REVA UNIVERSITY
Bengaluru, India

A Project Report on

Explaining Clustering using Decision Trees

Submitted in Partial Fulfilment for Award of Degree of

Master of Technology

In Artificial Intelligence

Submitted By

Yashaswini Viswanath

R20MTA12

Under the Guidance of

Dr. J.B. Simha

Chief Mentor, Artificial Intelligence, RACE, REVA University

REVA Academy for Corporate Excellence - RACE

REVA University

Rukmini Knowledge Park, Kattigenahalli, Yelahanka, Bengaluru - 560 064

race.reva.edu.in

August, 2022

REVA UNIVERSITY
Bengaluru, India

Candidate's Declaration

I, Yashaswini Viswanath hereby declare that I have completed the project work towards the Master of Technology in Artificial Intelligence at REVA University on the topic entitled **"Explaining Clustering using Decision Trees"** under the supervision of Dr. J.B.Simha, Chief Mentor, Artificial Intelligence, RACE, REVA University. This report embodies the original work done by me in partial fulfilment of the requirements for the award of the degree for the academic year 2022.

Place: Bengaluru

Name of the Student: Yashaswini Viswanath

Date: 20/08/2022

Signature of Student:

REVA UNIVERSITY
Bengaluru, India

Certificate

This is to Certify that the project work entitled **"Explaining Clustering using Decision Trees"** carried out by Yashaswini Viswanath bearing SRN R20MTA12, is a bonafide student of REVA University, is submitting the project report in fulfilment for the award of Master of Technology in Artificial Intelligence during the academic year 2022. The Project report has been tested for plagiarism and has passed the plagiarism test with a similarity score of less than 15%. The project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the said Degree.

Dr. J.B. Simha
Guide

Dr. Shinu Abhi
Director

External Viva
Names of the Examiners

1. Dr. Santosh Nair, Founder, Analytic Edgex
2. Rajkumar Dan, Data Scientist Consultant, Dell

Place: Bengaluru
Date: 20/08/2022

Acknowledgements

I am highly indebted to **Dr. Shinu Abhi**, Director, Corporate Training for the guidance and support provides throughout the course and my project.

I would like to thank **Dr. J. B. Simha** for the valuable guidance provided as my project guide to understand the concept and in executing this project.

It is my gratitude towards all other mentors for the valuable guidance and suggestions in learning Artificial Intelligence. I am thankful to my classmates for their support, suggestions, and friendly advice during the project work.

I would like to acknowledge the support provided by the founder and Hon'ble Chancellor, **Dr. P Shayma Raju**, Vice-Chancellor, **Dr. M. Dhanamjaya**, and Registrar, **Dr. N Ramesh**.
It is sincere thanks to all members of the program office of RACE who were always supportive in all requirements from the program office. It is my sincere gratitude towards my parents and my family for their kind co-operation. Their encouragement also helped me in the completion of this project.

This acknowledgement will not be complete without thanking the guiding light of my life, my grandfather **Dr. T. Basavaraju** who served as Professor in UVCE, Bangalore University for 35 years and held multiple positions as Director in various colleges in Bangalore.

Last but not the least, I would like to thank my child M.V. Vismaya who was the motivation for me to join this program and sacrificed a lot for fulfilment of my dreams.

Place: Bengaluru
Date:20/08/2022

REVA UNIVERSITY
Bengaluru, India

Similarity Index Report

This is to certify that this project report titled **"Explaining Clustering using Decision Trees"** was scanned for similarity detection. The process and outcome are given below.

Software Used: Turnitin

Date of Report Generation: 20th Aug 2022

Similarity Index in %: 5%

Total word count:

Name of the Guide: Dr. J. B. Simha

Place: Bengaluru

Date: 20/08/2022

Name of the Student:

Yashaswini Viswanath

Signature of Student

Verified by:

M N Dincy Dechamma

Signature

Dr. Shinu Abhi,

Director, Corporate Training

List of Abbreviations

Sl. No	Abbreviation	Long Form
1	AI	Artificial Intelligence
2	ML	Machine Learning
3	LIME	Local Interpretable Model-Agnostic Explanations
4	SHAP	Shapley Additives

List of Figures

Abstract

Explainability of AI models is the need of the hour as AI systems are proliferating into decision making institutions in the form of AI agents helping humans take data driven decisions. Explainability of decision taken by these agents will help the data scientist who models, regulatory bodies and public who are affected by it. Among the many types of machine learning models, unsupervised machine learning models do not require labelling of the dataset.

The objective of this project is to explore these explainability techniques for clustering. These are algorithms which do not need labelling and are at adept at finding hidden patterns in the dataset. One of the important categories is kmeans where we compute centroids and form clusters.

In this project decision tree as a surrogate model is explored and the number of clusters, type of loss of tree, depth of tree, number of leaves of the tree along with the number of features of data is considered for experimentation.

Pima Indians Diabetes dataset is chosen to empirically prove the research hypothesis. The dataset is specific to Indian women and is a matter of great pride to conduct research on it. The research hypothesis about percentage of explainability is formulated, verified, and accepted in this project. The highlight of this project is the novelty in tackling the percentage of explainability.

The research hypothesis about number of features and number of clusters required for explainability is experimented and results are analyzed. The conclusion is that a smaller number of clusters and a smaller number of features are adequate to explain the cluster assignment of kmeans. As a future scope it can be extended to number of leaves and depth of decision tree.

Keywords: Responsible AI, Explainable AI, clustering, decision tree

Table of Contents

Chapter 1: Introduction

Artificial Intelligence (AI) is evolving over time and new innovations are an integral part of this evolution. Similar to any new technology that humans have co-created, Artificial Intelligence use cases raise a lot of fundamental questions [1]. The black box nature of Artificial Intelligence raises alarms in many sectors of usage [2]. But the benefits of Artificial Intelligence and Machine Learning (ML) are necessary for the societal impact these algorithms manifest. The problem we are addressing in this project is how can we overcome the black box nature of AI and ML which leads to democratisation of these use cases [3]. This project provides an innovative solution in overcoming this black box nature of Artificial Intelligence Models, specifically clustering techniques using k-means algorithm.

Biases in the predictions of the algorithms used in AI systems raises a cause of concern and impedes the widespread adoption of AI at a rapid pace. There are two main categories of Bias which are data bias and algorithmic bias. Data Bias involves the dataset used for training the model largely depending on the collection mechanism and sources used to create the dataset. Algorithmic Bias is due to the data that is input into the training phase which leads to predictions which have the influence of a bias. The solution to overcome these two issues is by introducing explainability of models [1] which elaborates the biases and helps decision makers to make informed decisions based on AI and ML model prediction. This project puts forth an explainability mechanism for a subset of techniques used in practice.

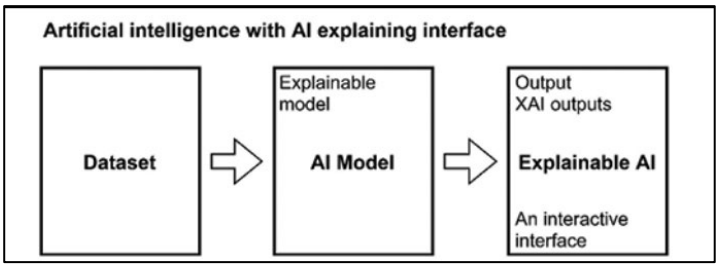


Figure No.1.1: Explainable AI Process [2]

The state of the art practice methodology for the explanatory process of AI predictions is illustrated in Figure No. 1.1. The term used henceforth to address this problem of explanation for bias mitigation will be explainable AI and is a term used by industry and academia [2]. Explainable AI is defined as that process which elucidates the reason behind the predictions of AI and ML models [1]. The various dichotomies of Explainable AI which enable differentiation of techniques in Explainable AI are: local vs global, where the user is interested in explainability of a single input data point or interested in the prediction by the overall model; exact vs approximate, where the explanation needs to be completely faithful to the model or there is scope for approximations for explanations; feature based vs sample based where explanations are generations with reference to the features of the model or the sample input data points are
used to generate explanations.

Various literature provides different taxonomies of techniques. The most widely used types of explainable AI are discussed here. Intrinsic explanations where the models have inherent explainability e.g. decision trees. Post hoc explanations where models which are complex need analysis post prediction e.g. stacked models. Model specific refers to explainability mechanisms which work only with certain models e.g. linear regression coefficients. Model agnostic explanation techniques consider only the input and output data points to provide the explanations. In Figure 1.1 we see that there is an explainability interface and this gives rise to important types of explanations such as textual, visual explanations [1]. The author of this project proposes newer types of explanation interfaces such as audio and multi-modal explanations. As explainable AI evolves, newer taxonomy has been introduced, one of which is: model-centric and data- centric explanations where the above discussed techniques were all model centric and data centric comes from data that is used for training. The suitability of the data for problems trying to be solved is the focus of such newer types of explainability [3].

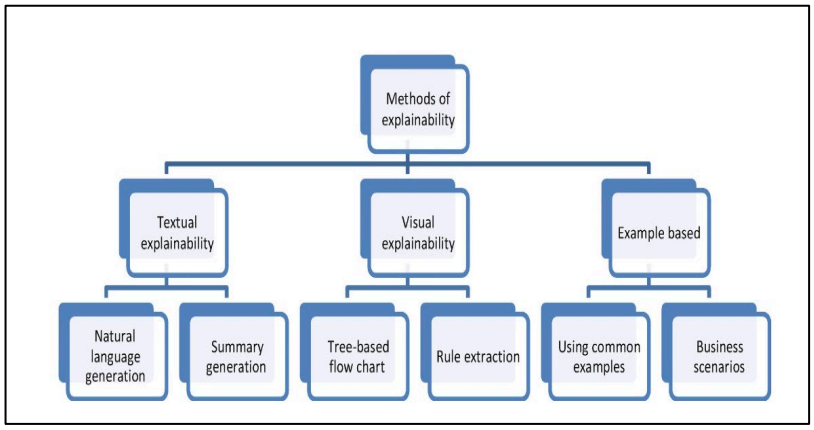


Figure No.1.2: Methods of Explainable AI [1]

The method of doing explainability can be classified into 3 types as textual explainability, visual explainability and example based as shown in Figure No.1.2. The outcome of a model parameter, or metrics defined by the model give raise to textual explanations. But in this type of explainability there is a need for Natural Language Generation to populate templates created based on storyline with updated parameters. If else statements can be used for visual explainability and tree charts for the same. In example based techniques the business scenario can be used or an analogy to common examples can be employed [1].

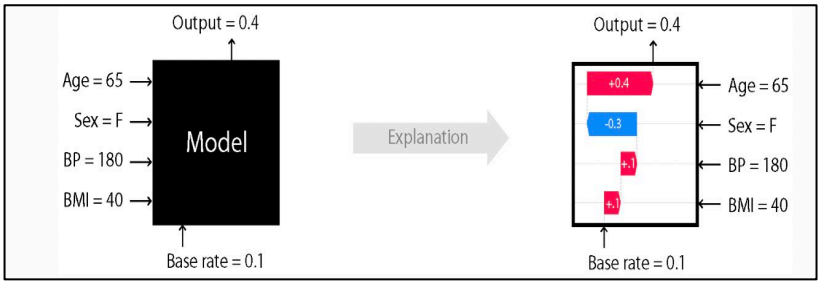


Figure No. 1.3: SHAP explanations [1]

In practice, the widely used tools are LIME and SHAP(Shapley Additives) which provide post-hoc analysis of the model. SHAP is based on Game Theory where certain features are added and removed to compute the factor [17]. SHAP provides visual explanation in the format shown in Figure 1.3. LIME also provides post hoc analysis and this brings up the need for pre modelling

explainability on unlabelled datasets. LIME is the abbreviation for Local Interpretable Model-Agnostic Explanations. **Local** refers to local fidelity - i.e., we want the explanation to really reflect the behaviour of the classifier "around" the instance being predicted. This explanation is useless unless it is **interpretable** - that is, unless a human can make sense of it. Lime is able to explain any model without needing to 'peak' into it, so it is **model-agnostic** as shown in Figure No. 1.4.

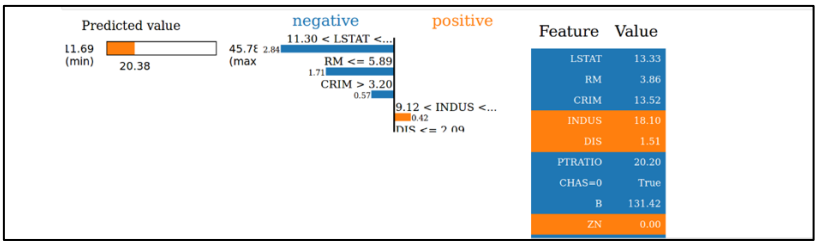


Figure No. 1.4: LIME explanation for ML models [1]

The challenge faced is the inclusion of all features for cluster assignment due to which we cannot have a concise way of cluster assignment. There is a need for research to come up with techniques which are simple but powerful to explain the cluster assignment of a data point [4].

Chapter 2: Literature Review

Recent Clustering techniques find patterns in data by quantization of data points which are unlabelled. The users of clustering algorithms want to know why a data point was assigned to a particular cluster. The challenge is due to consideration of all the features to form cluster assignments. The initial introduction of decision trees for solving this explainability problem was successful [23] [24]. An unsupervised decision tree was a canonical example of a clustering model with explainability. The improvement in explainability can be achieved by usage of a small decision tree to partition the data set into clusters. This is a straightforward approach as we can characterise each of the clusters using the decision tree [5]. This approach is independent of input size and dimensions and provides novelty compared to older approaches. The characteristics of a tree by restricting to k leaves where k is the number of clusters, the solution is independent of data dimension. This threshold tree uses k-1 features and any dataset can be quantized. Also, any new data point can be assigned a cluster using the tree.

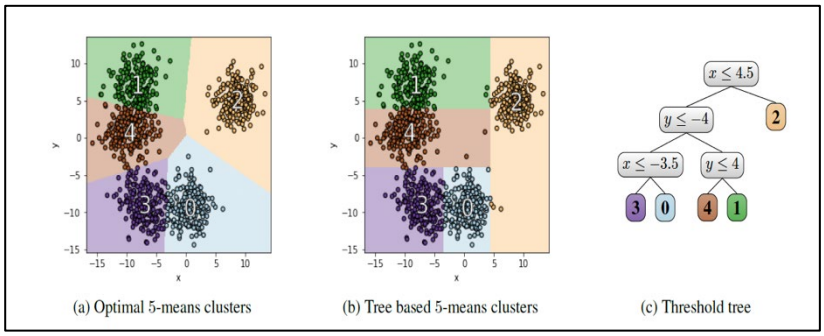


Figure 2.1: Example of a decision tree for cluster assignment problem [5].

The challenge is the price of explainability when the final clustering is forced to have an interpretable form. Also, the efficiency of finding such a tree in Figure 2.1 is based on the choice of features. The leaves of the decision tree is the collection of all clusters that we want to create from the unlabelled dataset. The mechanism to build the threshold tree can be inferred from Figure 2.2 and this forms the basis of this project.

The trade off between accuracy and explainability is taken into consideration while coming up with an optimal solution. The single feature thresholds are the highlight of these approaches which enables simple explanations using small height trees. ExKMC algorithm improvises on this approach by taking an additional parameter k' which results in a decision tree with k' leaves but still the number of clusters is k and we label the leaves with these clusters. The nodes of a tree in a binary threshold tree recursively splits the dataset into clusters each based on a single feature which results in simple explanations. The number of leaves are increased and experimented to find the impact [6]

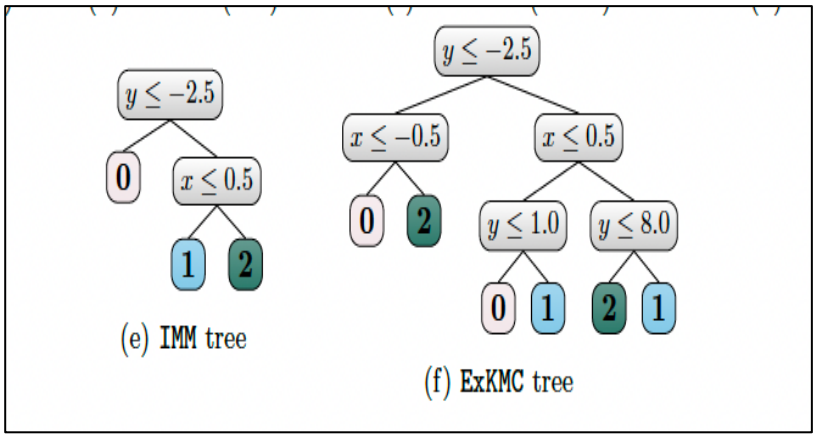


Figure 2.2: Tree with k' leaves > k clusters [6]

ExKMC approach uses a greedy method to increase the number of leaves as the clustering cost reduces with increase in number of leaves. Also, the increase in the number of thresholds give rise to better partitioning of the dataset. Each leaf is assigned a label of best centroid which determines the cluster assignment. We have a surrogate model which provides the decision tree and the cost of this surrogate model will be non-increasing [6]. This approach is explainable by design. These algorithmic problems try to come up with solutions which are approximate in nature by elimination of few outlier points [14].

The efficiency of k-means price of explainability was improved in near optimal solutions proposed by researchers [8][9]. A novel divide and share technique to

make these explainable trees efficient was proposed [7]. This was more efficient by using an algorithm to divide and share the region of data points but resulted in larger trees. These trees are proved to display lesser upper bounds for the price of explainability. This project experiments by considering the idea of larger trees (more leaves) vs the smaller trees we saw earlier. The depth of the surrogate decision tree is another parameter that can be varied during experimentation [10]. The literature has so far looked at IMM, ExKMC, ExGreedy and ExShallow methods and this project tries to amalgamate the best of all worlds to create a kmeans algorithm with surrogate model for the datasets we are considering.

All the literature review have considered the complexity of explainability and are comparing the time complexities and trying to be outperform each other. There is a gap in the literature review about the percentage of explainability which can be achieved using surrogate models. Also, the impact of different parameters which are prevalent in the model being explained and the surrogate model characteristics. This calls for experiments in this research direction.

Chapter 3: Problem Statement

Explainable AI improves trust and transparency of AI models. The accuracy and explainability of a model are inversely proportional. The methods of unsupervised learning forms the basis of many AI applications and there is a need for explainability of such models. The prominent algorithm technique is k means clustering where we have data points segregated into “k” clusters where each cluster is similar with data points within itself.

The research hypothesis for Surrogate models for explaining Clustering mechanisms is:

Research Hypothesis 1: Less than 50% of features are adequate to explain the cluster assignment

Research Hypothesis 2: Less than 5 clusters of kmeans clustering are adequate to explain the variance of dataset

Chapter 4: Objectives of the Study

The primary objective of this study is to understand the exploration of explainability techniques for clustering mechanisms such as kmeans. This study helps in understanding the underlying patterns of impact of number of clusters and number of features on the accuracy of explanations. This motivation for this objective is the gap we are addressing in the current research conducted in this domain.

The objective is to empirically prove the hypothesis of the project using Diabetes Dataset where kmeans clustering of the dataset enables unsupervised learning of patterns hidden in the dataset. The decision tree surrogate model is built for the clustering of Diabetes dataset and experiments are conducted to compute the percentage of explainability empirically.

Chapter 5: Project Methodology

The Pima Indian Diabetes dataset needs to be prepared for usage in this project. The dataset needs to be cleaned to be usable for further processing. Processing the dataset such that the models can find patterns is a crucial step in the project methodology. Exploratory data analysis is to be conducted to understand the diabetes dataset where the features of the dataset are the predictor variables. Once the data is ready, kmeans model needs to be built and cluster assignment is to be performed. The important step is to build a decision tree classifier with cluster segments as labels. The so created decision tree classifier will provide explanations where the leaf nodes are the clusters and the non-leaf nodes are the features and their ranges. The accuracy of the decision tree is the percentage of explainability. The project methodology is pictorially shown in Figure No. 5.1

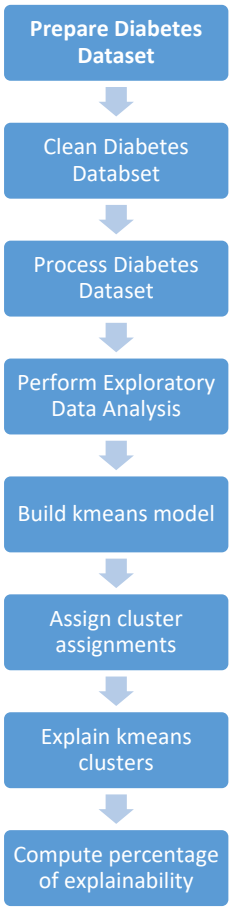


Figure No. 5.1: Method to be followed for project execution

Chapter 6: Experiment Design

Experiment Design for this project involves multiple steps shown in Figure No. 6.1 which are the standard processes in data driven experimentation.

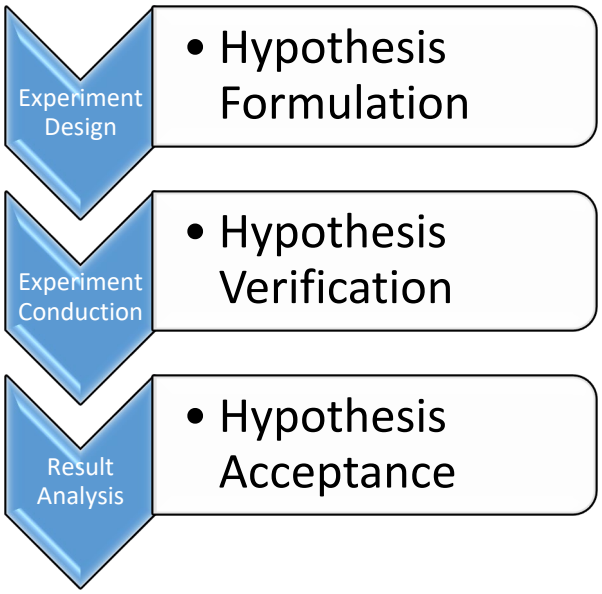


Figure No. 6.1: Research Methodology

Experiment Algorithm Selection

Unsupervised machine learning algorithms are adept at finding hidden patterns and this field of machine learning is considered as the starting point for this study. Among many techniques in the basket of unsupervised algorithms after thorough scrutiny of literature, k-means clustering is selected. The model which provides the explainability of kmeans is a surrogate model meaning they are created separately without modification of the technique being explained. The state of the art explainability techniques involve employing a decision tree to provide the explanations for cluster assignment. An example of the relationship between kmeans and decision tree used to explain the cluster assignment is shown in Figure No. 6.2.

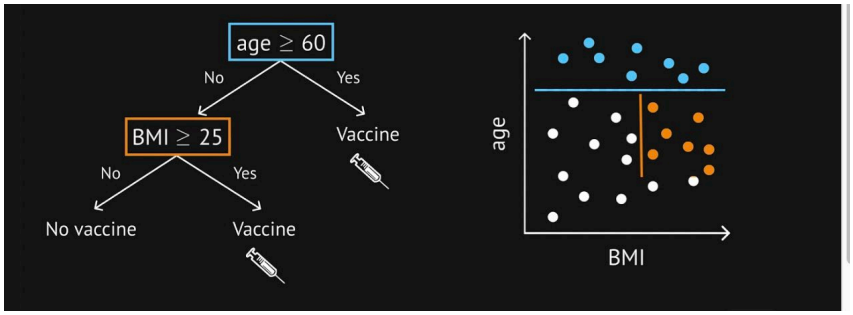


Figure No. 6.2: Clustering and decision trees relationship [5]

Experiment Parameter Selection

The experiment design involves deciding the domain and range of parameters used in the experiments. The parameters are selected based on research hypothesis that we are testing. The parameters selected for the experiment design for this project is shown in Figure No. 6.3.

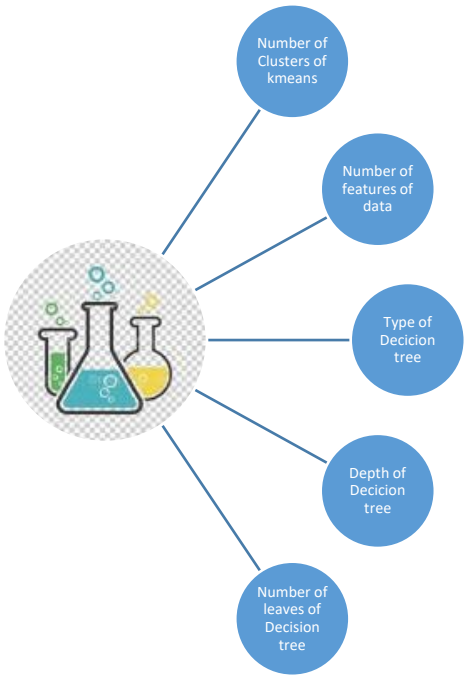


Figure No. 6.3: Five-Pronged parameter selection for this project

Number of Clusters of kmeans

The number of clusters of kmeans algorithm is chosen as the parameter which is varied from 2 to in steps of 1. This is the input to kmeans algorithm which uses unsupervised technique to find patterns and cluster. The significance of number of clusters is how more throughput can be achieved with a smaller number of clusters. In this experiment design the following number of clusters are considered: 2,3,4,5,6,7,8,9.

Number of features of data

The number of features required to explain the cluster assignment is a problem being experimented in this project. The relationship between this parameter and the percentage of explainability is important. The significance of number of features is how most explanations can be arrived at with lesser number of features. In this experiment design the following number of features is considered as the dataset chosen has a total of 8 features: 1, 2, 3, 4, 5, 6, 7, 8.

Type of Decision tree

The main model which is kmeans is unmodified to enhance it with explainability. The usage of surrogate model comes at an additional cost but gives flexibility. The significance of type of tree by using the type of loss as varying criteria is how the type of tree affect the percentage of explainability. In this experiment design the following types of loss as criteria is chosen: gini and entropy. These two criteria are used to split the node of the decision tree.

Depth of Decision tree

The depth of Decision tree and the complexity of explainability has been experimented in the literature review [7]. The novelty brought by this project is the experimentation of the depth to find a relationship with the percentage of explainability. The depth of decision tree is of significance as this does not impact the kmeans clustering and hence can be varied for maximum

explainability. In this experiment design the depths that are considered for experimentation are: 2,3,4,5,6,7,8,9.

Number of leaves of Decision tree:

The number of leaves of Decision tree and the complexity of explainability has been experimented in the literature review [11]. The novelty brought by this project is the experimentation of number of leaves to find a relationship with the percentage of explainability. The number of leaves is of significance as this does not impact the kmeans clustering and there can be multiple leaves for the same cluster, and this is not a constraint. In this experiment design the number of leaves that are considered are: 4,5,6,7,8,9,10,11,12.

The goal of this experiment design is to compute the percentage of explainability by varying the experimental parameters which are input to kmeans algorithm and corresponding decision tree as surrogate model.

Chapter 7: Resource Requirement

Dataset required for this project is sourced from National Institute of Diabetes and Digestive and Kidney Diseases. This research institution is instrumental in conduction of research on common and chronic health conditions. The mission of National Institute of Diabetes and Digestive and Kidney Diseases is dissemination of science-based information for improvement of quality of life and health of general public. As part of the Institution's research activities, this dataset was captured and released to further study of researchers. The significance of this dataset is the population under study is Indian citizens. This gives us pride to conduct research on this dataset.

Diabetes is a chronic illness due to sedentary lifestyle of patients. It is a condition where the insulin production is commensurate with the food intake for the proper functioning of the metabolism. Diabetes affects people of varied age groups and is monitored across the lifetime. This condition can affect kids, teenagers, young adults, middle aged and geriatric but is seen commonly in middle aged and geriatric population. This proliferation of diabetes is the motivation for usage of this dataset in this project.

Gender bias issues have been a cause of concern in data science and artificial intelligence. One of the main causes of Gender bias issues that are detected is the undersampling error while data is sourced. This calls for various bias mitigation algorithms.

The goal of this dataset is to predict occurrence of diabetes diagnostically based on the measures in the dataset. The variables which help in prediction are shown in Figure No. 7.1.

1. Pregnancies: Number of times a woman has been pregnant before collection of the data. A value of 0 indicates no pregnancy and any other number denotes the number of times.

2. Glucose: Glucose levels are an important diagnostic measure for diabetes prediction and one of the easiest to measure given the proliferation of portable battery-operated medical devices which measure the level of glucose. Here the type of glucose measured is Plasma glucose with a concentration of 2 hours in an oral glucose tolerance test.
3. Blood Pressure: One of the diagnostic measures of diabetes is Blood pressure where medical devices have been developed which are portable and battery operated. There are two types of Blood Pressure that is generally measured together which are Diastolic and Systolic. For this dataset Diastolic blood pressure in units of mmHg is considered.
4. Skin thickness: The Triceps skin fold thickness in units of mm is considered for diagnostics in this dataset.
5. Insulin: Production of Insulin and Diabetes are closely related where insulin levels are to be maintained for healthy functioning of the individual. Insulin levels are usually maintained with the help of oral medicines or injections. For the dataset under discussion measurement of serum insulin is at 2-Hour intervals in units of mu/ml.
6. BMI: Body Mass Index is the ratio of weight in units of kg over their height in units of meter squared. BMI is computed for the individuals using this formula and the computed ratio is available for research.
7. Diabetes Pedigree Function: This indicates a function which score likelihood of diabetes based on family history. The numeric nature of this formulation enables prediction using machine learning algorithms.
8. Age: The age of the woman in years is available in the dataset.

Chapter 8: Implementation

The project methodology discussed is implemented in python using libraries for scientific research.

Diabetes Dataset Preparation: The Pima Indians Diabetes dataset is downloaded as a comma separated values file which is loaded into a pandas dataframe for usage. The features of the dataset are captured for future processing. The sample of dataset is shown in Figure No. 8.1.

```
df = pd.read_csv("/content/umi_dia.csv")
features = ['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
       'BMI', 'DiabetesPedigreeFunction', 'Age']
target = "Outcome"
dia_data = df[features]
dia_data.head()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	6	148	72	35	0	33.6	0.627	50
1	1	85	66	29	0	26.6	0.351	31
2	8	183	64	0	0	23.3	0.672	32
3	1	89	66	23	94	28.1	0.167	21
4	0	137	40	35	168	43.1	2.288	33

Figure No, 8.1: Dataset sample

Diabetes Dataset Cleaning: All null valued rows of the dataframe are deleted as shown in Figure No. 8.2.

```
dia_data.dropna()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	6	148	72	35	0	33.6	0.627	50
1	1	85	66	29	0	26.6	0.351	31
2	8	183	64	0	0	23.3	0.672	32
3	1	89	66	23	94	28.1	0.167	21
4	0	137	40	35	168	43.1	2.288	33
...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63
764	2	122	70	27	0	36.8	0.340	27
765	5	121	72	23	112	26.2	0.245	30
766	1	126	60	0	0	30.1	0.349	47
767	1	93	70	31	0	30.4	0.315	23

768 rows × 8 columns

Figure No 8.2: Dataset after null removal

Diabetes Dataset Processing: The dataframe is scrutinized if all the data types are in the format for further processing as shown in Figure No. 8.3.

```
dia_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 8 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
dtypes: float64(2), int64(6)
memory usage: 48.1 KB
```

Figure No. 8.3: Datatypes of the features

Exploratory Data Analysis: During the EDA of the dataset the mean, standard deviation of the numeric data from the dataset is computed. The minimum and maximum values is checked. The quartile for each feature is computed. The EDA is displayed as a table as shown in Figure No. 8.4.

```
dia_data.describe()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000

Figure No. 8.4: EDA of the dataset

Model building – kmeans: The already implemented kmeans algorithm from the well-known library of sci-kit learn which is used in data science and AI research, is used to build the model.

The number of clusters is given as input to build the kmeans model. The distance measure used in kmeans for this implementation is Euclidean distance.

Cluster Assignment: kmeans fit method from sci-kit learn library is used to activate the unsupervised learning algorithm to find patterns and form clusters.

Build Decision tree surrogate model: The type of decision tree classifier implemented in sci-kit learn is an optimized version of the CART algorithm. It supports only numeric data and the processed dataset consists of only numeric data.

Explain kmeans cluster assignment: The clustering using kmeans will result in cluster assignments to each row of the dataframe. These cluster assignments are the target for the Decision Tree Classifier. The Classifier when used to classify builds a decision tree where the leaf nodes are the kmeans cluster and the non-leaf nodes explain the impact of the features on
the cluster assignment.

Compute percentage of explainability: The accuracy of decision tree provides the percentage of explainability of the clustering. The accuracy of the classifier is given by metrics module of sci-kit learn.

Chapter 9: Analysis and Results

Hypothesis verification is done by conduction of experiment which is designed for this project. The two research hypothesis regarding the percentage of explainability is tested.

The different experiment variables are as given in Figure No. 9.1.

```
#Experiment Variables

num_of_clusters = [2,3,4,5,6,7,8,9]
dia_tree_criteria = ["gini", "entropy"]
topN = [1, 2, 3, 4, 5, 6, 7, 8]
depth = [2,3,4,5,6,7,8,9]
num_leaf = [4,5,6,7,8,9,10,11,12]
```

Figure No. 9.1: Experiment Variables

The experiment is conducted using the iterative structure shown in Figure No. 9.2.

```
for clust in num_of_clusters:

  km_model = KMeans(n_clusters = clust)
  km_target = km_model.fit_predict(dia_data)
  dia_data["cluster"] = km_target

  X = dia_data[features]
  y = dia_data["cluster"]
  labels = []
  for i in range(0, clust-1):
    labels.append(i)

  for crit in dia_tree_criteria:
    for num_ft in topN:
      for dep in depth:
        for leaf in num_leaf:
          exp_tree = DecisionTreeClassifier(criterion= crit, max_features = num_ft, max_depth = dep,max_leaf_nodes = lea
          tree_model = exp_tree.fit(X, y)
          tree_predict = tree_model.predict(X)
          acc = sklearn.metrics.accuracy_score(dia_data['cluster'],tree_predict)
          print("Accuracy:"+ str(acc)+ "===" +str(clust)+","+ str(crit)+","+  str(num_ft)+","+  str(dep)+","+  str(leaf)
          results.append(str(acc) +"," + str(clust)+","+ str(crit)+","+  str(num_ft)+","+  str(dep)+","+  str(leaf) +","
          res_frame.loc[len(res_frame)] = [acc,clust, crit, num_ft, dep, leaf]
```

Figure No. 9.2: Experiment iteration

The resulting accuracy for each row of the dataframe is saved to enable further analysis.

The accuracy of the classifier is an indication to the goodness of explanations. Hence the experiment results are plotted in Figure 9.3.

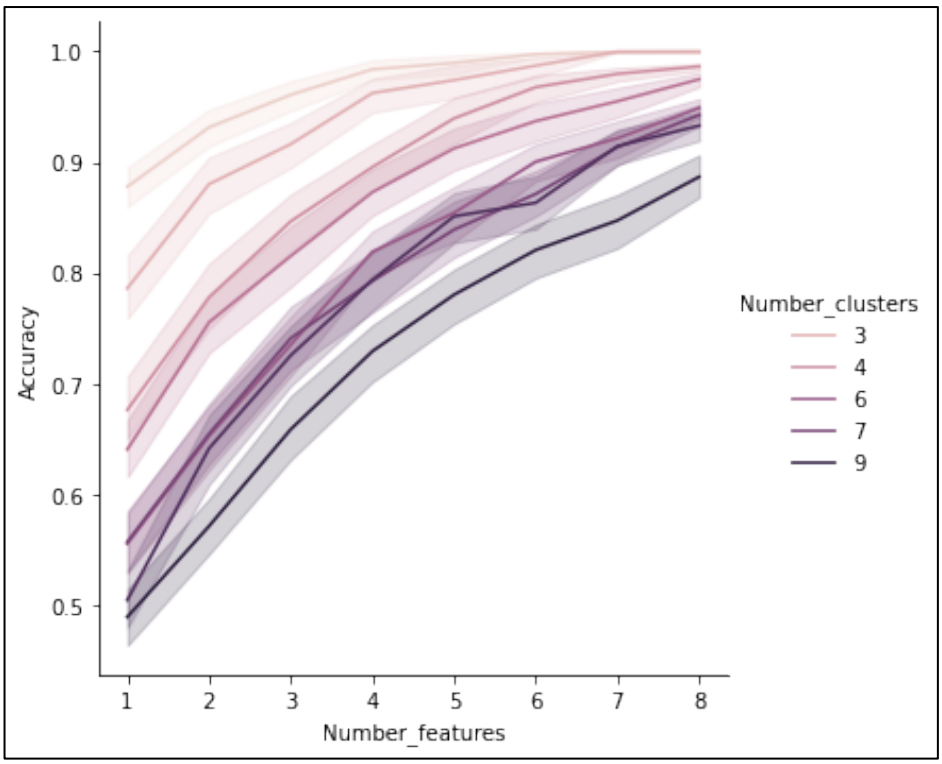


Figure No. 9.3: Accuracy plots of experiments for “gini” type of loss

When the number of features is less, correspondingly lesser number of clusters have higher percentage of explainability.

For 2 clusters, 1/8 features gives 90% explainability.
For 3 clusters, 1/8 features gives 80% explainability
For the entire domain of parameter selection for number of cluster at least 50% explainability is seen with just one feature.

The graph that is plotted above shows a very important observation:

Plateau of the curves as number of features increases keeping the number of clusters same.

The same set of experiments is conducted for “entropy” loss function and plotted in
Figure No. 9.4:

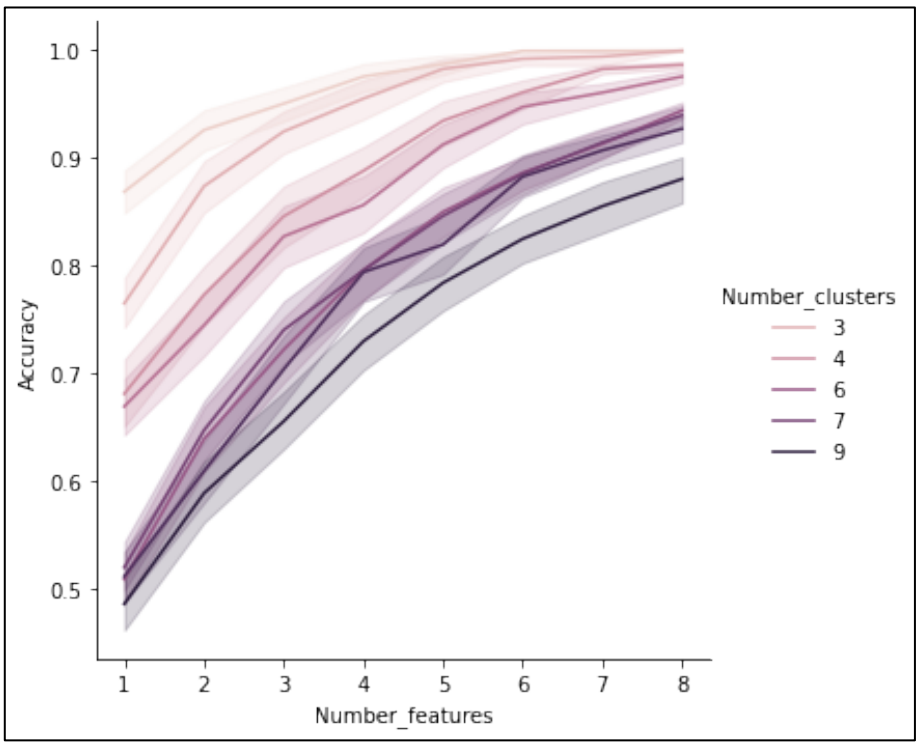


Figure 9.4: Accuracy plots of experiments for “entropy” type of loss

Similar pattern can be seen in “entropy” type of loss and hence we can conclude that the type of tree has very less bearing on the percentage of explainability. The research hypothesis that less than 50% of features are adequate to explain the cluster assignment is accepted. The Research hypothesis that less than 5 clusters of kmeans clustering is adequate to explain the variance of dataset is also accepted.

Chapter 10: Conclusions and Future Scope

This project started with an objective of exploring explanation mechanisms for unsupervised machine learning clustering techniques such as kmeans and the research conducted show how decision trees can be used as surrogate models for explainability. The high percentage of explainability of decision trees for kmeans is a testimony that this model is appropriate. The gap that was found in the literature review regarding the need for study on percentage of explainability is filled by this project. The variance of the dataset is considered and how much of the variance is explained by the surrogate model is measured and found to be encouraging.

The pride of using a dataset that is specific to India and conducting research for our mother country is noteworthy. This research work can be used by the healthcare industry for other ailments as well. AI in Healthcare is need of the hour and this research adds another brick to the wall. The important conclusion that a smaller number of features is adequate to explain the cluster opens to many doors of possibilities as each cost per feature capture in the Healthcare industry is high and sometimes the diagnostic itself can be expensive.

As a future scope of this project the effect of number of leaves and depth of decision tree on the percentage of explainability can be explored. Also, the mode of explanation delivery and interfaces for explanations have lot of room for innovation. Last but not the least, the application of these findings for a real-world practical use case such as a recommender system will be worth exploring.

Bibliography

[1] P. Mishra, Practical Explainable AI Using Python: Artificial Intelligence Model Explanations Using Python-based Libraries, Extensions, and Frameworks. Apress, 2021.

[2] D. Rothman, Hands-On Explainable AI (XAI) with Python: Interpret, visualize, explain, and integrate reliable AI for fair, secure, and trustworthy AI apps. Packt Publishing, 2020.

[3] A. Bhattacharya, Applied Machine Learning Explainability Techniques. Packt Publishing,2022.

[4] K. R. Varshney, Trustworthy Machine Learning. Chappaqua, NY, USA: Independently Published, 2022.

[5] Moshkovitz, M., Dasgupta, S., Rashtchian, C., & Frost, N. (2020, November). Explainable k-means and k-medians clustering. In International conference on machine learning (pp. 7055-7065). PMLR.

[6]Frost, N., Moshkovitz, M., & Rashtchian, C. (2020). ExKMC: Expanding Explainable k -Means Clustering. arXiv preprint arXiv:2006.02399.

[7] Makarychev, K., & Shan, L. (2022, June). Explainable k-means: don't be greedy, plant bigger trees!. In Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing (pp. 1629-1642).

[8] Makarychev, K., & Shan, L. (2021, July). Near-optimal algorithms for explainable k-medians and k-means. In International Conference on Machine Learning (pp. 7358-7367). PMLR.

[9] Charikar, M., & Hu, L. (2022). Near-optimal explainable k-means for all dimensions. In Proceedings of the 2022 Annual ACM-SIAM Symposium on

Discrete Algorithms (SODA) (pp. 2580-2606). Society for Industrial and Applied Mathematics.
[10] Dasgupta, S., Frost, N., Moshkovitz, M., & Rashtchian, C. (2020, July). Explainable k-means clustering: theory and practice. In XXAI Workshop. ICML.

[11] Laber, E., Murtinho, L., & Oliveira, F. (2021). Shallow decision trees for explainable k -means clustering. arXiv preprint arXiv:2112.14718.

[12] Gamlath, B., Jia, X., Polak, A., & Svensson, O. (2021). Nearly-tight and oblivious algorithms for explainable clustering. Advances in Neural Information Processing Systems, 34, 28929-28939.

[13] Laber, E. S., & Murtinho, L. (2021, July). On the price of explainability for some clustering problems. In International Conference on Machine Learning (pp. 5915-5925). PMLR

[14] Bandyapadhyay, S., Fomin, F., Golovach, P. A., Lochet, W., Purohit, N., & Simonov, K. (2022, June). How to Find a Good Explanation for Clustering?. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 4, pp. 3904-3912)

[15] Baralis, E., Pastor, D. E., & Cannone, M. Explainable AI for Clustering Algorithms.

[16] Kauffmann, J., Esders, M., Ruff, L., Montavon, G., Samek, W., & Müller, K. R. (2022). From clustering to cluster explanations via neural networks. IEEE Transactions on Neural Networks and Learning Systems.

[17] S. M. Lundberg και S.-I. Lee, ‘A Unified Approach to Interpreting Model Predictions’, Advances in Neural Information Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan,R. Garnett, Curran Associates, Inc., 2017, pp. 4765–4774.

[18] “LIME user manual LIME documentation,” https://lime.readthedocs.io/en/latest/usermanual.html#introduction (accessed Aug. 03, 2022).

[19] A. B. Arrieta κ.ά., ‘Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI’. arXiv, 2019.

[20] Cameron Craddock, Yassine Benhajali, Carlton Chu, Francois Chouinard, Alan Evans, András Jakab, Budhachandra Singh Khundrakpam, John David Lewis, Qingyang Li, Michael Milham, Chaogan Yan, Pierre Bellec (2013). The Neuro Bureau Preprocessing Initiative: open sharing of preprocessed neuroimaging data and derivatives. In Neuroinformatics 2013, Stockholm, Sweden

[21]Karegowda, A. G., Punya, V., Jayaram, M. A., & Manjunath, A. S. (2012). Rule based classification for diabetic patients using cascaded k-means and decision tree C4. 5. International Journal of Computer Applications, 45(12), 45-50.

[22] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S. I. (2019). Explainable AI for trees: From local explanations to global understanding. arXiv preprint arXiv:1905.04610.

[23] Molnar, C. Interpretable Machine Learning. Lulu. com, 2019. https://christophm.github.io/
interpretable-ml-book/.

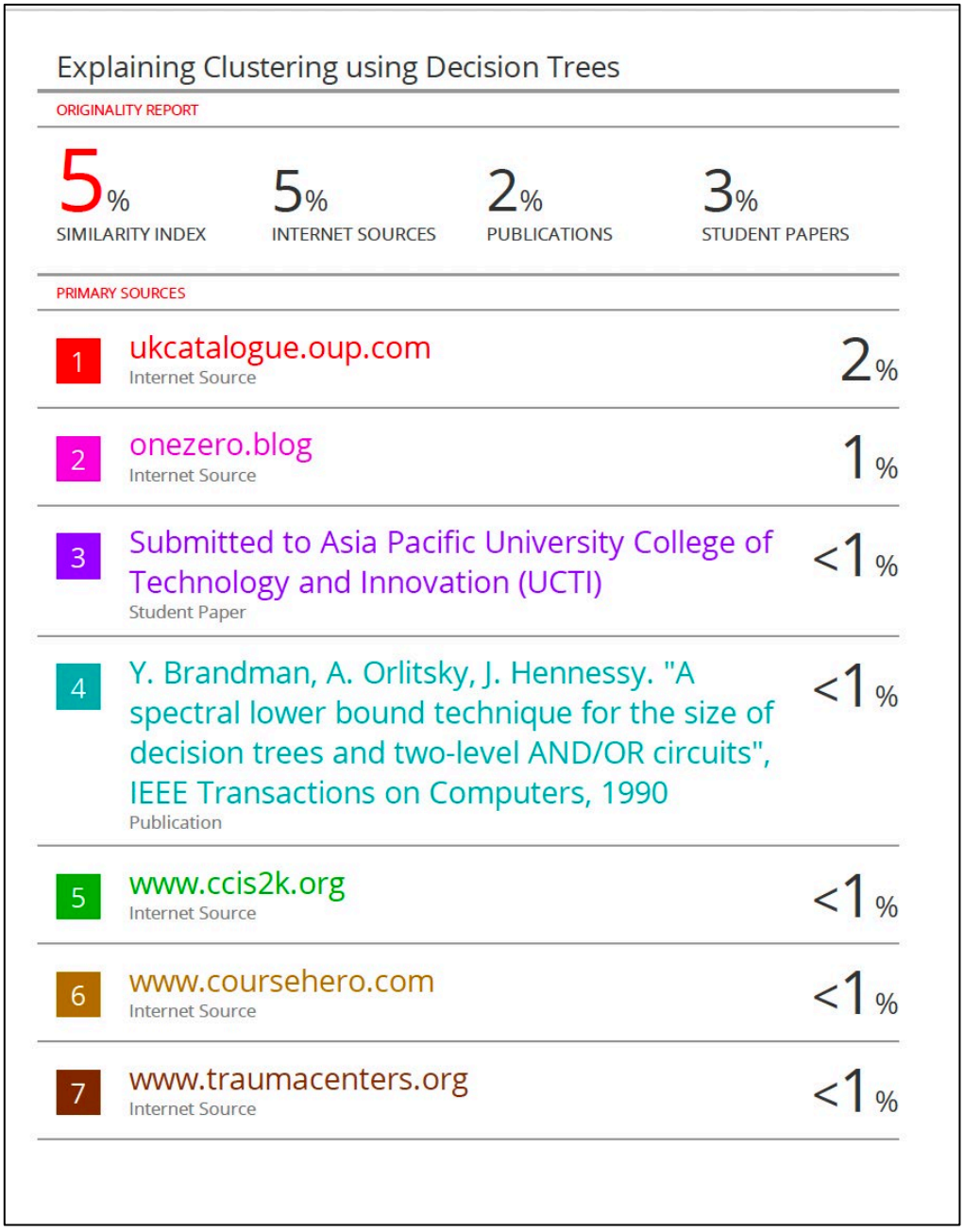
[24] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, και B. Yu, ‘Definitions, methods, and applications in interpretable machine learning’, Proceedings of the National Academy of Sciences, τ. 116, τχ. 44, σσ. 22071–22080, Οκτωβρίου 2019.

[25]Joshi RD, Dhakal CK. Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches. Int J Environ Res Public Health. 2021 Jul 9;18(14):7346. doi: 10.3390/ijerph18147346. PMID: 34299797; PMCID: PMC8306487.

[26] Smith JW, Everhart JE, Dickson WC, Knowler WC, Johannes RS. Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. Proc Annu Symp Comput Appl Med Care. 1988 Nov 9:261–5. PMCID: PMC2245318

Appendix

Plagiarism Report



Explaining Clustering using Decision Trees

ORIGINALITY REPORT

5%	5%	2%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ukcatalogue.oup.com Internet Source	2%
2	onezero.blog Internet Source	1%
3	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	<1%
4	Y. Brandman, A. Orlitsky, J. Hennessy. "A spectral lower bound technique for the size of decision trees and two-level AND/OR circuits", IEEE Transactions on Computers, 1990 Publication	<1%
5	www.ccis2k.org Internet Source	<1%
6	www.coursehero.com Internet Source	<1%
7	www.traumacenters.org Internet Source	<1%