Clustering Interpretation Optimization Using Decision Trees

Aryan Chaudhary, Arshin Jain, Shubham Dattatray Patil, Vimal Kirti Singh, Vinayak Abrol
IIIT Delhi
New Delhi
arshin22094, aryan22019, shubham22125, vimal22089, abrol, @iiitd.ac.in

Abstract

Current Clustering algorithms that we have, provide very less insight regarding cluster membership and provide very less interpretability. Due to this, we are not able to explain the recommendations made by the algorithm, whereas in the industry we often need a detailed explanation regarding the recommendations. In this project, we have tried to solve this problem by introducing a new clustering technique. This increases the interpretability of clusters forms. We will be leveraging the rules generated by the Decision Trees which forms the basis of our clustering algorithm. By doing this we provide a more clear understanding of why the specific data point belongs to the respective cluster by giving insights on every rule used.

1 Introduction:

Clustering is the task of dividing the population or data points into several groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is a collection of objects based on similarities and dissimilarities between them.

One such method for clustering is Decision trees. A decision tree is a supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical tree structure, which consists of a root node, branches, internal nodes, and leaf nodes. A decision tree starts with a root node, which does not have any incoming branches. The outgoing branches from the root node then feed into the internal nodes, also known as decision nodes. Based on the available features, both node types conduct evaluations to form subsets, which are denoted by leaf nodes, or terminal nodes. The leaf nodes represent all the possible outcomes in the dataset and also provides the number of clusters our data points can be segregated into and also makes it easy for user to interpret the rules every cluster.

Decision trees provides clusters. But in industry, we are often asked to provide more insights that would present a more clear understanding of the data. For example, from decision trees, we know that a particular point belongs to some cluster but we don't have exact knowledge of why that point went into that respective cluster and not to any other clusters. In such cases, we need to go a step further and make our clusters more interpretable. If we can figure out why a data point belongs to a particular cluster we can make a lot of business decisions with ease and it can help in the overall growth of the organization. In this project, we have tried to address this issue and have tried to test across various domains.

2 Literature Review

While the importance of cluster interpretability is well-understood, there has been limited success in addressing the issue. Several algorithms have been proposed to build interpretable clusters, where interpretability is a consideration during cluster creation rather than considered as a later analysis step introducing a tree-based approach, in which the clustering problem is translated into a supervised problem that is amenable to decision tree construction [2]. A modified purity criterion is used to evaluate splits in a way that identifies dense regions as well as sparse regions. However, this approach did not consider the unsupervised learning task as the primary objective [3]. Dimitris Bertsimas attempted to make clustering more interpretable by utilizing the rules and compared the performance of the same with other algorithms [1].

3 Methodology

The dataset we get is often not very clean and highly unbalanced with a lot of dimensions. Hence we are performing pre-processing steps on the data so that it becomes more consumable by our algorithm. In our clustering approach, the given dataset is translated into a supervised problem based on the merits of the features present in the data of a particular domain. We try to find the optimal decision tree by the hit and trial method. After performing a series of experiments involving hyperparameters like the number of leaves, depth of the tree, splitting criteria, etc. We achieve a decision tree that is more suitable to represent our data. This decision tree construction makes sure that the data is fully understood and grasped by the algorithm. Each leaf of the tree is equivalent to a cluster. Using this decision tree we extract the learned rules which help us in clustering the data itself. The number of rules generated represents the number of clusters. These rules then provide a more clear understanding of why and how the clusters are formed and in turn increases the interpretability.

3.1 Algorithm

Here we are using CART (Classification and Regression Trees) which is very similar to C4.5, but it differs in that it supports numerical target variables (regression) and does not compute rule sets. CART constructs binary trees using the feature and threshold that yield the largest information gain at each node. We are using the Scikit-Learn implementation which uses an optimized version of the CART algorithm.

We are using Gini index/entropy to split the tree depending on the dataset features.

Our decision tree has two types of nodes, the first node type is the decision node and the second type node is the cluster node. The decision node is the node where the splitting happens based on the features, which give the best results for our splitting criteria. Which can be again the Gini index or entropy. Cluster nodes are the leaf nodes where the final categorization happens. The path traversed from the root of the decision tree to the leaf node is called as decision path. Along this decision path, a set of rules are followed to partition the data. These rules eventually become the basis for our clustering. Each decision Path corresponds to one unique cluster. Each leaf of the tree is equivalent to a cluster. Observations in different leaves are not allowed to belong to the same cluster. We are generating the rules by recursively traversing the tree.

4 Results

4.1 Behavioral Risk Factor Surveillance System

4.1.1 Explain data

The Behavioral Risk Factor Surveillance System (BRFSS) is the United State's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and the use of preventive services. Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world. We are using the BRFSS 2015 data set for our analysis. We are using this data to analyze the health pattern in the population. Here we are focusing on diabetes.

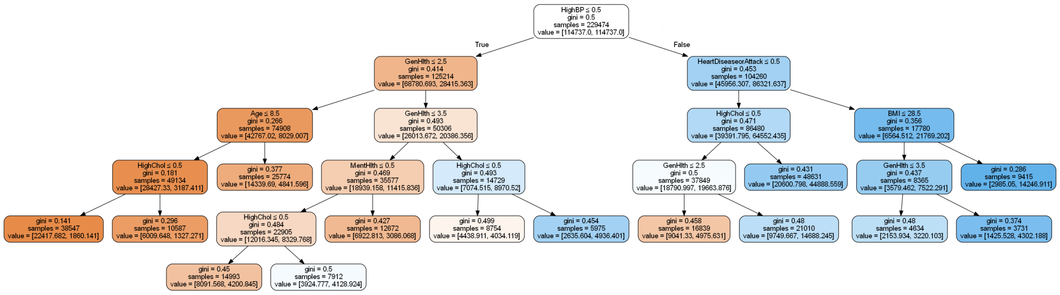
4.1.2 EDA & Preprocessing

The dataset originally has 330 features (columns), but based on diabetes disease research regarding factors influencing diabetes disease and other chronic health conditions, only select features are included in this analysis. Important risk factors for diabetes selected from the data are as follows :

blood pressure, cholesterol, smoking, diabetes, obesity, age, sex, race, diet, exercise, alcohol consumption, BMI, Household Income, Marital Status, Sleep, Time since the last checkup, Education, Health care coverage, Mental Health.

To do the analysis we have checked column names, the data type of each column, etc. Some attributes had data type as an object, we have converted it to int so that we can perform our analysis. We have also checked for null, unique value counts on each attribute.Also we analysed whether attributes are continuous or categorical. We have done binning on some of the attributes for better modelling of the data. We have removed columns that are not useful for our analysis. Analysis has been performed for checking Outliers. We have treated the dataset for null values, and duplicate values and have taken care of encoding.

4.1.3 Visual interpretation of the clusters



4.1.4 Rule-based interpretation of the clusters

Cluster1-> HighBP > 0.5 & HeartDiseaseorAttack ≤ 0.5 & HighChol > 0.5
Cluster2-> HighBP ≤ 0.5 & GenHlth ≤ 2.5 & Age ≤ 8.5 & HighChol ≤ 0.5
Cluster3-> HighBP ≤ 0.5 & GenHlth ≤ 2.5 & Age > 8.5
Cluster4-> HighBP > 0.5 & HeartDiseaseorAttack ≤ 0.5 & HighChol ≤ 0.5 & GenHlth > 2.5
Cluster5-> HighBP > 0.5 & HeartDiseaseorAttack ≤ 0.5 & HighChol ≤ 0.5 & GenHlth ≤ 2.5
Cluster6-> HighBP ≤ 0.5 & GenHlth > 2.5 & GenHlth ≤ 3.5 & MentHlth ≤ 0.5 & HighChol ≤ 0.5
Cluster7-> HighBP ≤ 0.5 & GenHlth > 2.5 & GenHlth ≤ 3.5 & MentHlth > 0.5
Cluster8-> HighBP ≤ 0.5 & GenHlth ≤ 2.5 & Age ≤ 8.5 & HighChol > 0.5
Cluster9-> HighBP > 0.5 & HeartDiseaseorAttack > 0.5 & BMI > 28.5
Cluster10-> HighBP ≤ 0.5& GenHlth > 2.5 & GenHlth > 3.5 & HighChol ≤ 0.5
Cluster11-> HighBP ≤ 0.5 & GenHlth > 2.5 & GenHlth ≤ 3.5 & MentHlth ≤ 0.5 & HighChol > 0.5
Cluster12-> HighBP ≤ 0.5 & GenHlth > 2.5 & GenHlth > 3.5 & HighChol > 0.5
Cluster13-> HighBP > 0.5 & HeartDiseaseorAttack > 0.5& BMI ≤ 28.5 & GenHlth ≤ 3.5
Cluster14-> HighBP > 0.5 & HeartDiseaseorAttack > 0.5 & BMI ≤ 28.5 & GenHlth > 3.5

4.1.5 Business value explanation via interpretability:

Here we will define only a few clusters which have high business value. Inference A: People with HighBP > 0.5 have a very high chance of having diabetes
Inference B: People with HighChol ≤ 0.5 are less likely to have diabetes, whereas HighChol > 0.5 are most likely to have diabetes
Inference C: People with Age$\leq$8.5 are very less likely to have diabetes.
Inference D: People with HighBP $>$0.5 are extremely likely to have a heart attack
Inference E: People with GenHlth > 3.5 are usually more likely to be unfit.
The above inferences can be used in the following scenarios:
Example if we are a pharmaceutical company and want to sell a diabetes testing kit we can do our analysis based on the above inference and pick the best possible target audience who will purchase

this product.
Even if catered to the general public, the above inference will help government organizations to spread awareness of diabetes.
Also every segment represents a different category of person who may/ may not be having diabetes and based on the above clusters different products can be developed catering to their needs. For example 1, for people belonging to Inference D, we can recommend them a Blood pressure monitoring machine. Example 2. People belonging to the Inference E category are fitness conscious so we can recommend fitness products.

4.2 Adidas US Retail Products Dataset

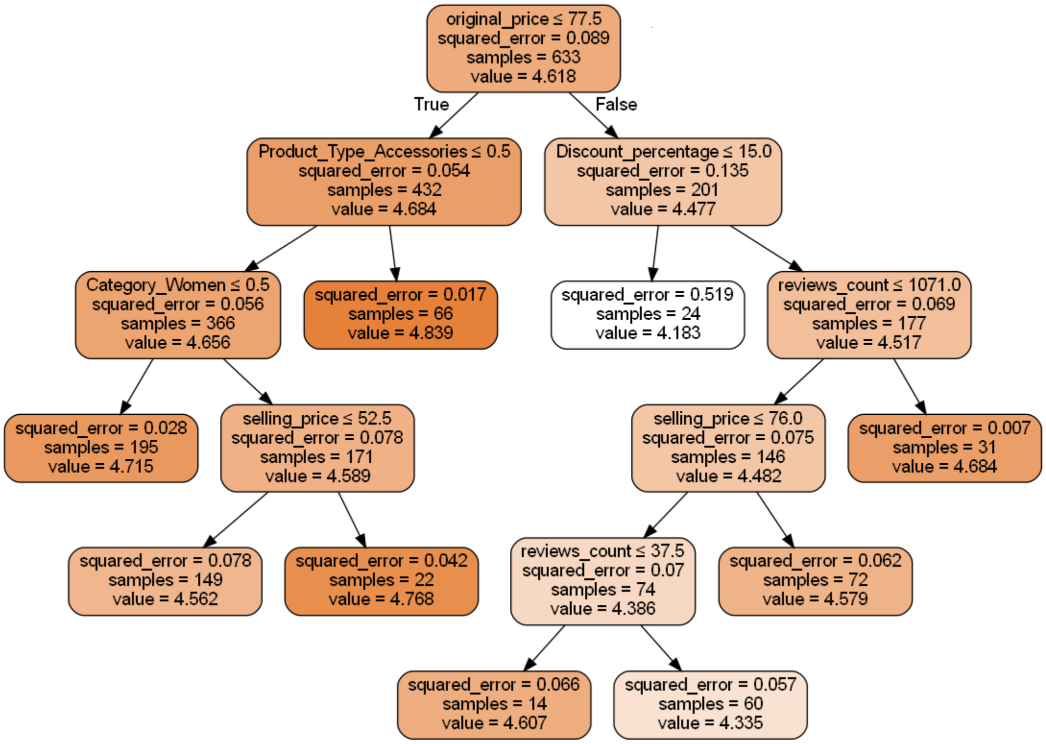
4.2.1 Explain Data

This dataset contains information on Adidas fashion products. The data includes fields such as name, selling price, original price, currency, availability, color, category, source website, breadcrumbs, description, brand, images, country, language, average rating and reviews count. This data was collected from a variety of sources and compiled into one dataset for research purposes It contains detailed information on product selling price, original price in multiple currencies (USD / EUR / GBP), product availability (in stock / out of stock), color, Category (such as Apparel / Footwear), source website, breadcrumbs, product description, brand name, link to product images, Country of origin and language. The average rating and reviews count are also included in the dataset so that researchers can study the correlation between them.

4.2.2 EDA & Preprocessing

After performing adequate EDA, we have treated the dataset for null values (by replacing them with mean), and duplicate values and have taken care of encoding. We have added new columns like discount percentage, product type, and category from existing attributes. We have removed columns with zero variance.

4.2.3 Visual interpretation of the clusters



4.2.4 Rule-based interpretation of the clusters

Cluster 1-> original price $\leq$ 77.5 & Product Type Accessories $\leq$ 0.5 & Category Women $\leq$ 0.5
Cluster 2-> original price $\leq$ 77.5 & Product Type Accessories $\leq$ 0.5 & Category Women > 0.5 & selling price $\leq$ 52.5
Cluster 3-> original price > 77.5 & Discount percentage > 15.0 & reviews count $\leq$ 1071.0 & selling price > 76.0
Cluster 4-> original price $\leq$ 77.5 & Product Type Accessories > 0.5
Cluster 5-> original price > 77.5 & Discount percentage > 15.0 & reviews count $\leq$ 1071.0 & selling price $\leq$ 76.0 & reviews count > 37.5
Cluster 6-> original price > 77.5 & Discount percentage > 15.0 & reviews count > 1071.0
Cluster 7-> original price > 77.5 & Discount percentage $\leq$ 15.0
Cluster 8-> original price $\leq$ 77.5 & Product Type Accessories $\leq$ 0.5 & Category Women > 0.5 & selling _price > 52.5
Cluster 9-> original _price > 77.5 & Discount _percentage > 15.0 & reviews _count $\leq$ 1071.0 & selling _price $\leq$ 76.0 & reviews _count $\leq$ 37.5

4.2.5 Business value explanation via interpretability

Here we will define only a few clusters that have high business value.
Inference A: products that belong to the category with Original _price $\leq$ 77.5, Discount percentage $\geq$ 15, and reviews count 1071 are products with very high ratings. (cluster 7)
Inference B: Products of Type Accessories are highest rated
Inference C: Products that belong to the women's category have high differences in original price and selling price and have high ratings The above inferences can be used in the following scenarios:
Products following inference A are better products and should be pushed ahead to the customers. These products also have fewer discounts, in turn giving more profits. Also, customers purchase such products irrespective of their selling price.
Products catered to women require high discounts.

4.3 Credit Card Customers

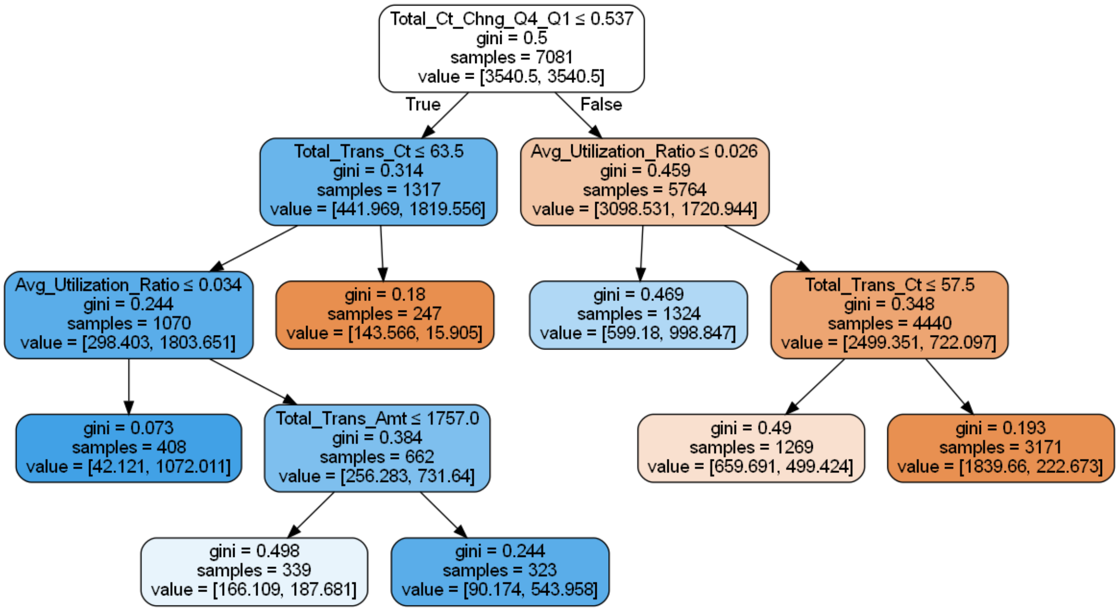
4.3.1 Explain Data

This dataset is addressing the problem that the bank is disturbed by more and more customers leaving their credit card services. They would really appreciate it if we could predict for them who is gonna get churned so they can proactively go to the customer to provide them better services and turn customer's decisions in the opposite direction.

4.3.2 Preprocessing

Now, this dataset consists of 10,000 customers mentioning their age, salary, marital_status, credit card limit, credit card category, etc. There are nearly 18 features. We have only 16.07 % of customers who have churned. We have performed ordinal encoding and one hot encoding. There was a class imbalance in the data which was taken was taken care of during model training by assigning weights to the classes along with the stratified k-fold strategy. To identify the classes we have kept retained customer as class zero and attrited customer as class one.

4.3.3 Visual interpretation of the clusters



4.3.4 Rule-based interpretation of the clusters

Cluster1-> Total _Ct _Chng _Q4 _Q1 > 0.537 & Avg _Utilization _Ratio > 0.026 & Total _Trans _Ct > 57.5
Cluster2-> Total _Ct _Chng _Q4 _Q1 > 0.537 & Avg _Utilization _Ratio ≤ 0.026
Cluster3-> Total _Ct _Chng _Q4 _Q1 > 0.537 & Avg _Utilization _Ratio > 0.026 & Total _Trans _Ct ≤ 57.5
Cluster4-> Total _Ct _Chng _Q4 _Q1 ≤ 0.537 & Total _Trans _Ct ≤ 63.5 & Avg _Utilization _Ratio ≤ 0.034
Cluster5-> Total _Ct _Chng _Q4 _Q1 ≤ 0.537 & Total _Trans _Ct ≤ 63.5 & Avg _Utilization _Ratio > 0.034 & Total _Trans _Amt ≤ 1757.0
Cluster6-> Total _Ct _Chng _Q4 _Q1 ≤ 0.537 & Total _Trans _Ct ≤ 63.5 & Avg _Utilization _Ratio > 0.034 & Total _Trans _Amt > 1757.0
Cluster7-> Total _Ct _Chng _Q4 _Q1 ≤ 0.537 & Total _Trans _Ct > 63.5

4.3.5 Business value explanation via interpretability

Here we will define only a few clusters, which have high business value. Inference A: When the total transaction rate from quarter 1 to quarter 4 is less than equal to 0.537 the customer is most likely to leave the credit card services.
Inference B: When the total transaction rate from quarter 1 to quarter 4 is less than equal to 0.537 and the total transaction count > 63.5 then those customers are retained.
Inference C: When the total transaction rate from quarter 1 to quarter 4 > 0.537 are very less likely to leave.
Inference D: When the total transaction rate from quarter 1 to quarter 4 > 0.537 and the Average Utilization Ratio ≤0.026 are likely to leave.
The above inferences can be used in the following scenarios:
Customer Following inference B are customers who are currently using the credit card services but are most likely to leave. Hence, the Bank must provide more cashback and rewards to these types of customers.
Customers following Inference C are the loyal customers who would stay with the bank and hence the bank can save money by giving them fewer rewards.
Customers following inference D are the ones which are leaving but were using the services heavily. The bank must try to retain these customers by giving them the best offers and rewards.

4.4 Mental Health Dataset

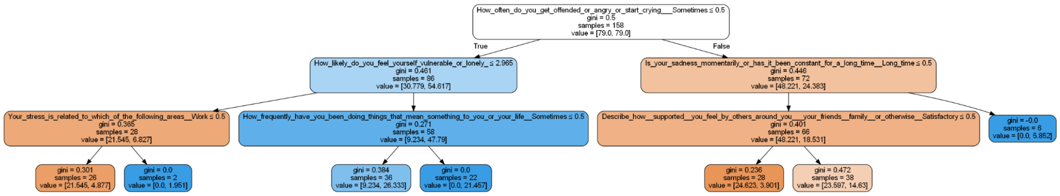
4.4.1 Explain the data

This dataset contains information on Mental Health Patients. The data includes fields that focus more on the patient's behavior, the patient's surroundings, and other habits which could be used for the analysis of his/her mental health. This is a synthetic artificial dataset used for analytics purposes.

4.4.2 EDA and preprocessing

After performing the adequate EDA, we dealt with duplicate rows which were increasing the redundancy in the data. After that, we replaced the NULL value with the mean value. Some features were going through a collinearity problem and were not adding value to the inference so we dropped those features and have taken care of encoding.

4.4.3 Visual interpretation of the clusters



4.4.4 Rules:

Cluster1-> How often do you get offended or angry or start crying Sometimes > 0.5 & Is your sadness momentarily or has it been constant for a long time Long time $\leq$ 0.5 & Describe how supported you feel by others around you your friends family or otherwise Satisfactory > 0.5

Cluster2-> How often do you get offended or angry or start crying Sometimes $\leq$ 0.5 & How likely do you feel yourself vulnerable or lonely > 2.965 & How frequently have you been doing things that mean something to you or your life Sometimes $\leq$ 0.5

Cluster3-> How often do you get offended or angry or start crying Sometimes > 0.5 & Is your sadness momentarily or has it been constant for a long time Long time $\leq$ 0.5 & Describe how supported you feel by others around you your friends family or otherwise Satisfactory $\leq$ 0.5

Cluster4-> How often do you get offended or angry or start crying Sometimes $\leq$ 0.5 & How likely do you feel yourself vulnerable or lonely $\leq$ 2.965 & Your stress is related to which of the following areas Work $\leq$ 0.5

Cluster5-> How often do you get offended or angry or start crying Sometimes $\leq$ 0.5 & How likely do you feel yourself vulnerable or lonely > 2.965 & How frequently have you been doing things that mean something to you or your life Sometimes > 0.5

Cluster6-> How often do you get offended or angry or start crying Sometimes > 0.5 & Is your sadness momentarily or has it been constant for a long time Long time > 0.5

Cluster7-> How often do you get offended or angry or start crying Sometimes $\leq$ 0.5 & How likely do you feel yourself vulnerable or lonely $\leq$ 2.965 & Your stress is related to which of the following areas Work $\geq$ 0.5

4.4.5 Business value explanation via Interpretability

Dealing with emotions is another task for today's generation hence we build a modal that serves to help people who are depressed, anxious, or chronically sad. Our idea helps the patient to detect his mental health in the early stages.

Inference A: How often do you get offended or angry or start crying Sometimes is less than equal to 0.5 and How likely do you feel yourself vulnerable or lonely is greater than equal to 2.96 this means that it is highly probable that an individual is a mental patient.

Inference B: How often do you get offended or angry or start crying Sometimes is greater than equal to 0.5 and Is your sadness momentarily or has it been constant for a long time Long time is greater than equal to 0.5 this means that it is highly probable that an individual is a mental patient.

Inference C: How often do you get offended or angry or start crying Sometimes is greater than equal to 0.5 and Is your sadness momentarily or has it been constant for a long time Long time is less than equal to 0.5 this means that an individual is not a mental patient.

Inference D: How often do you get offended or angry or start crying Sometimes is less than equal to 0.5 and How likely do you feel yourself vulnerable or lonely is less than equal to 2.96 and Your stress is related to which of the following areas Work is less than 0.5 this means that an individual is not a mental patient.

Inference E:How often do you get offended or angry or start crying Sometimes is less than equal to 0.5 and How likely do you feel yourself vulnerable or lonely is less than equal to 2.96 and Your stress is related to which of the following areas Work is greater than 0.5 this his means that it is highly probable that an individual is a mental patient.

The above inferences can be used in the following scenarios: For the patients from segments A, B and E we would suggest to them various ways to improve their mental health such as going to psychologist, meeting new people, sharing their problems, indulging in activities like yoga, walking, etc. For the patients from segment C and D are the one who does not have any mental condition but still suffering from some kind of stress so we would suggest that they take a break from their daily routine or meditate or join a new activity.

5 Conclusion

We have implemented a methodology of cluster creation that addresses the issue of cluster Interpretability. Our method extends the framework of Optimal Classification Trees in which we build trees that provide explicit separations of the data on the original feature set. Our results suggest that we can recover clusters similar to existing clustering techniques, but with the added advantage of interpretability. We have applied this method across various applications, such as in healthcare, finance, e-commerce, etc. We hope that this method can also be applied to other varied segments and give interesting insights. For future works, we can optimize decison trees more by introducing newer metrics and splitting techniques.

References

[1] Dimitris Bertsimas, Agni Orfanoudaki, and Holly Wiberg. Interpretable clustering via optimal trees. *arXiv preprint arXiv:1812.00539*, 2018.

[2] Dimitris Bertsimas, Agni Orfanoudaki, and Holly Wiberg. Interpretable clustering: an optimization approach. *Machine Learning*, 110(1):89–138, 2021.

[3] Luke Zappia and Alicia Oshlack. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *Gigascience*, 7(7):giy083, 2018.