Logo

Description automatically generated

**Module 5: Final Project**

**By**

**Elvis Opoku, Kevin Tushar Pandya, Neha Ajinkya Nagurkar, Vidhi Manojkumar Patani, Smit Pareshbhai Ranpariya**

**College of Professional Studies: Northeastern University**

**Prof. Shahram Sattar**

**Subject: ALY6040: DATA MINING APPLICATION**

**Date: May 10th, 2024**

**INTRODUCTION**

The Disability and Health Data System (DHDS) data set is an extensive online collection of state-level data on persons with impairments. It covers six functional impairment types: cognition, hearing, mobility, vision, self-care, & independent living. DHDS allows users to access critical data on the physical and psychological well-being of people with disabilities, providing insights into their difficulties and needs across a variety of disciplines.

Initially, the dataset had 644,356 items and 32 columns that included a wide variety of disability and health-related characteristics. However, following extensive data pre-processing as well as cleaning methods, such as resolving missing values and deleting extraneous columns, the collection of data was reduced to 644,356 items and 17 columns. This guaranteed that the data was ready for future analysis and research, allowing for important insights into the features and patterns of disability among people.

Our approach seeks to find patterns and insights into adult disability and wellness dynamics. We use a variety of machine learning approaches, including a Decision Tree Classifier, a Random Forest Algorithm, Support Vector Machine (SVM), Linear Discrimination Analysis (LDA), & Logistic Regression. Each model is assessed by considering its accuracy and effectiveness in predicting disability kinds and examining their links with medical and demographic factors.

**OBJECTIVE**

The primary goal of this research is to examine the DHDS dataset and uncover patterns, trends, and discoveries about disability and health in adults. Specifically, the analysis focuses at:

* Preprocess the dataset to verify its integrity and completeness.
* Conduct EDA to better understand the distribution of and interactions between variables.
* Identify any unusual values, missing numbers, or discrepancies in the dataset.
* Extract valuable information to help shape public health initiatives and laws.

With addition to exploratory research, multiple data mining models are going to be created and tested on the DHDS dataset. The models include:

* Decision Tree Classifier: A decision tree-based classification system that predicts handicap categories based on relevant information collected from the dataset.
* Random Forest Classifier: Uses ensemble learning to improve classification accuracy by combining numerous decision trees.
* Support Vector Machines (SVM): Using SVM using feature selection to categorize handicap categories and maximize decision boundary margins.
* Linear discriminating analysis (LDA): Using LDA to determine linear discriminant characteristics that best differentiate disability groups.
* Logistic Regression: Using logistic regression in tasks involving binary classification to predict disability status using demographic and health-related characteristics.

These models will be rigorously evaluated using measures that include precision, recall, precision, accuracy, and the F1-score. In addition, confusion matrices are going to be constructed to show how well each model predicts disability categories. The insights that result from these models will help to deepen our comprehension of disability and wellness dynamics, directing the development of tailored treatments and policies.

**SCOPE**

The study focuses upon the DHDS dataset, which provides detailed information about disability kinds, health indicators, demographic characteristics, and geographic areas. The scope includes complete data preparation, exploratory analysis of data (EDA), & the creation of data mining methods to obtain insight into trends in adult impairment and health.

More precisely, the scope includes:

* Data Preparation: Complete data preparation, including managing missing values, correcting inconsistencies, and identifying pertinent characteristics for analysis.
* Exploratory analysis of data (EDA): Comprehensive visualization & statistical analysis are used to comprehend the pattern of distribution of variables, detect relationships, and discover patterns throughout the dataset.
* Model Development entails creating and testing machine learning models that include decision tree models A classifier, Random Forest Classifier, the Support Vector Machine (the SVM), Linear Dis discriminant analysis (LDA), & Logistic Regression. These models seek to predict disability kinds and examine their correlations with various well-being and demographic variables.
* Performance evaluation includes evaluating the performance of generated models using measures such as precision, recall, precision, and F1-score. Confusion matrices are going to be used to demonstrate the model's prediction skills across various handicap categories.

The findings of this research project will contribute to an improved comprehension of adult impairment and medical dynamics, allowing policy makers, health care providers, and researchers to develop targeted policies and programs to improve the well-being of people with disabilities.

**DATA DESCRIPTION**

The DHDS dataset has 644,356 records and 32 columns. Before the cleaning process, the dataset comprised multiple columns, which include 'Year', 'LocationAbbr', 'LocationDesc', 'Category', 'Indicator', 'Response', 'Data\_Value\_Unit', 'Data\_Value\_Type', 'Data\_Value', 'Low\_Confidence\_Limit', 'High\_Confidence\_Limit', 'Number', 'WeightedNumber', 'StratificationCategory1', 'Stratification1', 'CategoryID', as well as 'IndicatorID'.

Following cleaning of data, the dataset was improved to contain 644,356 items and 17 columns. Missing values were found in multiple columns, notably 'Data\_Value', 'Low\_Confidence\_Limit', 'High\_Confidence\_Limit', 'Number', & 'WeightedNumber'. To solve the, a data imputation approach was used, in which missing values were supplied by indicator means.

A screenshot of a computer

Description automatically generated

*Fig: 1: Summary Statistics: Pre & Post-Data Cleaning and Imputation*

**DATA IMPUTATION: HANDLING MISSING VALUES**

Missing values are ubiquitous in datasets & can have a substantial influence on the accuracy and reliability of results. Missing values were found in several columns of the DHDS dataset, including 'Data\_Value', 'Low\_Confidence\_Limit', 'High\_Confidence\_Limit', 'Number', and 'WeightedNumber'. To resolve the missing values and verify the dataset's comprehensiveness, a systematic technique towards data imputation was used.

Initially, the 'Data\_Value' column was analyzed using the 'Indicator' to find missing data. It was thought that missing values in this field were inadvertent and needed imputation to ensure data integrity. The values that were missing for 'Data\_Value', 'Low\_Confidence\_Limit', 'High\_Confidence\_Limit', 'Number', and 'WeightedNumber' had been then imputed using the average values computed for each indicator.

The data imputation procedure included the following steps:

* Indicator Means Calculation: The groupby function in pandas was used to generate mean values for 'Data\_Value', 'Low\_Confidence\_Limit', 'High\_Confidence\_Limit', 'Number', and 'WeightedNumber'.
* Supplying Missing Value: The values that were missing in each of the columns were filled with the computed mean values for the related indicators. This was accomplished by mapping the average values to the relevant indicators and populating the values that were missing accordingly.

A screenshot of a computer

Description automatically generated

*Fig: 2: Summary Statistics Post-Cleaning and Imputation*

**EXPLORATORY DATA ANALYSIS (EDA)**

DISTRIBUTION OF DATA VALUES

Our investigation began by examining a histogram, which provided a complete picture of the variation of data values. This visualization allows us to determine the dataset's central tendency, dispersion, and skewness. We learned about the fundamental distribution pattern and probable outliers by dividing the information into intervals and indicating the frequency on occurrences inside each one.

A graph of data value

Description automatically generated

*Fig: 3: Histogram of Data Values*

TOP 10 LOCATIONS BY RESPONSE COUNT

Following on to the bar plots, we looked at the top ten locations according to response counts. This image allowed us to identify the geographic locations with the greatest levels of data gathering activities. By categorizing the locations in order of decreasing of response count, we identified areas of interest for additional investigation and resource allocation.

A colorful rectangular object with numbers

Description automatically generated with medium confidence

*Fig: 4: Bar plot showing top 10 locations by response count*

COUNT OF STRATIFICATION1 CATEGORIES

Using a count plot, we examined the frequency distribution of categories inside the 'Stratification1' variable. This visualization gave a thorough breakdown of each category's frequency, revealing the dataset's demographic mix. We obtained a better picture of this dataset's demographic variety by examining how numbers were distributed across different categories.

A graph of a bar graph

Description automatically generated with medium confidence

*Fig: 5: Count of Stratification1 Categories*

RESPONSE PERCENTAGE DISTRIBUTION BY STRATIFICATION1

Using violin plots, we investigated the pattern of distribution of response proportions among the 'Stratification1' categories. This depiction represented the central tendency, dispersion, and shape of answer percentage distributions across each demographic group. By displaying a density of data points throughout the violin plot, we were able to identify trends and variances in response percentages among demographic groupings.

A chart of a number of colored arrows

Description automatically generated with medium confidence

*Fig: 6: Violin plot of Response Percentage by Stratification1*

DISTRIBUTION OF CATEGORIES BY STRATIFICATION1

We used a count plot to evaluate the distribution among categories inside 'Stratification1' among various demographic groupings. This visualization allowed for the comparison during category frequencies inside each demographic category, revealing information on the relative frequency of distinct categories across various demographic groups.

A graph of different colored bars

Description automatically generated

*Fig: 7: Categories distributed by Stratification1*

PROPORTION OF STRATIFICATION1 CATEGORIES

Finally, we used a pie chart to determine the fraction of 'Stratification1' groups. This visualization gave a comprehensive perspective of the demographic makeup by displaying each category's proportionate contribution to the overall population. We learned about the distribution of different demographic groupings in the dataset by presenting proportions as pie slices.

A pie chart with different colored circles

Description automatically generated

*Fig: 8: Pie chart followed by Stratification1 Proportion*

**MODEL CREATION & ANALYSIS**

DECISION TREE CLASSIFIER

The Decision Tree model serves to predict handicap categories based on various characteristics in the dataset. We preprocess the information by encoding categorical parameters with one-hot encoding and dividing it into sets for training and testing. To prevent overfitting, the decision tree has a permitted depth of four. The model's accuracy is determined by a value of 90.43%, having recall, precision, and F1-score values presented for each handicap category.

A diagram of a company

Description automatically generated

*Fig: 9: Decision Tree Classifier Model*

INSIGHT THROUGH DECISION TREE CLASSIFIER:

* The decision tree reliably predicts handicap kinds, with excellent precision and recall in most categories.
* However, it struggles to classify cases of "Independent Living Disability," "Self-care Disability," & "Vision Disability," as seen by poor precision and recall scores.
* The decision tree visualization reveals the hierarchical arrangement of variables that help predict handicap categories.

A chart of disability

Description automatically generated with medium confidence

*Fig: 10: Confusion Matrix on Decision Tree Model*

RANDOM FOREST CLASSIFIER   
  
Next, we use a random forest classification model to increase the predicting accuracy of handicap kinds. Similarly to the decision- tree Classifier, the information being used is preprocessed and divided into sets for training and testing. The Random Forest framework with an optimal depth of 4 yields 92.38% accuracy. The categorization report shows the accuracy, recall, & F1-score for each handicap group.

RANDOM FOREST CLASSIFIER INSIGHTS:

* The randomly generated forest model surpasses the model based on decision trees in terms of overall accuracy, as well as recall and precision for most handicap categories.
* Despite progress, it remains difficult to reliably forecast categories which include "Hearing Disability," "Independent Living Disability," & "Self-care Disability."

A chart with text and numbers

Description automatically generated with medium confidence

*Fig: 11: Confusion Matrix on Random Forest Classifier Model*

SUPPORT VECTOR MACHINES (SVM)   
  
We investigate the usage of Support Vector Machines (SVMs) for classification tasks using the DHDS dataset. SelectKBest is used for feature selection, with the ANOVA F-value serving as the scoring mechanism. The SVM model using a basis function with radial basis (RBF) kernel and supplied hyperparameters achieves 91.67% accuracy. The categorization report includes specific performance indicators for each handicap group.

SVM INSIGHTS

* In comparison to Decision Tree & Random Forest models, SVM performs well, with good accuracy as well as balanced recall and precision across handicap categories.
* However, like with prior models, it suffers with categories like "Mobility Disability" & "Vision Disability," where accuracy and recall are poor.

A chart of disability

Description automatically generated with medium confidence

*Fig: 12: Confusion Matrix on SVM*

LINEAR DISCRIMINATORY ANALYSIS (LDA)   
  
LDA, or linear discriminant analysis, is used to minimize the complexity of the dataset & categorize handicap categories. Each feature is scaled with StandardScaler, while LDA is used to convert the features to a space with fewer dimensions. The model has a success rate of 92.85%, including precision, recall, and F1-scores provided for each handicap category.

LDA FINDINGS

* LDA performs well in categorizing handicap kinds, with good accuracy, balanced precision, and memory across categories.
* LDA's reduced-dimensional representation allows for a better separation of handicap categories, resulting in higher classification performance.

A screenshot of a computer

Description automatically generated

*Fig: 13: Confusion Matrix on LDA*

LOGISTIC REGRESSION   
  
Finally, Logistic Regression is used for binary categorization of handicap categories. Preprocessing of the dataset includes missing value imputation and one-hot categorical variable encoding. The Logistic Regression model yields an accuracy of 91.20%, and the confusion matrix visualization provides information about the model's prediction performance.

A screenshot of a graph

Description automatically generated

*Fig: 14: Confusion Matrix on Logistic Regression*

**RECOMMENDATION FOR SUBSEQUENT ANALYSIS AND IMPLEMENTATION**

* Enhanced Feature Engineering: The performance of LDA shows that the present collection of features utilized in models may be helpful in capturing fundamental patterns of disability and wellness dynamics. However, ongoing work in feature engineering may include including additional pertinent characteristics, which might include indicators of socioeconomic status, healthcare access measures, and environmental elements, to improve the models' prediction powers.
* Model Optimization: Although LDA worked well, additional modification of all models may be useful. This might include fine-tuning hyperparameters, investigating ensemble approaches, and rigorously validating external datasets to ensure resilience and generalizability across various circumstances.
* Integrating Temporary Analysis: Temporal analysis might reveal information on how rates of disability and health outcomes vary over time. Studying temporal trends can assist in identifying patterns, forecasting future trends, and guiding the creation of preventative and intervention methods.
* Ethics Considerations: As the simulations are put into effect to real-world scenarios including health and policy choices, ethical concerns take precedence. To guarantee ethical and responsible data usage, fairness, openness, and responsibility must be prioritized through the analytical process, with concerns regarding bias, privacy, and information protection addressed as needed.
* Collaboration and Stakeholder Engagement: Communicating with stakeholders, such as subject matter experts, lawmakers, healthcare professionals, & community representatives, is critical for validating results, gathering insights, & co-creating solutions. Teamwork guarantees that the results of research are pertinent, actionable, and culturally appropriate, resulting in more successful policies and initiatives.
* Continued Monitoring and Evaluation: Implementing methods for continual monitoring and assessment of handicap policies, initiatives, and interventions is critical. Frequent gathering and analyzing of data aids in determining the success of programs, identifying areas of improvement, and adapting methods to changing requirements and priorities.

Given these suggestions and the business problems raised, LDA (Linear Discriminant Analysis) appears as the best model for determining impairment classes in this situation. Its preciseness, balanced precision, and memory across several disability categories suggest excellent performance in disability classification. Other aspects to examine before finalizing the model selection include processing efficiency, interpretability, and special application needs.

**CONCLUSION**

Our examination of the Disability & Health Data System (DHDS) data provided us with useful insights into the patterns & dynamics of adult disability as well as health. Our models, which included the Decision Tree Classifier, Random Forest Classifier, Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), & Logistic Regression, produced reliable predictions and identified key factors determining handicap classifications.   
  
Overall, our analysis shows that, while some models predict disabilities categories alongside high accuracy and balanced precision and recall, it is difficult to accurately classify certain disability types, which include "Independent Living Disability," "Self-care Disability," along with "Vision Disability." These findings highlight the complexities of disability classification along with the need for more study and development of prediction algorithms.

Despite these obstacles, our study shed light on the links among demographic, health-related, and regional factors and disability categories. These findings can help to shape targeted policies, initiatives, and healthcare services that enhance the well-being and standard of life for persons with disabilities.

**REFERENCE**

The CDC (2024) is the Centers for Disease Control and Prevention. Regarding the Health and Disability Data System (DHDS). taken from the announcement page at [https://www.cdc.gov/dhds.](https://www.cdc.gov/dhds.%20)

The NCBDDD stands for the National Center on Birth Defects and Developmental Disabilities. (2024). Information & Figures Regarding Developmental Disability and Birth Defects. taken from data.html at <https://www.cdc.gov/ncbddd>