Prediction of Service Type Required for Persons Living with Intellectual and Developmental Disabilities

Submitted by: Austin Okafor

Student Number: 501143897

Supervisor Name: Ceni Babaoglu, Ph.D.

Course Name: Data Analytics, Big Data and Predictive Analytics

Submission Date: June 25, 2022

Ryerson University

Table of Contents

Abstract

Home- and Community-Based Services (HCBS) are types of person-centered care delivered in the home and community. A variety of health and human services can be provided. HCBS programs address the needs of people with functional limitations who need assistance with everyday activities at varying levels, like getting dressed or bathing. HCBS are often designed to enable people to stay in their homes, rather than moving to a facility for care.

Children, Community and Social Services of Ontario conducted a survey to gather client profiles who received developmental disability support in 2010 and 2013, in line with global aims to build data capacity for person-centered outcomes research for the population with intellectual and developmental disabilities (ID/DD) through the creation of a publicly accessible, de-identified, linked dataset of ID/DD relevant state-level data. The 2010 survey was completed by agencies providing residential services, while the 2013 survey was completed by agencies providing non-residential services. The dataset includes snapshot information for 33,615 individuals, with information gathered about their age, gender, living arrangement, income source(s), communication, use of disability aids, health and medical conditions, behavioural traits, and level of support needed. Responses from multiple agencies for the same client were consolidated to ensure that only one case existed for each client.

Despite significant investments from various organizations in services and supports for the ID/DD population, data infrastructure issues limit the ability of researchers to conduct person-centered outcomes research, which limits implementation of evidence-based person-centered practices,

programs, and policies to support persons with ID/DD. A lack of person-centered outcomes research also limits the ability of individuals with ID/DD and their family to strategically choose services and supports that promote prioritized outcomes and goals. The dataset that will be used in this research work can be found publicly here: https://data.ontario.ca/dataset/developmental-disability-support-client-profiles.

This project aims to conduct an exploratory analysis using the dataset to evaluate person-level predictors of outcomes prioritized by people with developmental disability and identify opportunities for Home- and Community-Based Services (HCBS) systems-level program improvements. To achieve this, I will investigate relationships between the dependent variable, Services (Service Type), and other independent variables in the dataset, to try to predict the classification of an individual requiring support services and determine if they will require either Home- or Community-based services, or both. This is significant because additional costs will certainly be incurred to provide both Home- and Community-based services.

My hope is that this research will help to establish milestones for advancing knowledge on the prevalence and health status of individuals with intellectual and developmental disabilities, especially around determining the service types required by such individuals and the cost implications of having to provide both service types to one individual, as well as help to analyze relationships between various sociodemographic information, need for home and community-based services.

In this research, I will attempt to answer questions like:

- What variables have the most influence in the prediction of the service type required by persons living with ID/DD?
- Can we predict the type of support that may be required by persons living with ID/DD using available data from the dataset?
- How accurately can we predict if a person living with ID/DD will require either "Home Based or Community based Support", or both?
- Which classification algorithm is best suited for predicting the type of support required by persons living with ID/DD based on the dataset?

Due to the diversity and nature of the variables of the dataset, multiple data preparation, descriptive analysis and machine learning techniques will be adopted for this project. Some of the procedures that will be used include, but not limited to, data preparation techniques, investigating the attribute types, finding any missing values and determining how best to handle them. I will attempt to determine which variables of the dataset are most relevant for this analysis and which to drop when preparing our prediction models. I will also be using exploratory data analysis and data descriptive techniques like DataExplorer and SmartEDA to investigate relationships between the data variables, determine which attributes seem to be correlated, perform feature selection to determine which attributes can be eliminated when creating samples from the dataset and which should be included in the Classification algorithms we will used for this analysis. The algorithms we will be using for this analysis are Decision Tree, Logistics Regression, and k-Nearest Neighbors (KNN). We will the conclude our investigation by comparing the predicted output of the three algorithms based on accuracy, sensitivity, and specificity, and identify the best algorithm.

GitHub Link

The GitHub ink for this project is:

https://github.com/AustinOkfor/CIND820-Project-Work-for-Austin-Okafor

Literature Review, Data Description, and Project Approach

Literature Review

Disability Inclusion Analysis of Lessons Learned and Best Practices of the Government of Canada's Response to the COVID-19 Pandemic.

This project was embarked upon in response to the COVID-19 crisis by the Live Work Well Research Centre and conducted in partnership with the DisAbled Women's Network of Canada (DAWN). The objective of the research was to assist Employment and Social Development Canada in identifying good or best practices in Canada and beyond to address the pandemic and people with disabilities, as well as lessons learned from the response to the COVID-19 pandemic in Canada. The research was conducted to help us better understand how diverse people with disabilities in Canada have been affected by the COVID-19 pandemic and the effects of government COVID-19 measures on diverse people with disabilities in Canada. This research was based on the United Nations Convention on the Rights of Persons with Disabilities (UNCRPD), which calls for the protection and safety of people with disabilities in situations of risk, such as the COVID-19 pandemic. Canada is among the many nations that has ratified the UNCRPD convention.

The research stated that the pandemic disproportionately affected people with disabilities in Canada negatively, due to the policies implemented in response to the pandemic. It was stated that the effects of such policies were profound not only for people with disabilities, but also for their families, and their support and care givers, and yet largely unnoticed by those outside of the

disability community. It was stated that diverse groups of disabled people experienced multiple and intersecting forms of discrimination, as well as barriers that shaped their experiences in the context of the pandemic. The research data used was gathered from sources including government policies, websites, interviews and focus groups with policy makers, community organizations, and disability leaders from BC, Alberta, Ontario, and Quebec.

The data collected was reviewed and analysed using an intersectional disability and gender analysis framework (iDGA), and the context of Canada's human rights commitments. Impacts were identified in 19 different thematic areas, ranging from employment and income to health care and community services, to isolation, social inclusion, and discrimination. These were then narrowed to three broad research findings as: 1). Evidence of consistent exclusion of people with disabilities arising from their invisibility, despite their increased risks. It results from inadequate data, lack of targeted polices, restrictive eligibility criteria, and disability definitions for programs, lack of policy coordination and complementarity, inaccessible communications and information, and actions or inactions which reinforce or exacerbate existing systemic inequities. 2). Experiences of exclusion were often cascading and cumulative. 3). Identify lessons and good practices for building back better post pandemic and to ensure greater disability justice in future disasters.

Some key takeaways from the research was that COVID-19 pandemic exposed pre-existing systemic exclusion of people with disabilities from policy planning and decision-making, and also that Inequalities faced by people with disabilities intensified during the pandemic. Also, it was found that disability leaders and organizations offer important insights for ensuring disability justice in the future. It also stated that governments and policy makers need to take concrete and

comprehensive actions immediately, such as establishing accessibility and inclusion standards related to emergency planning and management and national and disability benefit

Place of Residence and Preventive Health Care for Intellectual and Developmental Disabilities Services Recipients in 20 States

This research was conducted to identify trends in the receipt of preventive health care by adults with intellectual and developmental disabilities (ID/DD) by type of residential setting in the US. Data from the 2008 to 2009 collection round of the National Core Indicators (NCI) program was utilized for this research and included approximately 100 performance and outcome indicators that aim to measure and aid in improving system performance of state developmental disabilities authorities, drawn from a random sample of at least 400 individuals from everyone receiving services in the state. The indicators used were measured by multiple data sources, and the SPSS version 18 was used for performing a series of logistic regressions analysis on the dataset.

The analysis performed was able to verify that adults with intellectual and developmental disabilities living in different residential arrangements differ in terms of personal characteristics, such as age, level of intellectual disability, and mobility, and that some of these personal characteristics affect the likelihood of an individual receiving preventive care. It also observed that age, level of intellectual disability, mobility, and health status seem to particularly influence the receipt of preventive health for peoples with ID/DD. The researchers were also able to confirm their primary hypothesis that even after controlling for disability, a person's living environment influences whether he or she receives standard preventive care procedures. The only exception to this was with pneumonia vaccine, where controlling for disability had the effect of increasing the odds of receiving preventive examinations for people living in more restrictive environments.

The researchers found that the likelihood of a person receiving preventive care procedures was related to their age, level of intellectual disability, mobility, health status, and state. They also found that the type of living arrangement affected whether a person received these health services, even after controlling for state, level of disability, and other personal characteristics. They also found that people with ID/ DD were increasingly more likely to live in their family homes with parents or relatives rather than in traditional group settings, and that they were consistently the least likely to receive preventive health exams and procedures.

The research suggesting that efforts should be strengthened to improve preventive healthcare access for people living at home with family and those living independently, and that efforts should be made to ensure that people transitioning from institution to community-based settings maintain access to care, while also maintaining that it is not an option to return to a more institutional style service delivery system as a means of ensuring preventive healthcare access for people with ID/DD. It posits that instead; more effective supports need to be put in place that ensure that people with ID/DD living in integrated community settings have the choice and ability to receive quality preventive health care in timely fashion

Ethnic Origin and Disability Data Collection In Europe: Measuring Inequality – Combating Discrimination

This report tries to answer the question: “Do people who live with disabilities or come from an ethnic minority background suffer disadvantages on account of data shortages? Is there a way and is there a duty to remedy such disadvantages?”. It begins by describing how European antidiscrimination law cannot be effectively implemented without collecting equality data. It then states that although it is mandatory to collect disability data both at the national and the EU level,

disability categories are often medicalized, and the data sets that can best be used to indicate inequalities and inform policy making are yet to be established. This leads to the lack of proper data and as such statistical facts are not sufficient for determining measures to counter discrimination on the grounds of disability and ethnic origin. Consequently, this deprives people with disabilities, or an ethnic minority background of the tools needed to challenge discrimination. The report observed that although the European Union's Data Protection Directive permits collection of sensitive data provided safeguards are observed, the lack of a proper definition of disability, race and ethnic origin in European law, as well as the lack of legislation or case law resolving real or assumed conflicts between equality data collection needs and data protection duties, constitutes a bottleneck it terms "*equality data paralysis*".

The report alludes that the European Union's Data Protection Directive prohibits any collection of sensitive data pertaining to disability and ethnic origin but permits the collection of equality data in compliance with the exemptions enumerated in Article 8 of Directive 95/46/EC. It believes that most a narrow interpretation of national data protection laws is responsible for most EU Member States refusal to collect disaggregated equality data. They however collect data that reveal disability and ethnic origins, on the basis of third-party identification and proxies. It also concludes that disability and ethnic minority communities are never consulted on their data needs and that public debate on the issue is almost non-existent and affirms that there is no European union legislation that imposes a straightforward obligation on Member States to collect equality data, with the only exception being disability data (Article 31 CRPD).

The report makes various recommendations at both the European level, national level, and country-specific level. It suggests that there should be development of de-medicalized disability and

composite ethnic origin categories that also capture discrimination experiences in surveys, and a uniformity of categories across administrative units. The report also suggests amending or changing categories where required, including the de-medicalization of disability categories. It advises to ensure that children are registered in the relevant categories; and to strive to build trust through the involvement of equality bodies and ombudsmen institutions. It also suggests that at the national level, the data protection authorities need to collaborate with equality bodies

in resolving data collection issues.

Financial and environmental challenges of implementing inclusive education for disadvantage student in Mongolia

This research aimed as analysing the effect of government polices over the years and its impact on the children with special needs, the research was conducted on the Mongolia education system, it found the following, government's policy to incentivise school teacher with a 30% bonus was not yielding the desire result, instead of it creating an inclusive environment for the children to thrive in it led to discrimination and segregation, it created something often seen within the care community for people with disability which is the creation of interest groups solely propelled by financial gains.

The writer concluded that incentivising care givers, in this case the teachers in the special need school did not yield the desired result, the solution therefore lies with the redirection of funding. Funding should be directed towards regular schools ensuring that they have right equipment and personnel to keep the children with disabilities fully integrated and accepted within the school system and this in turn will help to reduce the stigmatization around special needs children.

Bipartisan Solutions to Improve the Availability of Long-term Care

This research looks at the challenges associated with providing long-term services and support (LTSS) in the US, particularly for home and community-based services (HCBS). It states that LTSS refer to a broad range of paid and unpaid medical and nonmedical services for individuals with functional limitations due to age, chronic illness, or disability. It includes assistance with activities for daily living and instrumental activities for daily living. Those requiring LTSS support include children, adults, or seniors with physical, cognitive, developmental, mental, or other chronic health conditions.

It states that the cost of care and the shortage of caregivers relative to need are some challenges facing the provision of care for people who require LTSS and that a combination of public- and private-sector options, and an investment of federal resources is required to improve access to these services. The report recommends an expansion of access to home- and community-based services by making them available to individuals with long-term care needs who are ineligible for Medicaid and the development of a transitional program to support the expansion and development of an integrated delivery models where they are unavailable.

The report also recommends that data should be collected on the disparities in access to HCBS and that recommendations should be made to Congress to address inequities discovered. It advises that a refundable tax credit for caregivers should be established to help with out-of-pocket costs for paid LTSS-related care, and that long-term care insurance needs to be simplified and standardized, so as to achieve an appropriate balance between coverage and affordability, among others.

Taking Away Medicaid for Not Meeting Work Requirements Harms People with Disabilities

This report looks at the effects of allowing states in the US to impose work requirements on adult Medicaid enrollees other than those who are 65 or older, pregnant, or who qualify for Medicaid because they receive disability benefits through the Supplemental Security Income (SSI) program. It states that millions of low-income adults with disabilities and serious illnesses qualify for Medicaid because they receive Supplemental Security Income (SSI), while even more are covered under the Affordable Care Act's (ACA) Medicaid expansion and longstanding coverage for low-income parents. It however states that nearly 5 million non-elderly adult Medicaid enrollees with disabilities do not receive SSI.

It argues that since Medicaid beneficiaries with disabilities or illnesses are far likelier than other beneficiaries to be unemployed, sporadically employed, or to work less than full time, they are more likely to lose coverage. It notes that the resulting loss to Medicaid coverage will be especially harmful to people with disabilities and serious illnesses, who typically need regular care to manage their conditions, resulting in worsening of the health of people with disabilities, further making it impossible for them to work.

This report is important to my research because it highlights how unavailability of predictive models which can help determine the outcome of certain actions can result of harmful policies for people living with ID/DD.

Guidelines on mental health promotive and preventive interventions for adolescents

The research alludes that mental health is the leading cause of disability in young people and this portends a range of high-risk behaviours, including self-harm, tobacco, alcohol and other substance use, risky sexual behaviours and exposure to violence. Also, it states that suicide is one of the three leading causes of death among older adolescents. The guidelines provided in this report provide evidence-informed recommendations based on studies of interventions delivered to 10–19-year-olds, on psychosocial interventions to promote positive mental health and prevent mental disorders among adolescents.

The guidelines were developed in conformance to standard WHO procedures for developing guidelines. Recommendations provided in this work was based on the evidence synthesis and Evidence to Decision frameworks and developed five recommendations for mental health promotive and preventive interventions for adolescents. It recommends that universally delivered psychosocial interventions need to be provided for all adolescents and advises that the interventions should cover social and emotional learning. The recommendation goes further to emphasise that psychosocial interventions should especially be provided for adolescents affected by humanitarian emergencies, for pregnant adolescents and adolescent parents, for adolescents with emotional symptoms, and for adolescents with disruptive/oppositional behaviours.

The relevance of this research is that it shows how using scientific methods of data analytics can result in guidelines that can define how services can be rendered effectively for people living with ID/DD.

Conclusion

These literatures, tie into my research in that they all stress the need for providing, and improving access to services for people living with ID/DD in home- and/or Community based environments. They also stress on the need to improve the level or degree of availability of these services, and the kind of services available. The outcomes of this research to discover what factors can influence the type of service type or level of support required by persons living with ID/DD, and attempt to predict the level of support or the type of service that may be required by persons living with ID/DD, could prove vital for planning for availability of such services, developing and implementing policies for people with ID/DD.

Brief Descriptive Statistics

Data Upload and Cleanup

A revision of the imported data showed a UID column which represented an index column, this was removed also as part of the data cleaning process as this is not useful for analysis. I also examined the number of NA's in each column as this will help determine which columns may need to be dropped due to too many missing data in them. It was noticed that quite a umber of columns have rather large missing values.

From checking the sums of NA's for each column, we can observe that all columns except "Services" have missing data represented by either -88 or -99, depending on why the data was missing. This information is useful in performing further analysis as it would better help understand outliers and data distributions in our charts.

```
#Checking NA count for each column

colSums(is.na(dssData))
```

```
        SERVICES            PR_Q3D           PR_AGE1           PR_QF1A
               0                74               101              1862
        PR_QF1B           PR_QF1C           PR_QG1A           PR_QG1B
            8369             28215              1891              3501
        PR_QG1C           PR_QH1A           PR_QH1B           PR_QI1
           32609              2290             26274              2614
        PR_QJ1            PR_QK1            PR_QL1          PR_QL1A_1
            2591              2408              1949             22776
      PR_QL1A_2         PR_QL1A_3         PR_QL1A_4          PR_QM1A
           22776             22776             22776              3357
        PR_QM1B           PR_QM1C           PR_QN1_A          PR_QN1_B
            4087             10376              3841              4978
        PR_QN1_C          PR_QN1_D          PR_QN1_E          PR_QN1_F
            4849              4957              4151              4638
        PR_QN1_G          PR_QN1_H  PR_QO1_A_COMBINE  PR_QO1_B_COMBINE
            4464              4736              3168              3362
PR_QO1_C_COMBINE  PR_QO1_D_COMBINE  PR_QO1_E_COMBINE             PR_QQ
            3369              3590              3221              2568
    C_BEHAVETHER     C_AUDIOSPEECH      C_COMMCONSUL              C_PT
            4946              5159              5111              5483
            C_OT       C_DIETICIAN    C_PSYCHIATRIST    C_PSYCHOLOGIST
            5453              5526              5781              5927
        C_QE2_2           C_QE2_3           C_QE2_4           C_QE2_5
           15529             31133             29206             28759
           C_QRA            C_AQRA2           C_AQRA3           C_QRA4
           31133             31288             31289             31292
       C_QRB1_1          C_QRB1_2          C_QRB1_3          C_QRB1_4
           20539             24157             32094             29743
       C_QRB1_5          C_QRB1_6            C_QRB2          C_QRB3_1
           26635             32481             18317             30696
       C_QRB3_2          C_QRB3_3            C_QRB4          C_QRC1_1
           31864             29714             26733             29823
       C_QRC1_2          C_QRC1_3          C_QRC1_4            C_QRC2
           30428             31153             32536             28759
         C_QS1             C_QS2_1           C_QS2_2           C_QS2_3
            9256             29752             32173             29912
       C_QS2_4         C_QS2_CODE             C_QS3
           31658             32023             24560
```

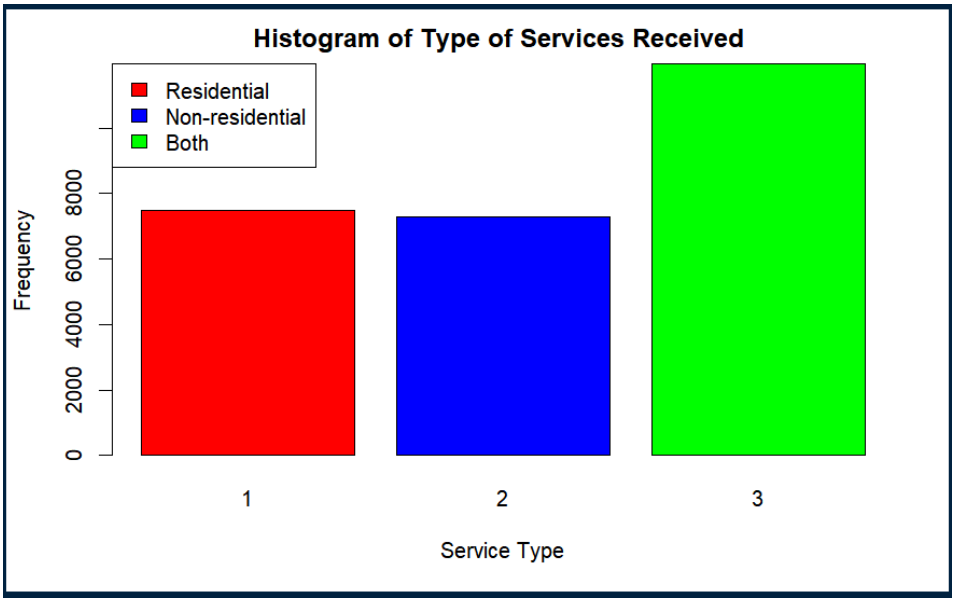
I also checked for the composition of the data, which revealed that there are 33615 observations of 75 variables in the dataset, and that all the categorical variables were stored as integers, and the PR_QQ variable was stored as a character variable. These were fixed using appropriate

codes. Since my dataset was so large, I then checked for the count of NA's for each column and dropped all columns containing 10,000 or more NA's, as well as rows with 5 or more NA's. this resulted in a dataset with 26732 observations of 35 variables.

Descriptive Statistics

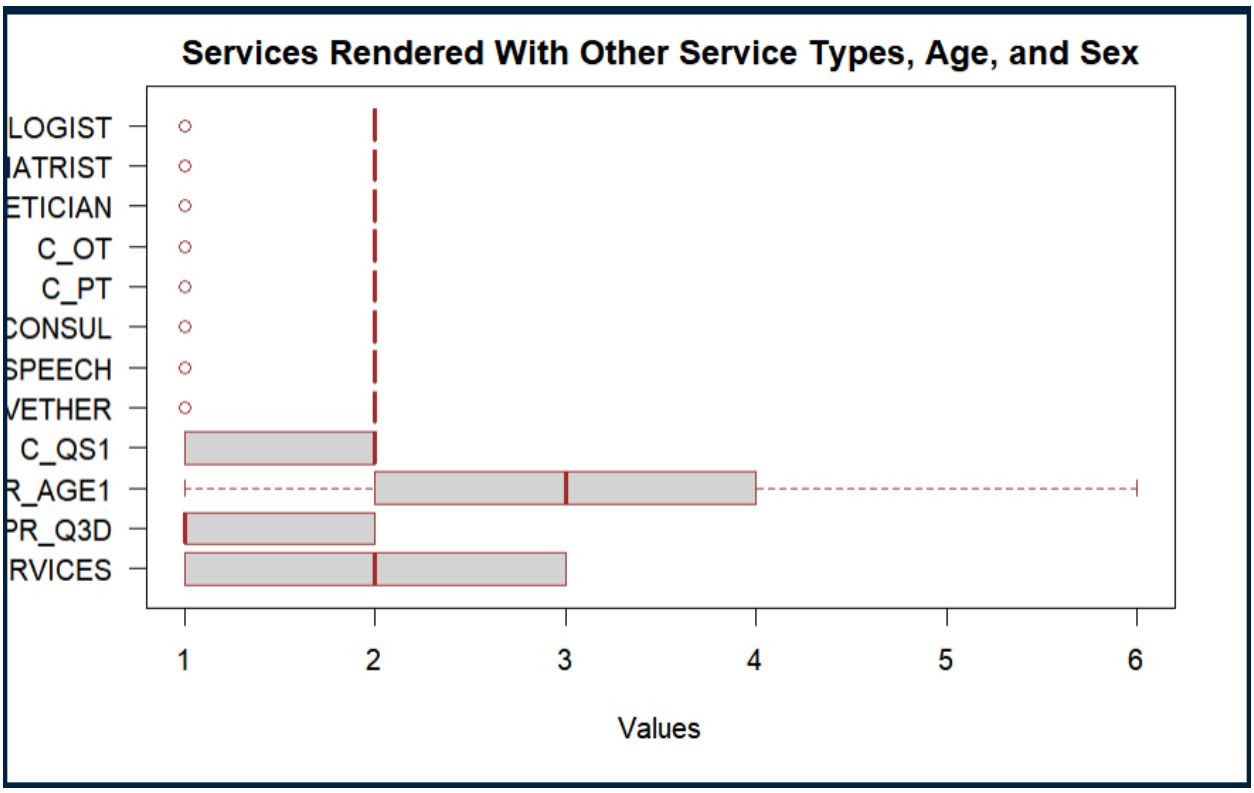
For this phase, we start by first examining a possible relationship between "epilepsy, seizures or convulsions cases" and "acquired brain injuries reported". The result obtained suggested that 373 patients were diagnosed for both Epilepsy, seizures or convulsions and Acquired brain injury. Also, a further 70 patients had both Epilepsy, seizures or convulsions and Acquired brain injury but have not been diagnosed yet. This number is significant because it represents a portion of people who may have suffered brain injury because of not receiving proper care for their Epilepsy, seizures, or convulsion condition. It also represents the number of people who may not have been able to receive required services for their condition.

Next, we plot a Bar chart of service type received,

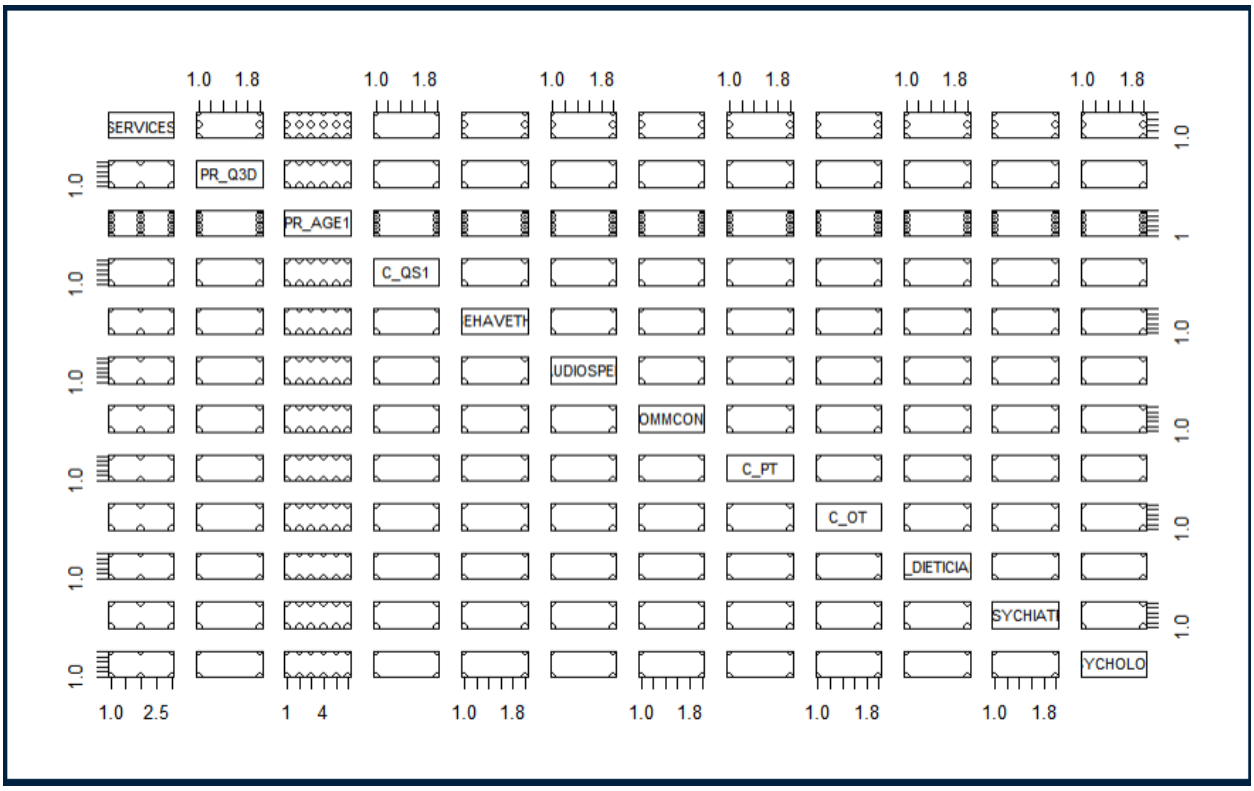


This shows the three service types provided to people living with ID/DD. Our interest in this project is to determine those who require either service type 1 or 2 vs those who require service type 3.

Next we create a boxplot of a subset of the data representing the Service type, age, gender, and all other categories of services provided, and then finally, a plot of this sample to investigate any correlations between them.



Creating a plot of the sample subset, to check for any correlation relationships between the variables for services rendered. Since these are categorical variables, no relationship is expected.



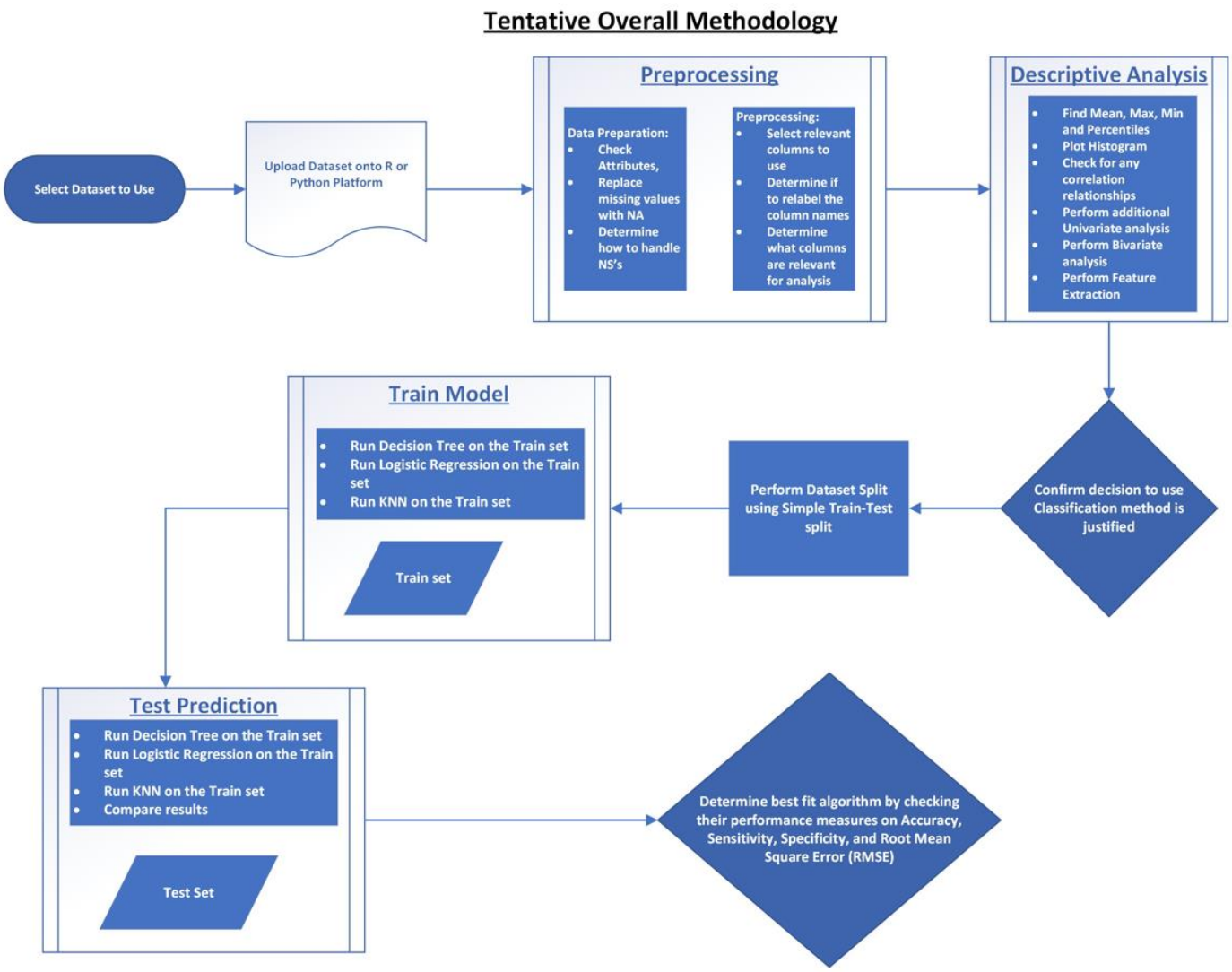
As expected, no significant correlation relationships were found.

GitHub Link

The full work done with results are hosted on the GitHub link below.

https://github.com/AustinOkfor/CIND820-Project-Work-for-Austin-Okafor

Tentative Overall Methodology



Initial Results and the Code

Exploratory Data Analysis (EDA)

For his phase of the project, we will utilize some R packages for Automated Exploratory Data Analysis. This is especially useful because of the size of the data we are working with, the complexity of the variables, and the number of columns in the dataframe. We start by installing Tidyverse, DataExplorer, and SmartEDA. These will be used for our EDA work. Then, leveraging our previous knowledge of the data from the Descriptive Statistics analysis conducted in our last work, we will load the dataset and assign column types automatically. We will however not change -88 and -99 to NA's even thought these represent unknown and missing values respectively. We, however, capture each variable based on the variable types they should be represented as. Also, there was no need to convert any values to NA's since the package we will be using will have no problem handling the data as is. That aside, I believe that changing the original values may distort the true understanding of how the data is distributed, because certain levels will be lost, since majority of variables are factor variables.

As a side note, I should mention that due to the size of the data and number of columns (33,615 rows and 76 columns), it was very tedious trying to obtain and code each column type, while ensuring no error was made during assignment. I actually had to redo this step several times to ensure accuracy.

After reading the file, I removed the UID column as this is not useful for any analysis, and then ran "str" and "summary" to understand the distribution of the data better and confirm the variable types were properly captured during upload. Irregular data was then rectified.

Now we move on to more advanced EDA tools by creating two separate EDA reports using DataExplorer and SmartEDA using appropriate codes. This produced as extensive diagrammatic analysis of the data, plotting bar chart, correlation, scatter plots, and many more. The HTML output of these codes can be viewed on the GitHub link provided earlier. The HTML output displays among other: Basic Statistics, Raw Counts, Percentages, Data Structure, Missing Data, Profile, Univariate Distribution, Histogram, Bar Chart (with frequency), Bar Chart (by SERVICES), QQ Plot (by SERVICES), Correlation Analysis, Principal Component Analysis, Bivariate Distribution, Boxplot (by SERVICES), and Scatterplot (by SERVICES). It also shows that 33,615 rows, 75 Columns, 73 Discrete columns, and 2 Continuous columns were analyzed.

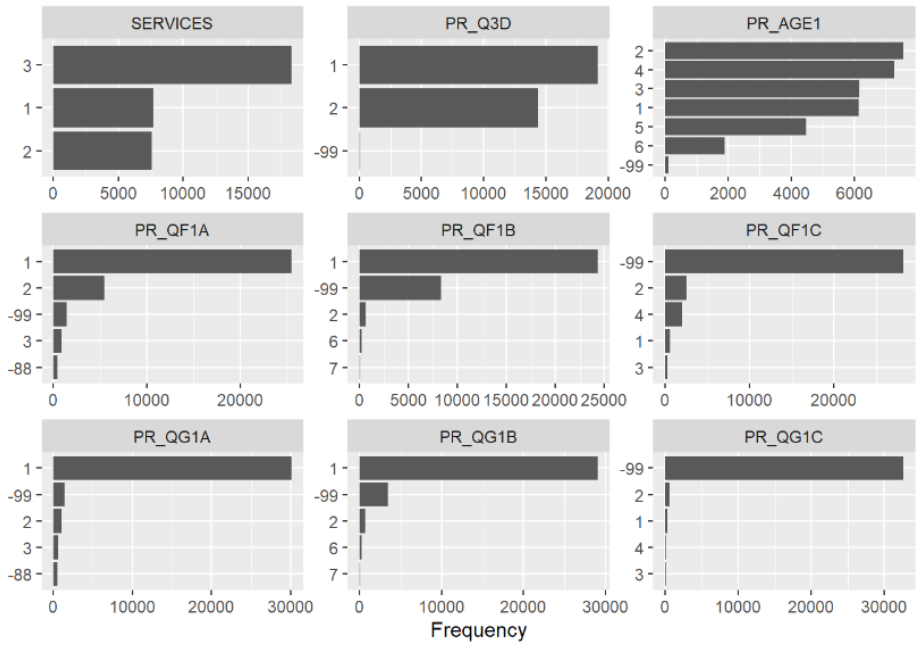
The correlation analysis also shows that there are some correlations between variables but due to the number of variables and the volume of the data, this was difficult to read, and this further highlights the extent of work I had to perform just for data cleanup and preparation. The principal component analysis sections display feature importance, and this is useful for feature selection when creating out samples. A snippet of the data description is provided below:

Descriptions <chr>	Value <chr>
Sample size (nrow)	33615
No. of variables (ncol)	75
No. of numeric/interger variables	2
No. of factor variables	44
No. of text variables	29
No. of logical variables	0
No. of identifier variables	0
No. of date variables	0
No. of zero variance variables (uniform)	0
%. of variables having complete cases	100% (75)
%. of variables having >0% and <50% missing cases	0% (0)
%. of variables having >=50% and <90% missing cases	0% (0)
%. of variables having >=90% missing cases	0% (0)

13 rows

```
# view bar charts using DataExplorer

plot_bar(dataframe)
```



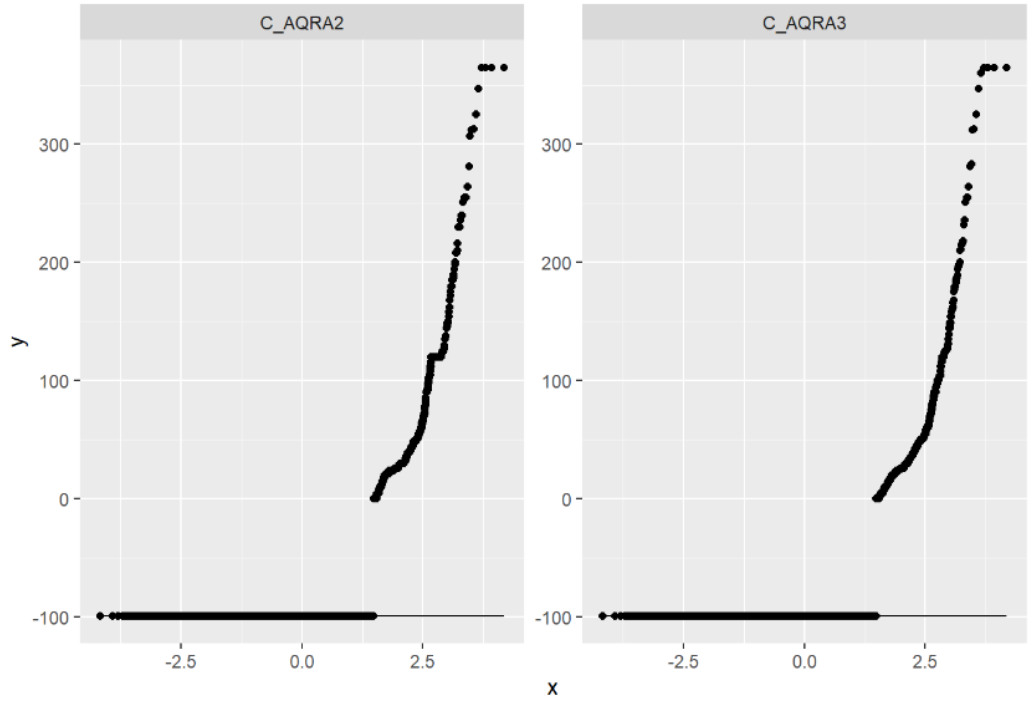
Page 1

```
# view bar charts with Services using DataExplorer

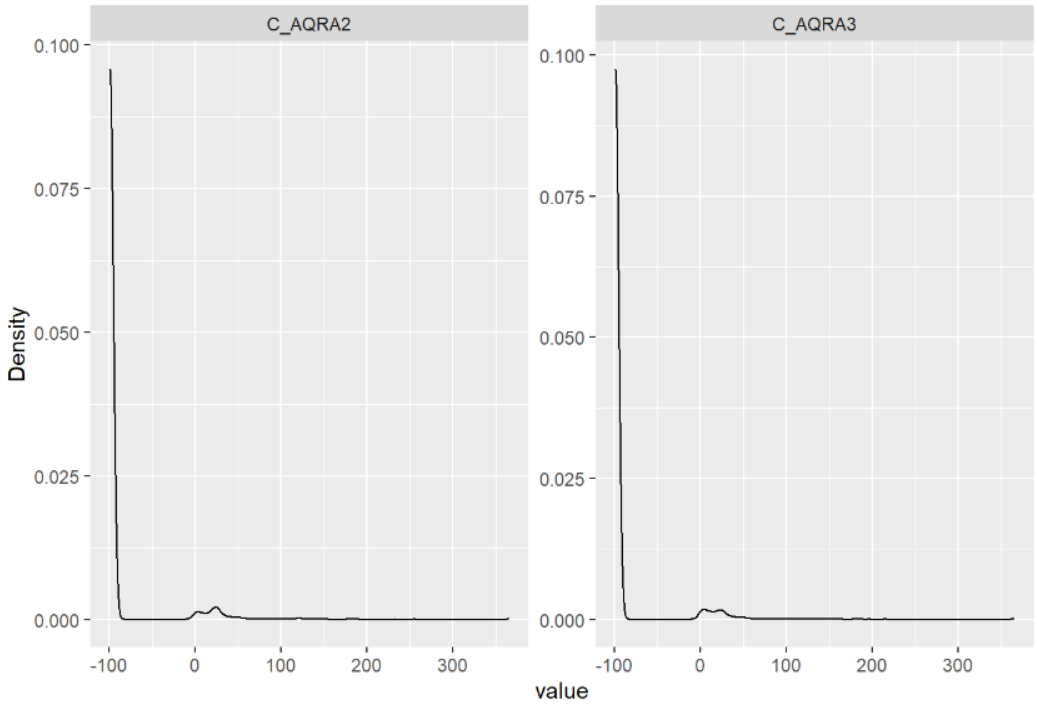
plot_bar(dataframe, by="SERVICES")
```

Page 1

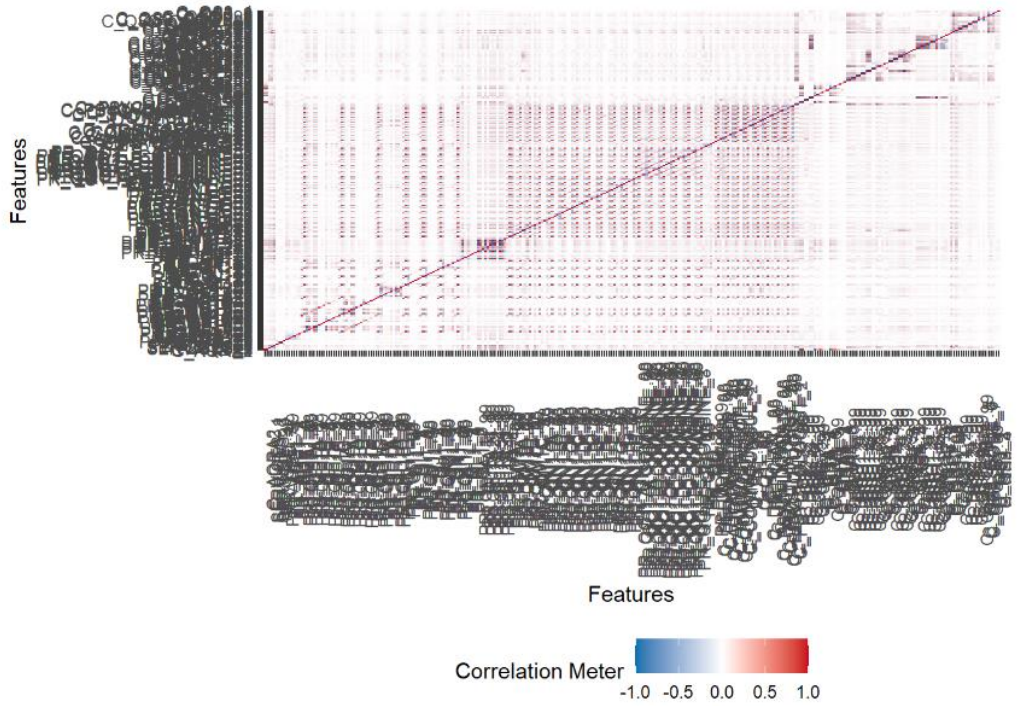
```
plot_qq(dataframe)
```



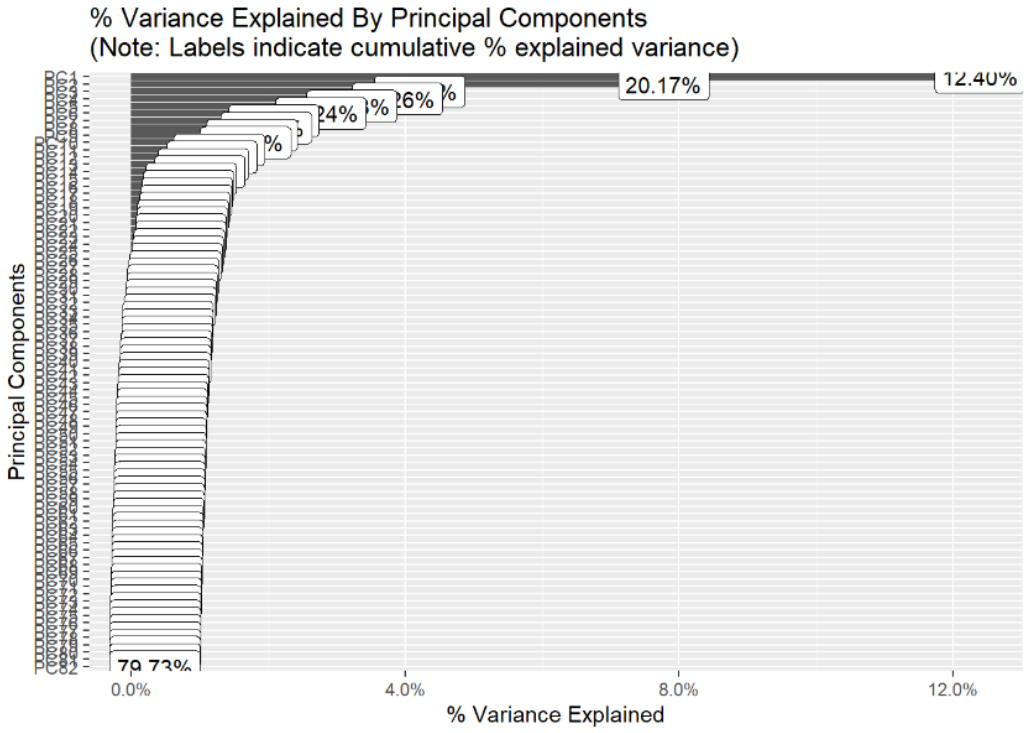

```
plot_density(dataframe)
```

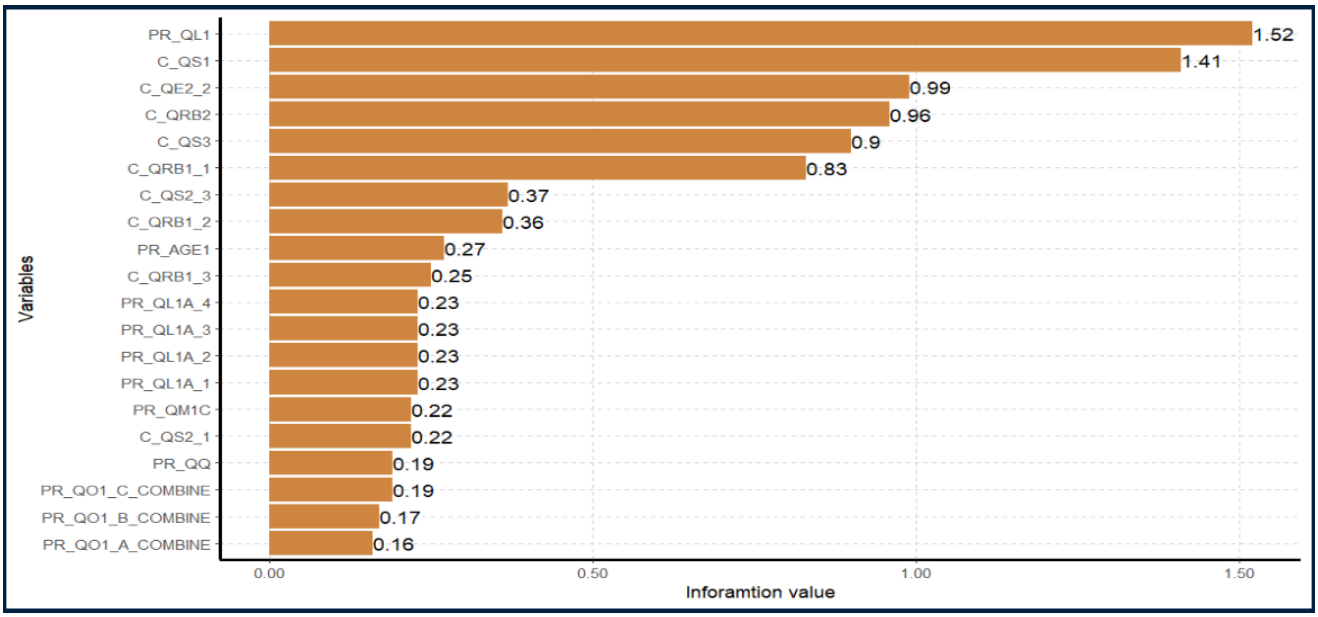


```
plot_correlation(dataframe)
```




```
plot_prcomp(dataframe)
```





From the SmartEDA and DataExplorer report, I reviewed the variable importance based on information value, and this was particularly useful for feature selection, as it displayed the various variables based on their predictive strength and relevance. It also helped answer the first question for this project: "**What variables have the most influence in the prediction of the service type required by persons living with ID/DD?"**

Referencing the outputs of SmartEDA and DataExplorer for feature selection, I then created a subset of the original data by reducing the number of columns from 75 to 19, using only the columns of high and medium importance. This will then be used for the remaining stages of our analysis, and to answer the remaining project questions.

After Creating a Subset dataframe using columns with High and Median importance features, and adding labels for variables in services, I prepared the data for modeling by splitting the dataset into training and test sets, checking dimension of train and test set to confirm accuracy of split.

```
# check dimension of train and test set to confirm accuracy of split
dim(subset_train)

dim(subset_test)
```

```
[1] 26892    19
[1] 6723   19
```

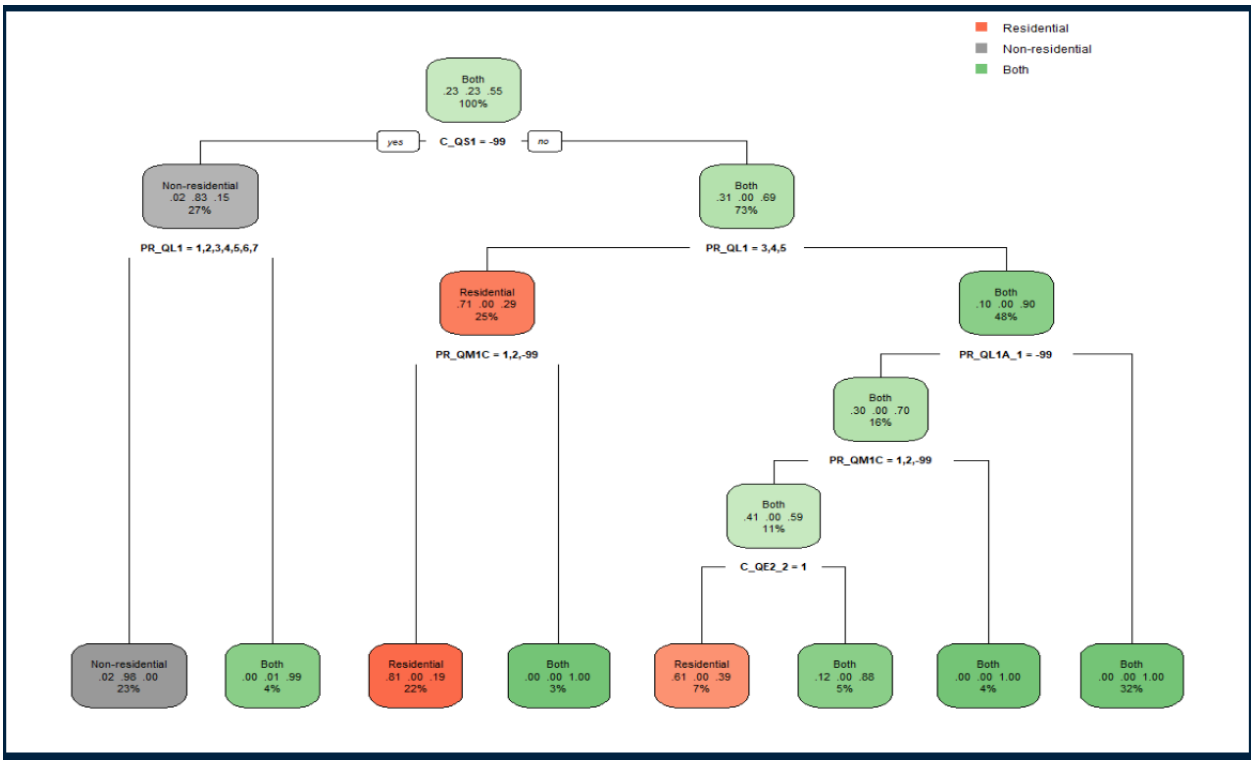
We will now attempt to answer he second question of this project: **"Can we predict the type of support that may be required by persons living with ID/DD using available data from the dataset?"**

Since our dependent variable, Services, has more than 2 possible outcomes, a multi-class classification algorithm is required for the first modeling operation and, since we are performing Classification, the Decision Tree Model was used.

We used rpart.plot for the Decision Tree classification procedure to build the model using type = class. As this is a preliminary analysis, no further investigation was performed. I however tested the model by using the output of the classification to build a model.

This result confirms that we can predict the type of support that may be required by persons living with ID/DD using available data from the dataset, and thus answers the second question in our project.

Below are the outputs obtained:



	predict_services		
	Residential	Non-residential	Both
Residential	1462	38	38
Non-residential	0	1513	4
Both	488	5	3175

Final Results and Project Report

Final Results and Project Report

In this final phase of this project work, we shall attempt to answer the remaining questions in this project by running classification algorithms for Decision Tree, Logistic Regression and k-Nearest Neighbors by leveraging on the results of the EDA analysis performed in the previous section for performing our analysis. We will also be grouping the Services column data based on those who require **either Home-based or Community based services**, vs those who require **both Home-based and Community based services**. This will result in a column with binary data.

After loading the dataset based on previous knowledge, we then proceed to create a subset of the data using only the columns of high and medium importance as we did for the EDA phase, and as before, we do not replace the missing values with NA.

Data Preparation

First we create a new column called “Either_Or_Both”, where we represent service requirements for either Home-based or Community-based with “0” an service requirement for both Home-based and Community-based with “1” and assign this to the dataframe. Then we drop the “Services” column.

Then we prepare data for modeling by splitting the dataset into training and test sets, ensuring to set seed so the data is reproducible. Next we remove the class column, Either_Or_Both, from our training and test datasets.

Using the Logistics Regression algorithm

After creating our logistic regression model, running our code and analysing the output, we can notice that not all variables are significant for prediction, C_QRB2, C_QS1, C_QRB1_3, PR_QM1C, and PR_QO1_C_COMBINE are highly significant, while C_QS2_3 and C_QS3 and less significant in that order. We will however not be investigating this further as it does not help in answering the questions of this research.

```
Call:
glm(formula = Either_Or_Both ~ ., family = "binomial", data = newDF_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.3839  -0.0605   0.0000   0.0001   3.7712

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.937e+01  3.673e+02   0.053 0.957956
PR_QL1         -4.912e-02  3.121e-03 -15.739  < 2e-16 ***
C_QE2_2        -1.838e-02  9.402e-04 -19.552  < 2e-16 ***
C_QRB1_1       -2.594e-03  1.025e-03  -2.530 0.011403 *
C_QRB1_2       -1.200e-03  6.253e-04  -1.919 0.054992 .
C_QS2_3        -2.382e-03  8.010e-04  -2.973 0.002946 **
C_QRB2          2.479e-03  1.298e-03   1.910 0.056155 .
C_QS1           7.200e-02  2.042e-03  35.263  < 2e-16 ***
C_QS3          -1.198e-03  7.740e-04  -1.548 0.121528
C_QRB1_3       -3.517e-03  9.457e-04  -3.718 0.000201 ***
C_QS2_1        -8.850e-06  8.032e-04  -0.011 0.991209
PR_AGE1        -5.234e-03  9.294e-03  -0.563 0.573326
PR_QL1A_1       2.128e-02  8.868e+02   0.000 0.999981
PR_QL1A_2      -1.358e-01  4.117e+02   0.000 0.999737
PR_QL1A_3       1.021e+00  1.060e+03   0.001 0.999231
PR_QL1A_4      -6.941e-01  3.029e+02  -0.002 0.998172
PR_QM1C        -3.826e-02  1.080e-03 -35.439  < 2e-16 ***
PR_QO1_C_COMBINE -1.545e-02  1.928e-03  -8.016 1.09e-15 ***
PR_QQ           4.966e-06  1.580e-05   0.314 0.753223
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 37049  on 26891  degrees of freedom
Residual deviance: 10760  on 26873  degrees of freedom
AIC: 10798

Number of Fisher Scoring iterations: 19
```

We shall then produce the confusion matrix and statistics for the prediction performed using this model for further review. We notice from the output that the Accuracy level is quite high at 91.4%, Sensitivity comes in at 0.8802, and Specificity is at 0.9462. Since our data doesn't necessarily place much priority on sensitivity or specificity, this is a good result as both tend towards 1.

```
Confusion Matrix and Statistics

          Predicted
Actual    0    1
     0 2859  187
     1  389 3288

               Accuracy : 0.9143
                 95% CI : (0.9074, 0.9209)
    No Information Rate : 0.5169
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8281

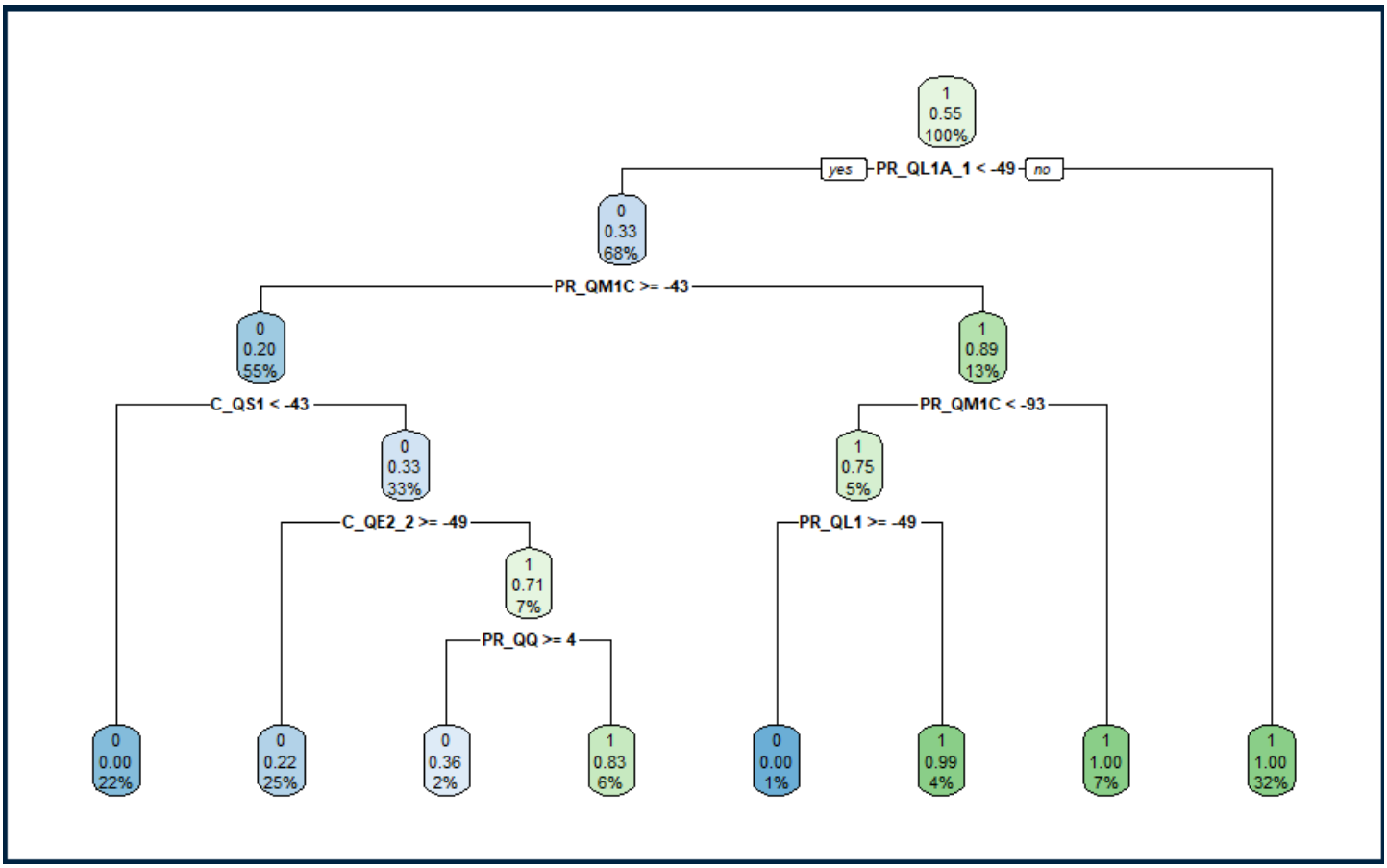
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8802
            Specificity : 0.9462
         Pos Pred Value : 0.9386
         Neg Pred Value : 0.8942
             Prevalence : 0.4831
         Detection Rate : 0.4253
   Detection Prevalence : 0.4531
      Balanced Accuracy : 0.9132

       'Positive' Class : 0
```

Next, we check efficiency by measuring the time to train and test the model. This reveals that it takes 0.67 seconds to generate a model, 0.01 second to predict an outcome, and a total time of 0.7 seconds to train the model and predict an outcome using a logistic regression algorithm

Efficiency of the Logistic Regression model

```
```{r Compute LM Runtimes}

tic("Total_glm")
tic("glm_train")
glm_model <- glm(Either_Or_Both ~.,newDF_train, family = "binomial")
toc()
tic("glm_test")
predict_glm_model <-predict(glm_model, newDF_test, type = "response")
toc()
toc()

```
```

```
glm_train: 0.67 sec elapsed
glm_test: 0.01 sec elapsed
Total_glm: 0.7 sec elapsed
```

Using the Decision Tree algorithm

Since we previously generated the decision tree algorithm using the original values in the "Services" column (1,2,3) during the EDA stage, we will simply rerun the codes used for that generation on the new dataset we created for this phase, where we grouped the "Services" column now only contains "0" and "1" as values. This yields below decision tree.



We shall then produce the confusion matrix and statistics for the prediction performed using the new model and review the output. We notice from the output that the Accuracy level is quite high at 92.5%, Sensitivity comes in at 0.8717, and Specificity is at 0.9806. Since our data doesn't necessarily place any priority on sensitivity or specificity, this is also a good result as both tend towards 1.

```
Confusion Matrix and Statistics

          Predicted
Actual       0    1
     0    2982   64
     1     439 3238

               Accuracy : 0.9252
                 95% CI : (0.9186, 0.9314)
    No Information Rate : 0.5089
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8506

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8717
            Specificity : 0.9806
         Pos Pred Value : 0.9790
         Neg Pred Value : 0.8806
             Prevalence : 0.5089
         Detection Rate : 0.4436
   Detection Prevalence : 0.4531
      Balanced Accuracy : 0.9261

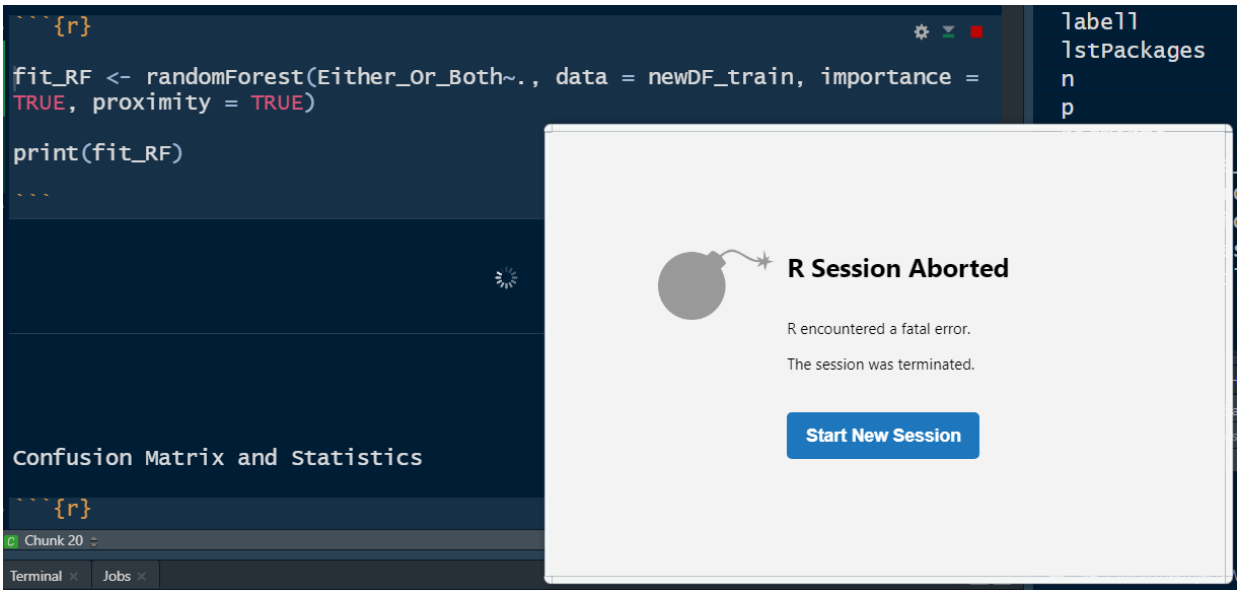
       'Positive' Class : 0
```

Next, we check efficiency by measuring the time to train and test the model. This reveals that it takes 0.47 seconds to generate a model, 0.01 second to predict an outcome, and a total time of 0.48 seconds to train the model and predict an outcome using a decision tree algorithm.

Efficiency of the Decision Tree model

```
```{r Compute DT Runtimes}

tic("Total_dTree")
tic("dTree_train")
fit2 <- rpart(Either_Or_Both~., data = newDF_train, method = 'class')
toc()
tic("dTree_test")
predict_class_model <-predict(fit2, newDF_test, type = 'class')
toc()
toc()

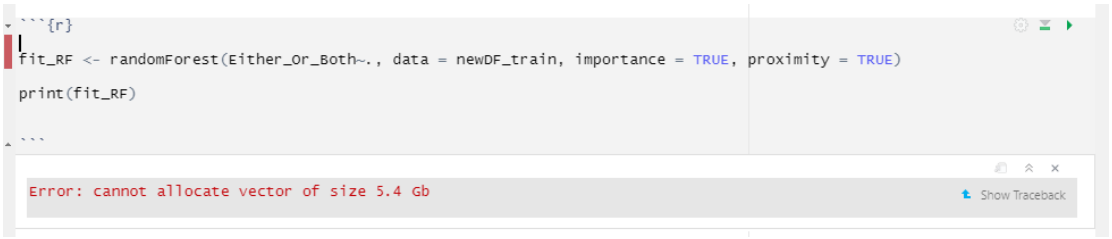

```
```

```
dTree_train: 0.47 sec elapsed
dTree_test: 0.01 sec elapsed
Total_dTree: 0.48 sec elapsed
```

Next, let us attempt to generate the decision tree using Random Forest method, as this provides a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of overcoming over-fitting problem of individual decision tree. I installed and loaded the required package (randomForest) and tried running appropriate codes but encountered a resource problem. It aborted on my desktop and never ran, but on the azure site, it returned an error.



Please note that since all attempts to generate a Classification type model failed, displaying "Error: Cannot allocate vector of size 5.4Gb", before terminating the process, we will not be exploring the Random Forest procedure any further.



Using the k-Nearest Neighbors (KNN) algorithm

First, we create a normalization function for normalizing the data, to eliminate a biased outcome. Then we remove the dependent variable from the dataset and normalize the data by applying the normalization function. We then split the normalized data set into training and testing data set and create the train and test labels for comparing results. Next, we calculate the square root of the number of observations in the training data set to determine the optimal value of 'k' to use for the model. This yields 164 as optimal value for k.

```r
round(sqrt(NROW(DF.n_train_labels)))
```

```
[1] 164
```

Finally, we generate the model, and use the model to make a prediction.

```r
DF.n_test_pred <- knn(train = DF.norm_train[,2:19],
test = DF.norm_test[,2:19],
cl = DF.norm_train[,1], k=164)

head(DF.n_test_pred)
```

```
[1] 1 1 1 1 1 1
Levels: 0 1
```

We then produce the confusion matrix and statistics for the prediction performed using the model for review. We notice from the output that the Accuracy level is again high at 91.2%, Sensitivity comes in at 86.2%, and Specificity is at 96.4%. Since our data doesn't necessarily place any priority on sensitivity or specificity, this is equally a good result as both tend towards 1.

```
Confusion Matrix and Statistics

      Predicted
Actual    0    1
     0 2927  118
     1  468 3210

               Accuracy : 0.9128
                 95% CI : (0.9058, 0.9195)
    No Information Rate : 0.505
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8258

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8622
            Specificity : 0.9645
         Pos Pred Value : 0.9612
         Neg Pred Value : 0.8728
             Prevalence : 0.5050
         Detection Rate : 0.4354
   Detection Prevalence : 0.4529
      Balanced Accuracy : 0.9133

       'Positive' Class : 0
```

Next, we check efficiency by measuring the time to train and test the model. This reveals that it takes 4.3 seconds to generate a model, 0.01 second to predict an outcome, and a total time of 4.31 seconds to train the model and predict an outcome using a decision tree algorithm.

Efficiecy of the k-Nearest Neighbors model

```{r Compute KNN Runtimes}

tic("Total_KNN")
tic("KNN_train")
DF.n_test_pred <- knn(train = DF.norm_train[,2:19], test =
DF.norm_test[,2:19], cl = DF.norm_train[,1], k=164)
toc()
tic("KNN_test")
table_KNN <- table(Actual=DF.n_test_labels, Predicted=DF.n_test_pred)
toc()
toc()

```

```
KNN_train: 3.5 sec elapsed
KNN_test: 0.01 sec elapsed
Total_KNN: 3.51 sec elapsed
```

Performing Cross-Validation on the Three Procedures

We will now compare the performance of the three algorithms using the repeated k-fold cross-validation method. To achieve this, we start by defining a train control model with K = 10 with 5 repeats, then we apply the train control to each of the models. From the outputs, we will extract the root mean squared error (RMSE), which measures the average difference between the predictions made by the model and the actual observations, R-squared, which is a measure of the correlation between the predictions made by the model and the actual observations, and the mean absolute error (MAE), which is the average absolute difference between the predictions made by the model and the actual observations. These values are of interest because the lower the RMSE, the more closely a model can predict the actual observations, while the higher the R-squared, the more closely a model can predict the actual observations and the lower the MAE, the more closely a model can predict the actual observations.

Repeated K-fold cross-validation - Decision Tree Model output:

```
CART

26892 samples
   18 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 24203, 24202, 24203, 24203, 24203, 24203, ...
Resampling results across tuning parameters:

  cp          RMSE       Rsquared   MAE
  0.05881095  0.3045250  0.6248430  0.1854077
  0.19959693  0.3534965  0.4913519  0.2505833
  0.39819685  0.4502927  0.3871880  0.4105938

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was cp = 0.05881095.
```

From the output, the optimal model has RMSE value as 3045250, an R-squared of 0. 0.6248430, and an MAE of 0.1854077'

Repeated K-fold cross-validation - Linear Regression Model output:

```
Linear Regression

26892 samples
   18 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 24202, 24203, 24202, 24203, 24203, 24203, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.2872896  0.6669755  0.2094923

Tuning parameter 'intercept' was held constant at a value of TRUE
```

This model yields an RMSE value of 0.2872896, R-squared of 0.6669755, and MAE of 0.2094923.

Repeated K-fold cross-validation - Decision Tree Model output:

```
k-Nearest Neighbors

26892 samples
   18 predictor

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 26892, 26892, 26892, 26892, 26892, 26892, ...
Resampling results across tuning parameters:

  k  RMSE       Rsquared   MAE
  5  0.2494504  0.7524600  0.09755850
  7  0.2450836  0.7599921  0.09855168
  9  0.2428542  0.7637671  0.09968792

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was k = 9.
```

Here the RMSE value is 0.2428542, R-squared is 0.6669755, and MAE is 0.2094923.

Comparing Accuracy, Sensitivity, Specificity, R-squared, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE)

```
GLM -   Accuracy  : 0.9143
      Sensitivity : 0.8802
      Specificity : 0.9462
          RMSE    : 0.2872896
        Rsquared  : 0.6669755
            MAE   : 0.2094923


DTree -  Accuracy : 0.9252
      Sensitivity : 0.8717
      Specificity : 0.9806
          RMSE    : 0.3045250
        Rsquared  : 0.6248430
            MAE   : 0.1854077


KNN -    Accuracy : 0.9128
      Sensitivity : 0.8622
      Specificity : 0.9645
            RMSE  : 0.2428542
        Rsquared  : 0.7637671
            MAE   : 0.09968792
```

We notice from above display that we are consistently **able to predict if a person living with ID/DD will require either (Home-based or Community-based Support), or both (Home-based and Community based Support) to least 91.28% Accuracy**. This result provides the answer to our third research question: "How accurately can we predict if a person living with ID/DD will require either "Home Based or Community based Support", or both?"

We can also observe that Decision Tree produces the highest accuracy, but also the largest RMSE and he lowest R-squared value. The KNN model on the other hand has a slightly lower Accuracy than the Decision Tree and the Linear regression models, however, it out-performs both by yielding the lowest root mean square error, the highest R-squared value, as well as the lowest mean absolute error. Therefore**, the KNN classification algorithm is best suited for predicting the type of support required by persons living with ID/DD based on the dataset**, and this answers our final research question "Which classification algorithm is best suited for predicting the type of support required by persons living with ID/DD based on the dataset?"

Conclusion and Summary

Although the KNN algorithm took by far the longest time to run at 3.51 seconds in total, compared to 0.48 seconds for decision tree and 0.7 seconds for linear egression, it's performance on it's Accuracy, Sensitivity, Specificity, RMSE, MAE and R-Square scores make it the preferred model, as it closely matches the other models in Accuracy, Sensitivity and Specificity but outperforms the other algorithms in RMSE, MAE and R-Square scores.

With regards to related research work and opportunities for further research, the data used for this research only provides data for already identified persons living with ID/DD. There wee no records for persons who do not require any services. This made it impossible to predict who may require ID/DD services or who might not. The data used, was also as raw as they come, and I could not find any analysis work performed on the dataset. This meant that I had to spend a lot of time on data preparation ad processing.

This research was also unable to investigate the Decision Tree algorithm further due to a resource issue which prevented me from running the Random Forest algorithm for classification, so as to get more accurate predictions. A successful run may have changed the final outcome, if we were able to discover a prediction with better cross validation results. I was only able to run the algorithm for the regression model, which was not useful for my work.

Also, due to the size of the data, a lot of time was spent trying to clean and prepare he data for analysis. This took away from time that could have been used to perform further investigations on each of the models, like rerunning the models using the most significant variables, and comparing outcome and performance, running the Naïve Bayes algorithm and comparing it's performance against the others, and many more options I would have loved to investigate.

These can form the basis for further research on this work since the data preparation portion of the work has been completed by me.

References

Data source: https://data.ontario.ca/dataset/developmental-disability-support-client-profiles

Ontario Government. (2016, June 28). *Developmental disability support: Client Profiles - Ontario Data Catalogue*. Developmental disability support: client profiles - Datasets - Ontario Data Catalogue. Retrieved May 13, 2022, from https://data.ontario.ca/dataset/developmental-disability-support-client-profiles

Assistant Secretary for Planning and Evaluation. (2021, June 1). *Dataset on intellectual and developmental disabilities: Linking data to enhance person-centered outcomes research*. ASPE. Retrieved May 13, 2022, from https://aspe.hhs.gov/dataset-intellectual-developmental-disabilities-linking-data-enhance-person-centered-outcomes

Centers for Medicare and Medicaid Services. (2021, January 12). *Home- and community-based services*. CMS. Retrieved May 13, 2022, from https://www.cms.gov/Outreach-and-Education/American-Indian-Alaska-Native/AIAN/LTSS-TA-Center/info/hcbs#:~:text=Home%2D%20and%20Community%2DBased%20Services%20(HCBS)%20are%20types,like%20getting%20dressed%20or%20bathing

Stienstra, D., Grand'Maison, V., Pin, L., Rodenburg, E., Garwood, K., & Reinders, K. (2021). *Disability inclusion analysis of lessons learned and best practices of the Government of Canada's response to the COVID-19 pandemic*. Live Work Well Research Centre. Retrieved June 9, 2022, from https://liveworkwell.ca/disability-inclusion-analysis-covid-19

Bershadsky, J., Taub, S., Engler, J., Moseley, C. R., Lakin, K. C., Stancliffe, R. J., Larson, S., Ticha, R., Bailey, C., & Bradley, V. (2012). *Place of Residence and Preventive Health Care for Intellectual and Developmental Disabilities Services Recipients in 20 States*. Public Health Reports (1974-), 127(5), 475–485. http://www.jstor.org/stable/41639541

Chopin, I., Farkas, L., & Germaine, C. (2014). Conclusions and recommendations. In C. Hermanin & A. Atanasova (Eds.), *Ethnic Origin and Disability Data Collection In Europe: Measuring Inequality – Combating Discrimination (pp. 64–78).* Open Society Foundations. http://www.jstor.org/stable/resrep27134.9

Otgonlkhagva, S. (2012). *Financial and Environmental Challenges of Implementing Inclusive Education for Disabled Children in Mongolia. In D. Pop (Ed.), Education Policy and Equal Education Opportunities (pp. 287–301).* Open Society Foundations. http://www.jstor.org/stable/resrep27130.16

Bipartisan Policy Center. (2021). *Bipartisan Solutions to Improve the Availability of Long-term Care*. Bipartisan Policy Center. http://www.jstor.org/stable/resrep35530

Center on Budget and Policy Priorities. (2020). *Taking Away Medicaid for Not Meeting Work Requirements Harms People With Disabilities*. Center on Budget and Policy Priorities. http://www.jstor.org/stable/resrep23747

World Health Organization. (2020). *Guidelines on mental health promotive and preventive interventions for adolescents*. World Health Organization. ttp://www.jstor.org/stable/resrep27861

Dhruv. (2020, June 5). *Random Forest approach in R programming*. GeeksforGeeks. Retrieved July 15, 2022, from https://www.geeksforgeeks.org/random-forest-approach-in-r-programming/

Bhalla, D. (2014). *A complete guide to Random Forest in R*. ListenData. Retrieved July 15, 2022, from https://www.listendata.com/2014/11/random-forest-with-r.html

Classification using random forest in R. Blog about iOS, Swift, Android, Kotlin and Python. (2017, January 24). Retrieved July 15, 2022, from https://en.proft.me/2017/01/24/classification-using-random-forest-r/

Dhruv. (2021, July 13). *Naive Bayes classifier in R programming*. GeeksforGeeks. Retrieved July 14, 2022, from https://www.geeksforgeeks.org/naive-bayes-classifier-in-r-programming/#:~:text=Naive%20Bayes%20is%20a%20Supervised,between%20the%20features%20or%20variables.

Finnstats. (2021, April 9). *Naive Bayes classification in R: R-bloggers*. R. Retrieved July 15, 2022, from https://www.r-bloggers.com/2021/04/naive-bayes-classification-in-r/

Lateef, Z. (2020, May 26). *A step by Step Guide to implement naive bayes in R*. Edureka. Retrieved July 13, 2022, from https://www.edureka.co/blog/naive-bayes-in-r/

Lateef, Z. (2020, May 27). *Comprehensive guide to logistic regression in R*. Edureka. Retrieved July 14, 2022, from https://www.edureka.co/blog/logistic-regression-in-r/

Lateef, Z. (2022, March 29). *A complete guide on KNN algorithm in R with examples*. Edureka. Retrieved July 14, 2022, from https://www.edureka.co/blog/knn-algorithm-in-r/

Lateef, Z. L. (2020, November 25). *Decision tree algorithm tutorial with example in R*. Edureka. Retrieved July 14, 2022, from https://www.edureka.co/blog/decision-tree-algorithm/

Morgun, I. (2017, January 24). *Classification using random forest in R*. Blog about iOS, Swift, Android, Kotlin and Python. Retrieved July 21, 2022, from https://en.proft.me/2017/01/24/classification-using-random-forest-r/

Shams. (2018, April 17). *How can I measuring running time of R code*. Edureka Community. Retrieved July 20, 2022, from https://www.edureka.co/community/1460/how-can-i-measuring-running-time-of-r-code

Zach. (2020, November 4). *K-fold cross validation in R (step-by-step)*. Statology. Retrieved July 20, 2022, from https://www.statology.org/k-fold-cross-validation-in-r/

Abdo, A. (2021). *Regression Trees*. RPubs by RStudio. Retrieved July 20, 2022, from https://rpubs.com/tanboi/715151