR images. The International journal of neuroscience, 2020, 132 (7): 11 - 13.

[2] Wang Yinhua, Ji Yong. Current Development Status of Alzheimer's Disease in the World. Chinese Journal of Contemporary Neurology and Neurosurgery, 2015, 15 (07): 507 - 511.

[3] VINCENT D, et al. Aβ deposition, cortical thickness, and memory. Jama Neurology, 2013, 70 (7): 1 - 9.

[4] Sun Minglang. Research on Early Diagnosis of Alzheimer's Disease Based on Deep Learning. Dalian University of Technology, 2021.

[5] Abiad E, Kuwari A, Aani A U. et al. Navigating the Alzheimer’s Biomarker Landscape: A Comprehensive Analysis of Fluid-Based Diagnostics. Cells, 2024, 13 (22): 1901 - 1901.

[6] Han Bei, Dai Linlin, Li, et al. Significance of Cognitive Intervention on Cognitive Dysfunction and Quality of Life of Patients with Alzheimer's Disease. China Medical News, 2019.

[7] Kumar H, Datusalia K A, Kumar A, et al. Network pharmacology exploring the mechanistic role of indirubin phytoconstituent from Indigo naturalis targeting GSK-3β in Alzheimer's disease. Journal of biomolecular structure & dynamics, 2025, 11 - 14.

[8] Du Yifeng. Progress in the Diagnosis and Treatment of Alzheimer's Disease in 2018. Journal of International Geriatrics, 2017, 38 (2): 52 - 55.

[9] Gao F. Integrated Positron Emission Tomography/Magnetic Resonance Imaging in clinical diagnosis of Alzheimer's disease. Eur J Radiol, 2021, 145: 110017.

[10] Talwar P, Kushwaha S, Chaturvedi M, et al. Systematic Review of Different Neuroimaging Correlates in Mild Cognitive Impairment and Alzheimer's Disease. Clin Neuroradiol, 2021, 31 (4): 953 - 967.