جامعة الملك سعود
King Saud University
1957

KIDS HOBBY PREDICTION DATASET

Data Mining- 2023

PROBLEM

- Understanding children's interests and guiding them towards suitable hobbies can be challenging for parents.

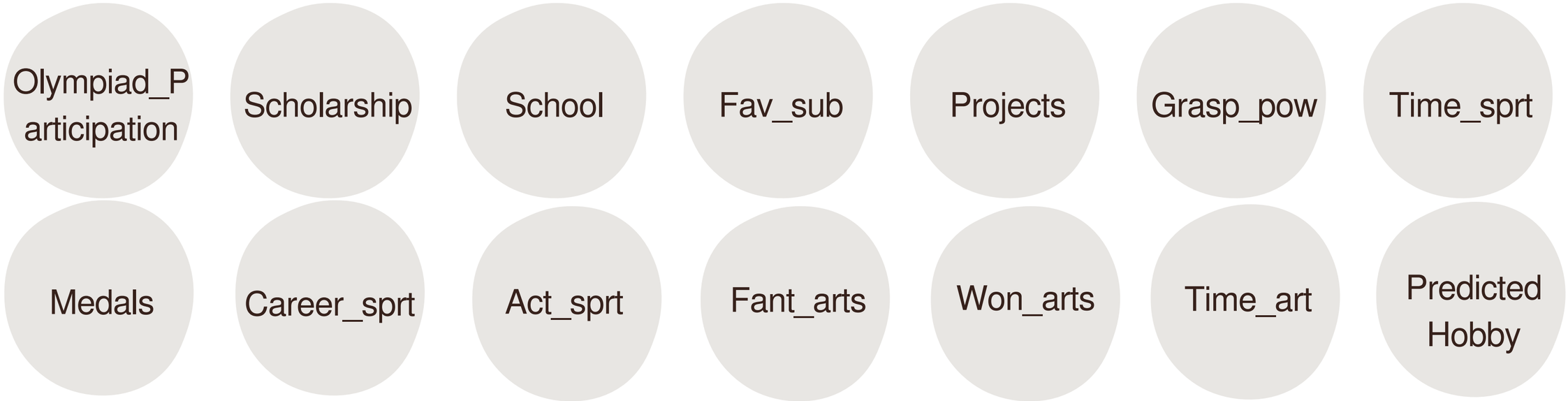
- In our analysis of the 'Hobby Kids' dataset, we aim to identify hobbies that match children's interests. This helps parents better guide their kids towards activities they'll enjoy.

KIDS HOBBY DATASET

We applied data mining for kids hobby dataset that have

Number of objects: 1601

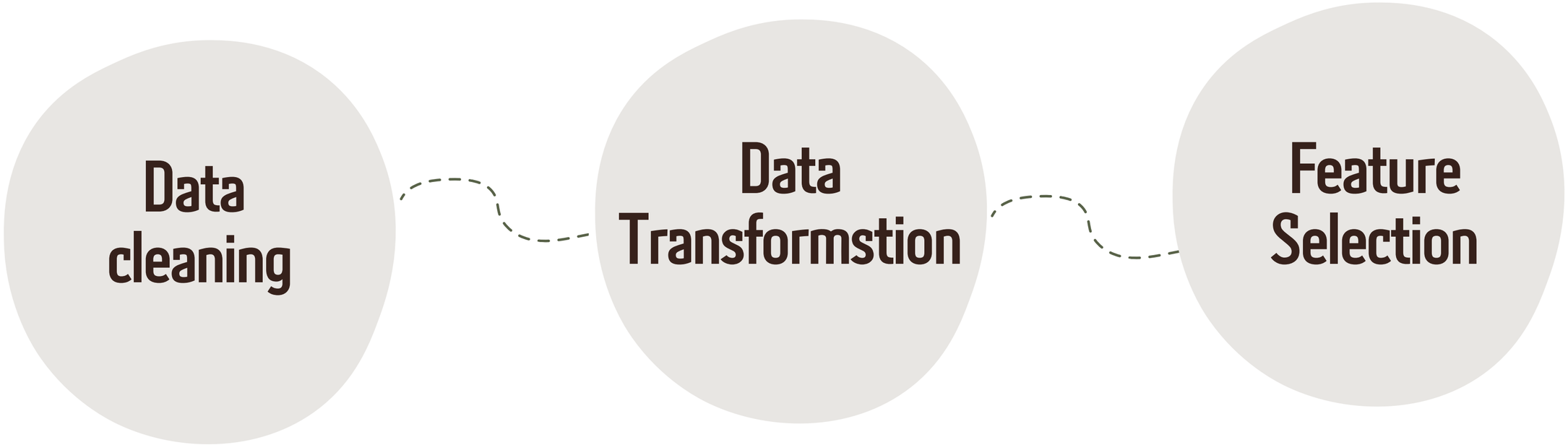
Number of attributes: 14



The source: https://www.kaggle.com/datasets/abtabm/hobby-prediction-basic

DATA PREPROCESSING

- To enhanced quality of results we tried to apply different tasks of data Preprocessing



DATA CLEANING

We simply looked for missing values, According to our investigation, the dataset does not contain any outliers since it doesn't have a numerical data type.

Additionally, there are no inconsistent values or other errors.

```{r}
sum(is.na(Hobby_Data))
```

```
[1] 0
```

DATA TRANSFORMSTION

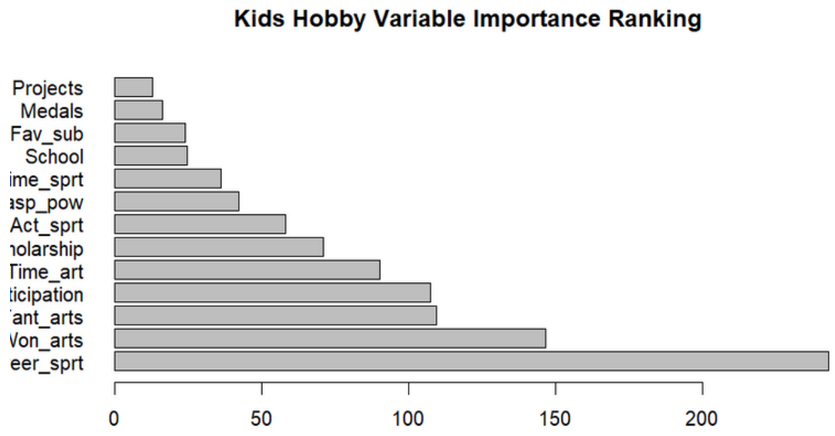
- We don't need to use normalization and discetization in our dataset. Since our dataset doesn't have numeric attributes and normalization involves mathematical operations.

Encoding

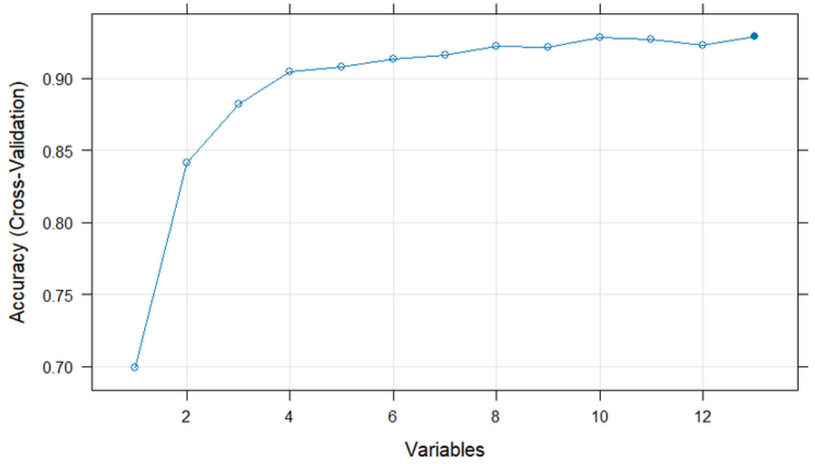
- We applied encoding for categorical and boolean data to preparing the data for analysis by making it consistent and suitable for a variety of analytical techniques.

FEATURE SELECTION

- we utilize feature selection techniques. These techniques enable us to eliminate redundant or irrelevant attributes from the dataset
- we used two feature selection methods:



Rank Features by Importance



Recursive Feature Elimination (RFE).

DATA MINING TECHNIQUES

- we used two data mining tasks

CLASSIFICATION

Supervised learning

CLUSTRING

Unsupervised learning

CLASSIFICATION

- Classification as part of supervised learning, we will apply a classification algorithm to assign each data point into predefined categories based on its attributes. This involves selecting the most relevant features that have been cleaned and formatted during preprocessing

- we trained our model to be able to predict kids hobby if it is: (Academics,Arts,Sports).

CLASSIFICATION

- To bulild our modlel we creating a decision tree algorithm which is a recursive algorithm produces a tree with a leaf node representing the final decision, using three measure selection (Information gain-Gini index-Gain ratio).

- The final decision for class “Predicted Hobby” is either (Academics,Arts,Sports).

the prediction is based on set of attributes :

(Scholarship , Fav_sub, Projects, Grasp_pow , Time_sprt , Career_sprt , Act_sprt , Fant_arts , Won_arts , Time_art ,Olympiad_Participation)

CLASSIFICATION

In this Technique Data set dividing to :

- Training set: help to creating a decision tree.
- Testing set: used to evaluate the result model.

We tried 3 different sizes of Training and Testing sets :

1

70%Training set
30%Testing set

2

80%Training set
20%Testing set

3

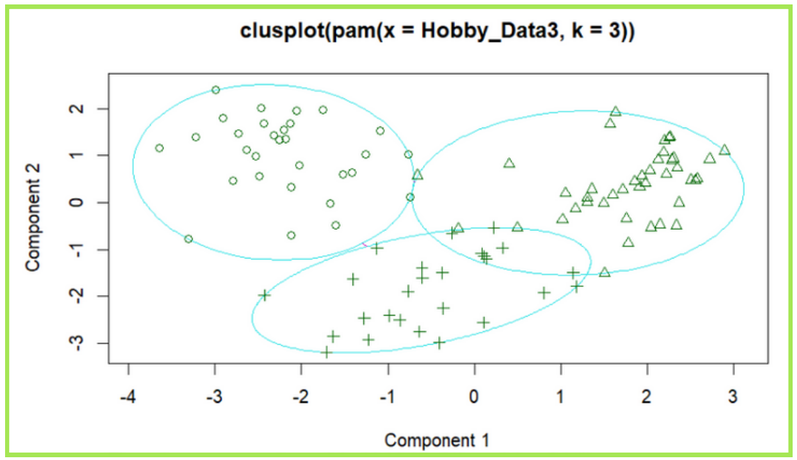
90%Training set
10%Testing set

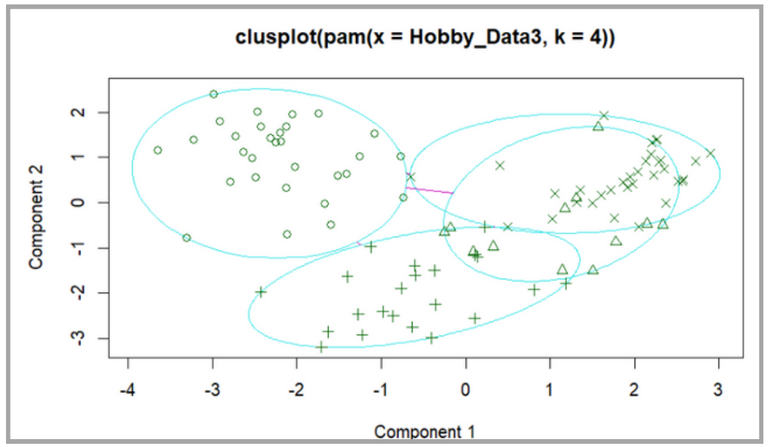
CLUSTRING

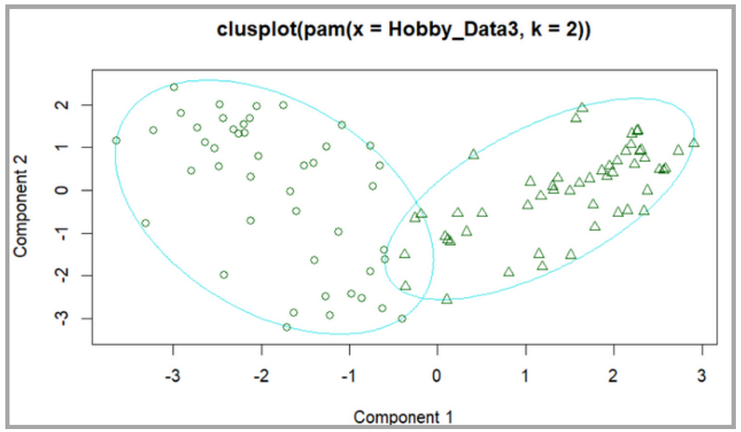
- We using k-medoids instead of k-mean
- Partitioning algorithms aim to partition a dataset into a predetermined number of clusters, where each data point belongs to exactly one cluster.
- K-medoids is a variation of the popular k-means algorithm, but instead of using centroids (mean values) to represent the clusters, it uses medoids (actual data points) as the cluster representatives.

CLUSTRING

- We choose 3 different number of cluster(2 , 3 ,4) based on Elbow method, Silhouette coefficient, silhouette score
- k=3 seems to be a reasonable choice. It has a good visual and silhouette width, besides its BCubed Precision and Recall values strike a balance.







RESULTS AND FINDINDS

- In classification the best attribute selection measures was Gini index with a 90% of training set and 10% for testing.
- Cluster using k-means algorithm it is not applicable to our data.
- Alternative clustering techniques, specifically designed for categorical data, such as partitioning around medoids are used.
- we tried different values of K, such as 2,3,and 4. The best was 3.
- Classification is better than clustering since it provides more effective prediction.

Do you Have Any Questions?

THANK YOU FOR LISTENING

MADE BY:

Wiam Baalahtar	443200416
Dana Aldawood	443200510
Reema Alkhaldi	443201003
Hind Alhijailan	443200971

Supervised By: D. Hanan Altamimi