**CPSC 5310 01 Machine Learning**

**Project\_Part\_5**

**Team 4 – Priyadarshini Shanmugasundaram Murugan, Archana Janardhana Vinoda Kumari,**

**Priyanka Vysyaraju**

**REPORT ON OUR PROJECT – PREDICTING ALZHEIMER**

**Our Topic:**

Alzheimer's disease is a neurodegenerative disorder causing memory loss, cognitive decline, and behavioral changes. Diagnosis involves medical history, cognitive tests, and, in some cases, brain imaging. While there is no cure, ongoing research seeks to develop effective treatments. Predictive modeling using machine learning involves creating a model to predict the likelihood of Alzheimer's based on specific input features.

We've opted to analyze the "predict\_alzheimers" dataset, encompassing diverse demographic and clinical characteristics related to Alzheimer's disease. This dataset comprises 33 columns, encompassing both categorical and numerical data.

**Data Preprocessing:**

Data preprocessing plays a pivotal role in readying data for machine learning models. This process encompasses understanding the dataset, rectifying missing values and outliers, encoding categorical variables, normalizing numerical features, and conducting feature engineering. Furthermore, it entails data splitting, tackling class imbalances, handling text data, as well as normalizing and checking data skewness, while meticulously documenting all preprocessing steps. The specific actions undertaken depend on the dataset's characteristics and the machine learning task's requisites, often involving iterative refinement based on model performance. Initially, we planned to preprocess the data by addressing missing values, encoding categorical variables, and scaling numerical features. However, since our dataset contains no missing values, there is no need for data cleaning.

**Regression:**

The objective is to utilize machine learning regression techniques for predicting cognitive decline, focusing on a target variable with clinical relevance to Alzheimer's disease and cognitive decline. The target variable should display substantial correlations with multiple features, suggesting their potential predictive value for changes in the target variable. In light of these criteria, MMSE (Mini-Mental State Examination) scores (mmse.bl) are commonly employed as a standard cognitive function measure in Alzheimer's disease research, demonstrating broad clinical relevance. Following their statistical significance concerning the target variable (MMSE baseline score), the features nlabel, adas13.bl, cdrsb.bl, adas11.bl, faq.bl, and ravlt.immediate.bl are identified as the top six features for predicting cognitive decline. This selection indicates their potential as robust predictors of cognitive decline and underscores their importance in regression analysis.

We have explored several regression models—linear regression, random forest regression, support vector regression, and ridge regression—to forecast cognitive decline using our dataset.

The Linear Regression Model demonstrates moderate predictive capability, yielding an R-squared value of 0.6439, signifying that approximately 64.39% of the variance in the target variable is accounted for by the model. Its RMSE of 1.7166 indicates that, on average, predictions are within about 1.71 units of the actual values.

Surpassing the Linear Regression model, the Random Forest Regressor exhibits superior performance, boasting a higher R-squared value of 0.6632, suggesting it can explain around 66.32% of the variance. Moreover, its RMSE is lower at 1.6694, indicating more precise predictions.

In contrast, the Support Vector Regression (SVR) Model displays notably inferior performance compared to the other models, with an R-squared value of 0.6163. This indicates it can only elucidate approximately 61.63% of the variance in the target variable, with its higher RMSE of 1.7819 implying less accurate predictions.

Similarly, the Ridge Regression Model shows performance akin to the Linear Regression model, yielding an R-squared value of 0.644. This indicates it can account for about 64.4% of the variance in the target variable, with its higher RMSE of 1.716 suggesting less precise predictions.

Based on these observations, the Random Forest Regressor emerges as the most favorable model for this task, effectively balancing predictive precision and the capacity to explain variance in the target variable. Conducting a feature importance analysis for the Random Forest Model reveals that 'nlabel' stands out as the most influential feature for predicting cognitive decline among the top 10 features chosen for prediction.

**Clustering:**

We believe that utilizing clinical scores such as the Clinical Dementia Rating Scale, Alzheimer’s Disease Assessment Scales, Mini-Mental State Examination score, and Functional Activities Questionnaire score can be beneficial in identifying subgroups of patients with similar disease severity or cognitive function. Employing the elbow method, we determined the optimal number of clusters to be three. Subsequently, we computed the silhouette score for four distinct clustering techniques—K-means, Agglomerative, Mini batch K-means, and mean shift—to classify patients into these three clusters. The silhouette score gauges the similarity of an object to its assigned cluster in comparison to other clusters. A higher silhouette score signifies superior clustering performance, with a score nearing 1 indicating a perfect match.

Based on the scores, K-Means exhibits the highest silhouette score, indicating its suitability as the most appropriate clustering method for this dataset among those examined. The scatter plots depicted the clustering outcomes for each method (K-Means, Agglomerative, Mini-Batch K-Means, and Mean-Shift) after reducing the dimensionality of the clinical scores data to two principal components for enhanced visualization. Each point denotes a patient, with cluster assignments indicated by distinct colors.

**Classification:**

The task is to classify whether an individual has Alzheimer’s disease or not using machine learning for binary classification. Our steps include collecting a relevant dataset, exploring its structure, preprocessing by handling missing values and outliers, encoding variables, and splitting data into training and testing sets. We will choose an appropriate classification algorithm, train the model, and evaluate its performance using metrics. Optionally, we would like to tune hyperparameters for optimization, interpret model decisions and validate predictions.

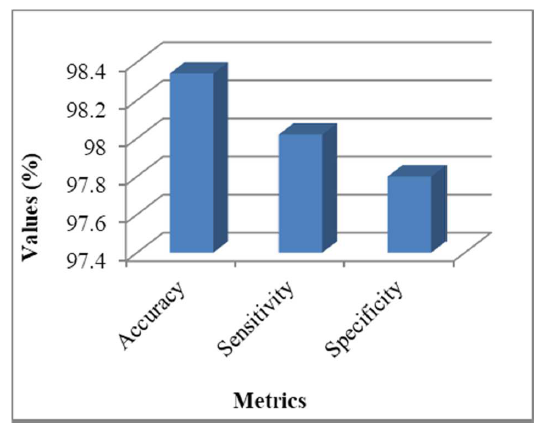
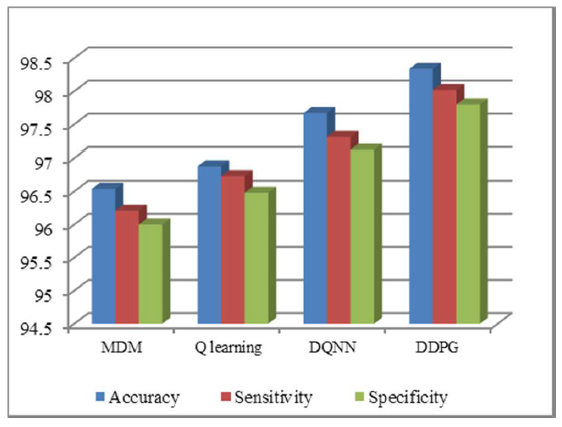
We assessed the accuracy of five classification models: logistic regression, K-nearest neighbors, decision trees, support vector machine, and naive Bayes. By employing both the elbow method and Grid Search CV, we determined the optimal k value for the K-Nearest Neighbors (KNN) algorithm to be 8. This refinement ensures a finely tuned model that maximizes predictive accuracy for our dataset.

Among the models tested, the naive Bayes classifier exhibited the highest accuracy, achieving 74.4%, rendering it the most effective model for this task. The notably high accuracy of the Decision Trees model suggests potential overfitting to the training data, thereby diminishing its reliability on unseen data. The success of the naive Bayes model may stem from its adeptness at managing independence assumptions between features, a trait that appears well-suited to this particular dataset.

**REPORT ON THREE TECHNICAL RESEARCH PAPERS**

**Alzheimer’s Disease Prediction with K-means Clustering and Reinforcement Learning Approach (2023)**

In this paper, a system was developed to aid in the early detection of Alzheimer's disease (AD) through the analysis of MRI images of the brain. The process involves five key phases: data collection, preprocessing, image segmentation utilizing a traditional k-means clustering approach augmented with mathematical morphology, feature extraction employing Convolutional Neural Networks (CNN), and classification employing reinforcement learning methods. To evaluate the efficacy of the proposed system, comprehensive evaluation metrics such as accuracy, specificity, and sensitivity are utilized. These metrics provide insights into the system's ability to accurately diagnose different stages of Alzheimer's disease, thereby assessing its clinical utility and reliability. Results indicate that the reinforcement learning approach, particularly the DDPG algorithm, achieves notable improvements in terms of reward by learning optimal policies. Moreover, comparative analyses with other models demonstrate the superior performance of the DDPG algorithm, characterized by a more refined policy space and lower error values in classification tasks. Overall, the developed system showcases promising potential in aiding early detection efforts and advancing our understanding of Alzheimer's disease pathology.

**Data-Driven Approach Based on Feature Selection Technique for Early Diagnosis of Alzheimer’s Disease (2020)**

The paper proposes a data-driven approach using feature selection techniques for early diagnosis of Alzheimer's Disease (AD). It utilizes neuropsychological scores like the Mini-Mental State Examination (MMSE) along with structural MRI features from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. Various machine learning algorithms like Support Vector Machines (SVM), Decision Trees, Logistic Regression, and ensemble methods like AdaBoost, XGBoost, and Random Forests are employed and compared. The ADNI dataset containing 119 cognitively normal, 155 mild cognitive impairment and 46 AD subjects is used. Preprocessing steps include z-score normalization to scale different feature ranges and the SMOTE technique to handle class imbalance. A filter-based feature selection method is applied to identify the most relevant features like hippocampal volumes and MMSE scores for the different classification tasks of CN vs AD, MCI vs AD, and CN vs MCI. Among the machine learning models, SVM with an RBF kernel achieved the highest accuracies of 99.2% for CN vs AD, 78.5% for CN vs MCI, and 91.3% for MCI vs AD when MMSE scores were included along with MRI features. The proposed approach outperformed other methods from the literature in terms of accuracy, sensitivity, and specificity, highlighting the significance of combining neuropsychological tests with imaging data for AD diagnosis. One of the key implementation challenges was tuning the hyperparameters of the various machine-learning algorithms through techniques like grid search and cross-validation to obtain optimal performance. The authors also note the need to explore other neuropsychological tests besides MMSE and investigate how brain atrophy patterns correlate with these test scores for more accurate and robust results.

**Table 1: Performance of SVM based on feature selection with and without MMSE**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Performance Measure | CN vs AD Without MMSE | CN vs AD With MMSE | CN vs MCI Without MMSE | CN vs MCI With MMSE | MCI vs AD Without MMSE | MCI vs AD With MMSE |
| Accuracy | 84.89 | 99.2 | 67.56 | 78.5 | 75.15 | 91.3 |
| Precision | 0.85 | 0.99 | 0.68 | 0.79 | 0.75 | 0.91 |
| Recall | 0.85 | 0.99 | 0.67 | 0.78 | 0.75 | 0.91 |
| F1-Score | 0.85 | 0.99 | 0.67 | 0.78 | 0.75 | 0.91 |
| AUC | 0.849 | 0.992 | 0.674 | 0.784 | 0.752 | 0.913 |
| Sensitivity | 0.84 | 1.00 | 0.59 | 0.72 | 0.81 | 0.92 |
| Specificity | 0.86 | 0.98 | 0.76 | 0.85 | 0.70 | 0.90 |

**Early Diagnosis and Prevention of Alzheimer’s Disease Based on Regression Prediction Model (2023)**

This Paper proposes a novel approach to diagnosing Alzheimer's disease by combining MLR, PCA, and Spearman correlation analysis. The objective is to identify highly correlated indicators for more accurate diagnoses. The study preprocesses Alzheimer's disease data and creates a holistic evaluation model using a multidimensional analysis method. The MLR model is established using PCA, MATLAB's regression function, and ROC curve analysis. The proposed model shows promising results across various stages of Alzheimer's disease, enhancing predictive accuracy, refining the feature space, and improving patient outcomes and clinical decision-making.

It highlights the importance of early detection and intervention, emphasizing the irreversible nature of AD progression and the significance of prevention over cure. It underscores the need for identifying risk factors, early symptoms, and timely diagnosis to facilitate effective disease management. Preventing Alzheimer's disease involves three levels of intervention. Primary prevention focuses on delaying cognitive dysfunction symptoms through lifestyle modifications, medication, and educational activities. Secondary prevention prevents the progression of dementia in individuals with mild cognitive impairment (MCI). Tertiary prevention aims to control disease progression and complications through Cholinesterase inhibitors (ChEIs) treatment, memory, and strategy training.

Overall, this paper proposes a multiple linear regression model for early Alzheimer's disease prevention, identifying key factors influencing symptoms like ADAS13, FAQ, and MMSE. By analysing these factors, it aims to advance the understanding of AD pathogenesis and develop effective interventions to delay disease progression, offering practical solutions for early AD prevention in the medical field.

