UAS Machine Learning

Name : Muhammad Choirul Imam

NIM : 2502395595

Due to uploads size limit, full answer and source code can be found in https://github.com/mcimam/rplay-ml

1. Regression

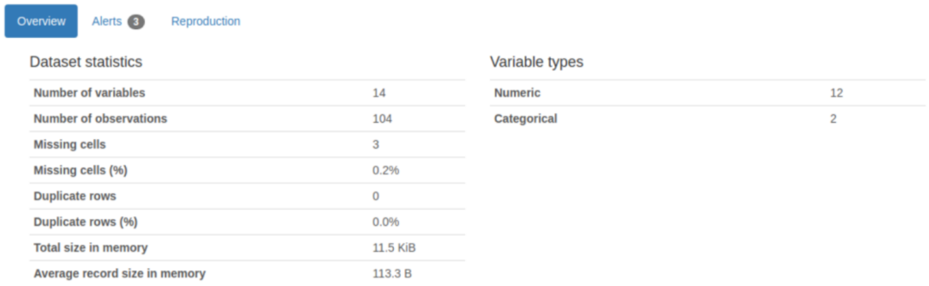
Dataset

Dataset that used in this problem is dataset about possum animal that can be found publicly in https://www.kaggle.com/datasets/abrambeyer/openintro-possum. The regression model created this dataset aims to predict possum's age based on 11 parameters:

- Population category
- Gender
- Head length
- Skull width
- Total length
- Tail length
- Foot length
- Ear length
- Distance between eyes
- Chest girth
- Belly girth

Explanatory Data Analysis

First step, explanatory data analysis is conducted. Image 1 shows an overview of datasets. There are 14 parameters in total in this dataset and 104 records. From that 14, 12 are numerical where the rest are categorical. Of those 14 parameters, case has unique value and uniformly distributed, indicating that it's an id of data.



Dataset statistics

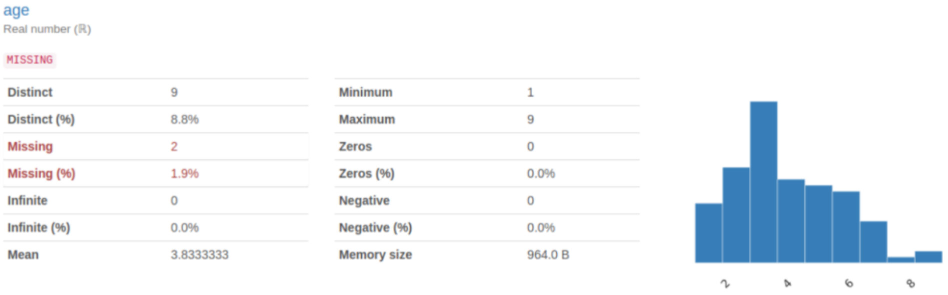
Number of variables	14
Number of observations	104
Missing cells	3
Missing cells (%)	0.2%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	11.5 KiB
Average record size in memory	113.3 B

Variable types

Numeric	12
Categorical	2

Image 1.1. Overview of Datasets

Analysis is carried out to all parameters of datasets one by one. For example, in age parameters, we detect 2 missing. As for the age, it varied from 1 to 9 years old. With details as shown in image 1.2. Age also has 2 missing values or 1.9% from overall data. As you can see from image 1.2, age doesn't imbalance dataset. Sample of some first data also view as in image 1.3.



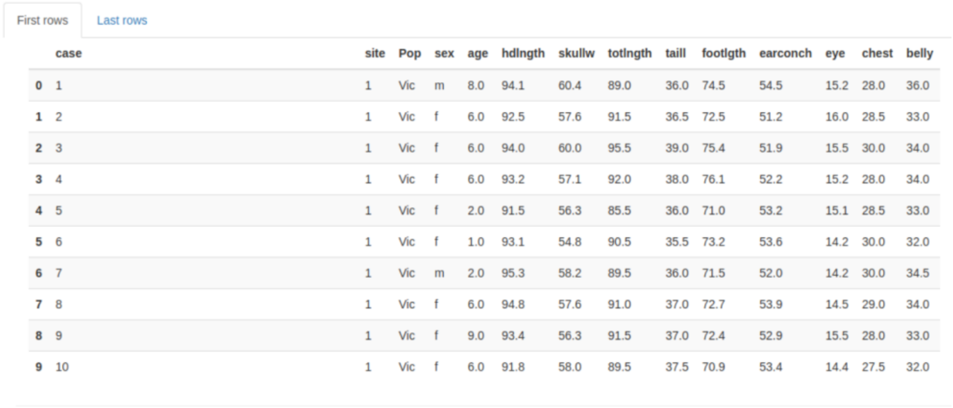
age
Real number (ℝ)

MISSING

Distinct	9
Distinct (%)	8.8%
Missing	2
Missing (%)	1.9%
Infinite	0
Infinite (%)	0.0%
Mean	3.8333333

Minimum	1
Maximum	9
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	964.0 B

Image 1.2. Age analysis



First rows | Last rows

	case	site	Pop	sex	age	hdlngth	skullw	totlngth	taill	footlgth	earconch	eye	chest	belly
0	1	1	Vic	m	8.0	94.1	60.4	89.0	36.0	74.5	54.5	15.2	28.0	36.0
1	2	1	Vic	f	6.0	92.5	57.6	91.5	36.5	72.5	51.2	16.0	28.5	33.0
2	3	1	Vic	f	6.0	94.0	60.0	95.5	39.0	75.4	51.9	15.5	30.0	34.0
3	4	1	Vic	f	6.0	93.2	57.1	92.0	38.0	76.1	52.2	15.2	28.0	34.0
4	5	1	Vic	f	2.0	91.5	56.3	85.5	36.0	71.0	53.2	15.1	28.5	33.0
5	6	1	Vic	f	1.0	93.1	54.8	90.5	35.5	73.2	53.6	14.2	30.0	32.0
6	7	1	Vic	m	2.0	95.3	58.2	89.5	36.0	71.5	52.0	14.2	30.0	34.5
7	8	1	Vic	f	6.0	94.8	57.6	91.0	37.0	72.7	53.9	14.5	29.0	34.0
8	9	1	Vic	f	9.0	93.4	56.3	91.5	37.0	72.4	52.9	15.5	28.0	33.0
9	10	1	Vic	f	6.0	91.8	58.0	89.5	37.5	70.9	53.4	14.4	27.5	32.0

Image 1.3. Sample Data

Aside from analysis for each individual column. Relations between each column are checked too. Image 1.4 shows an example of interaction between age and chest width. From this plot of interaction, patterns start to emerge.

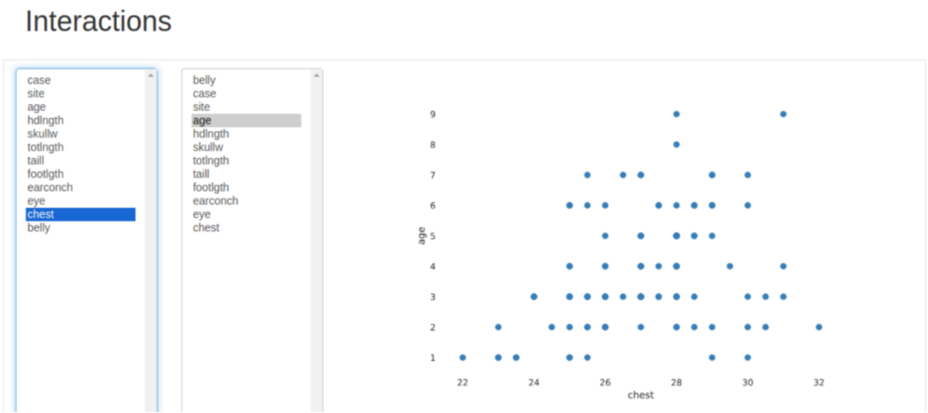


Image 1.4. Interaction between data

Data Preprocessing

From our explanatory data analysis, some steps need to be carried out before data can be processed. The process that needs to work on are:

- Feature selection

  Of the 14 parameters available, only 11 are being used. The reason is varied. Case parameters is an id, Pop and sex are both categoric value. While the rest is plainly numerical value with real number. Leaving all 11 of parameters.

- Solve missing value

  Considering that only 1.9% of overall records are missing, we can safely delete the data and consider it as invalid record.

  ```
  df = df.dropna()
  ```

- Normalization

  After all missing values are solved, we need to solve normalization. This step is necessary to ensure that all features have fair treatment, and all features are within comparable range. Considering that the data is continuous value, it decided that Standards Scaler. This makes data have a mean of 0 – and standard deviation of 1.

  ```
  ss = StandardScaler()

  for i in df.drop(columns=['age']).columns.to_list():

  df[i] = ss.fit_transform(df[[i]])
  ```

- Data split

We split dataset to 2 trains and test with ration of 8:2. Chosen dataset is split randomly. Train dataset will be used to fit model while test is being used as test dataset to measure model performance.

```
x_train, x_test, y_train, y_test = train_test_split(df.drop(columns=['age']), df[['age']], test_size = 0.2, random_state=2)
```

Regression Results

Two algorithms are being used to make regression models, namely linier regression and decision tree regression. Linier regression is chosen due to its simplicity to handle linier regression. While decision tree is chosen mainly because of its ability to handle more complex and non-linear regression.

After training both data with training data (x_train), we predict it using x_test to get y_pred variable. Than we compare y_pred with y_test to see how it's performed. Image1.5 show comparison between y_pred and y_test.

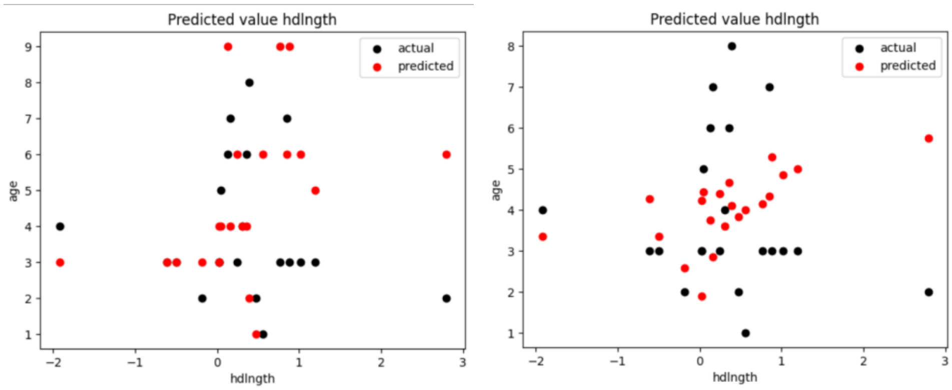


Image 1.5. Actual vs Predicted value of head length vs age using decision tree(left) and linier regression (right)

There are multiple matrix that can be used using prediction result and actual result, In this case, result of both regressions are measured in Mean Square Error and R2. (Table1.1). Mean Square Error can measure how big the error between predicted results and actual results. Lower value means lower error. From those results, we can conclude that Linier regression bring out better performance rather than decision tree regression. Mainly because easy to use model of linier regression.

Table 1.1. MSE and R2 Score

	Mean Square Error	R2 Score

Linier Regression	4.488	-0.298
Decision Tree Regression	9.476	-1.742

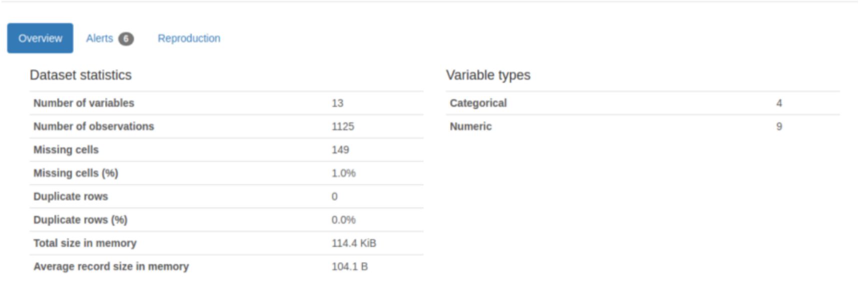
2. Classification

Dataset

Dataset that used in this question is https://www.kaggle.com/datasets/praveengovi/credit-risk-classification-dataset . The aim for this model is to predict customer's credit risk based on ten parameters explained in explanatory data analysis part.

Explanatory Data Analysis

Image 2.1 gives an overview of overall dataset condition. Nine features are numerical and continues value while the other four are categorical. 149 data are found missing within the dataset, but all come from one column named fea_2 (Image, 2.3).



Dataset statistics

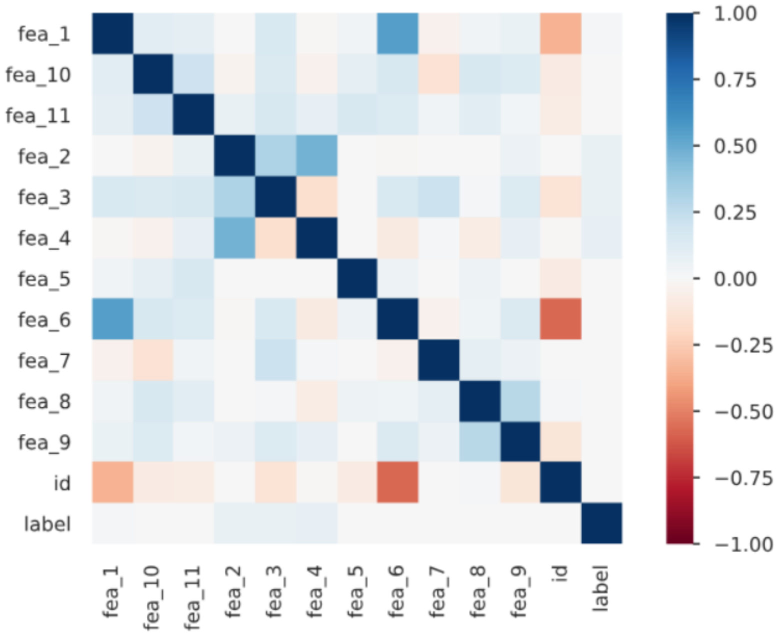
Number of variables	13
Number of observations	1125
Missing cells	149
Missing cells (%)	1.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	114.4 KiB
Average record size in memory	104.1 B

Variable types

Categorical	4
Numeric	9

Image 2.1. Overview Of Datasets

Next, Image 2.2 explains correlation between each variable. Fea_1 and fea_6 have high correlation with each other.



L

Image 2.1. Correlation between variables

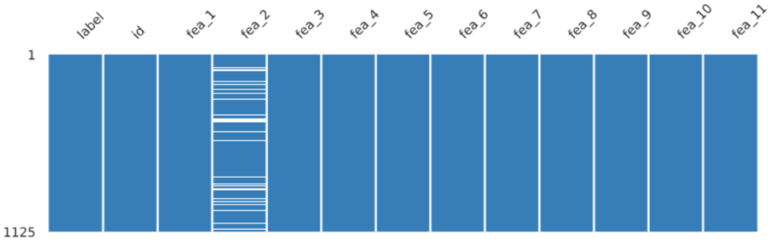
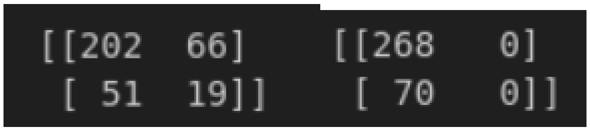


Image 2.3. Nullity matrix foreach variable

Classification Results

The algorithms used in classification are decision tree and Support Vector Machine (SVM). The result is Support Vector Machine (SVM) shown overall good results in accuracy (Table 2.1). However, SVM shown is fall behind to guess false true negative based on confusion matrix result (Image 2.4)



```
[[202  66]     [[268   0]
 [ 51  19]]     [ 70   0]]
```

Image 2.4. Confusion Matrix (decision tree left; SVM right)

	Accuracy	Precision	Recall	F1 Score
Decision Tree	0.635	0.169	0.191	0.18
SVM	0.791	-	-	-

Table 2.1. Accuracy, Precession, Recall, F1 Score values

3. Clustering

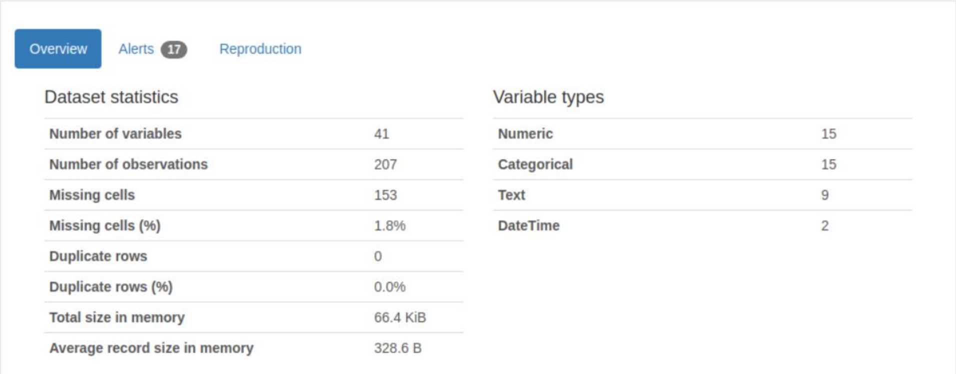
Dataset

Dataset that used in this problem is dataset about coffee quality can be found publicly in https://www.kaggle.com/datasets/volpatto/coffee-quality-database-from-cqi. The clustering is based on 12 parameters:

- 'Variety'
- 'Processing Method'
- 'Aroma',
- 'Flavor'
- 'Aftertaste'
- 'Acidity',
- 'Body'
- 'Balance','Overall'
- 'Moisture Percentage'
- 'Category One Defects'
- 'Category Two Defects'
- 'Color'

Explanatory Data Analysis

Image 3.1. show overall conditions oof this dataset. Of the 41parameters available, 9 of them are text, while the other 2 are datetime value. The rest are filtered to see which ones are probably relevant or have correlation with each other (Image 3.2).



Dataset statistics

Number of variables	41
Number of observations	207
Missing cells	153
Missing cells (%)	1.8%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	66.4 KiB
Average record size in memory	328.6 B

Variable types

Numeric	15
Categorical	15
Text	9
DateTime	2

Image 3.1. Overall Data

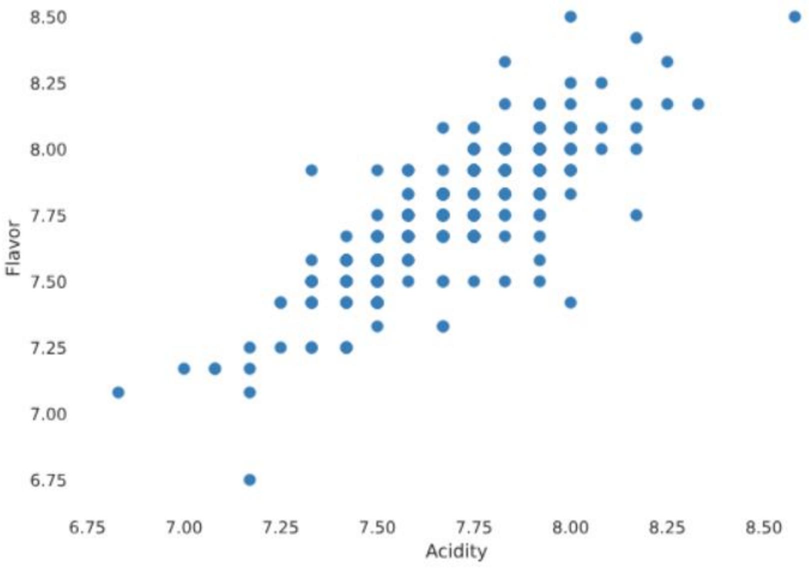


Image 3.2. Acidity-Flavor plot

Data Preprocessing

Based on mapping done in table 3.1, several data preprocessing are conducted. The first one is label encoding to turn categorical data into discrete numeric data. Label encoding is used due to no order really matter in categorical data.

Second, to fix imbalance, we MinMaxScaler is conducted across all real continuous data such as aroma, flavor, and aftertaste. This is done to counter algorithms that are sensitive to large data range.

Table 3.1. Parameters Used

Name	Type	Missing	Imbalance	Zero
Variety	Category	X		
Processing Method	Category	X	X	
Aroma	Real			
Flavor	Real			
Aftertaste	Real			
Acidity	Real			
Body	Real			
Balance	Real			
Moisture Percentage	Real			
Category One Defects	Real			X
Category Two Defects	Real			X
Color	Category			

Clustering

The algorithms that are used in this clustering are K-means and OPTICS. K-means is used because it can handle general clustering with widely use case. While optics is being used because it can handle bigger data. To find optimum K or cluster value, we use elbow method that measures within cluster minimum square. For clustering performance, we use silhouette score. Results are as follows (Image 3.3). To visualize data, dimensional reduction techniques is used such as pca (Image 3.4).

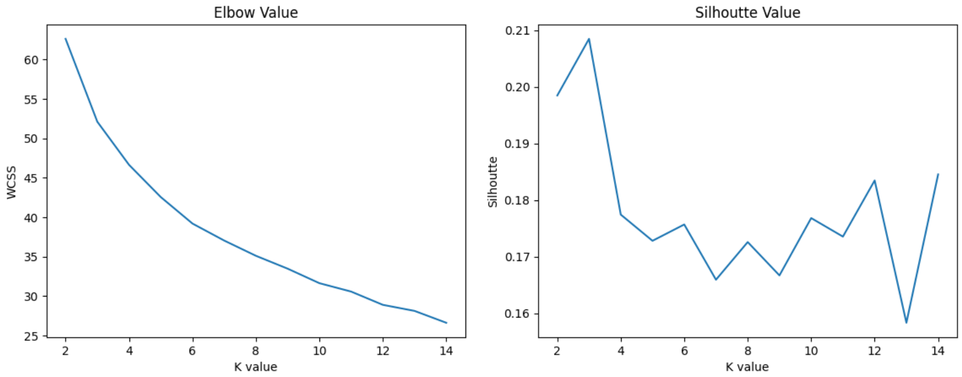


Image 3.3. Elbow and silhouette score compared to K value.

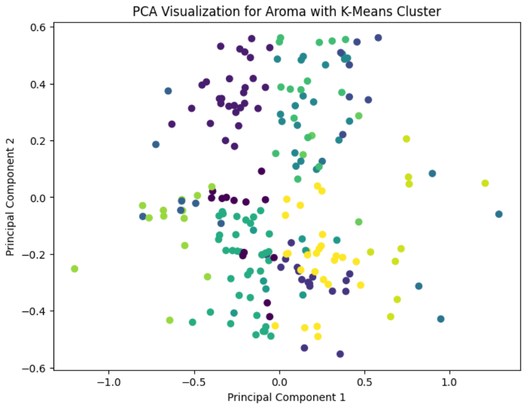


Image 3.4. PCS Visualization

Table 3.2. K value and Silhouette Score of K-means and OPTICS

	K-MEANS	OPTICS
K value	3	2
Silhouette Score	0.20877024996116478	-0.045546776438058115

4. Titanic Classification

Dataset

The main problem that needs to be solved is creating a model that predicts if passenger are survived or not survived based on Titanic disaster dataset. The dataset that used in titanic classification problem come from https://www.kaggle.com/datasets/yasserh/titanic-dataset .

Algorithm

Algorithm than being used in titanic classification is Decision tree classifier. First attempt, decision tree without any adjustment to its hyperparameter nor any pruning conducted result in more overfit model. Plotting the tree (image 4.1) explains why the model is overfit.

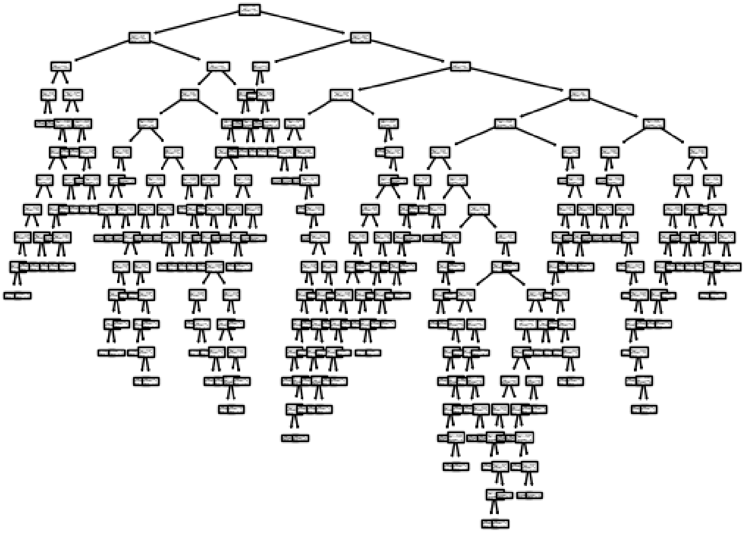


Image 4.1. Decision tree without pruning and limiter

To combat this problem, we tried to use cross validation technique, specifically grid search validation technique. With parameters grid of max_depth, min_samples_split, min_samples_leaf with accuracy as scoring metrics. The results are 3,1,1 as max_depth, min_samples_spli and min_samples_leaf respectively. Redrawing the tree explains that the model is much better (Image 4.2.)

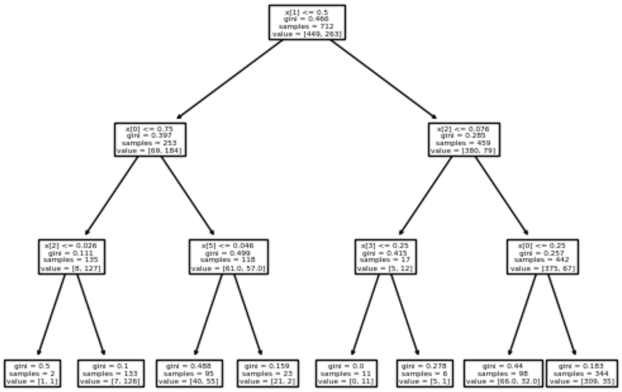

x[1] <= 0.5 gini = 0.466 samples = 712 value = [449, 263]
x[0] <= 0.75 gini = 0.397 samples = 253 value = [69, 184]
x[2] <= 0.076 gini = 0.285 samples = 459 value = [380, 79]
x[2] <= 0.026 gini = 0.111 samples = 135 value = [8, 127]
x[5] <= 0.046 gini = 0.499 samples = 118 value = [61.0, 57.0]
x[3] <= 0.25 gini = 0.415 samples = 17 value = [5, 12]
x[0] <= 0.25 gini = 0.257 samples = 442 value = [375, 67]
gini = 0.5 samples = 2 value = [1, 1]
gini = 0.1 samples = 133 value = [7, 126]
gini = 0.488 samples = 95 value = [40, 55]
gini = 0.159 samples = 23 value = [21, 2]
gini = 0.0 samples = 11 value = [0, 11]
gini = 0.278 samples = 6 value = [5, 1]
gini = 0.44 samples = 98 value = [66.0, 32.0]
gini = 0.183 samples = 344 value = [309, 35]

Image 4.2. Decision tree after changing hyperparameters.

Table 4.1. Test Results

Metrics	Results
Accuracy	0.6582

Precision	0.6582
Recall	0.6582
F1 Score	0.7428

Hyperparameters

Hyperparameters are defined as parameters that are used to control the training process of machine learning models. In this attempt, hyperparameters that being used are:

- Max Depth : Specify maximum depth of tree made
- Min Sample Split : Specify minimum sample need to split the leaf node
- Min Sample Leaf : Specify minimum sample inside leaf node

Bias vs Variance Tradeoff

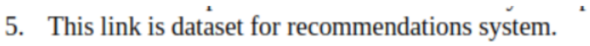
Bias in this case is defined as inability to capture true relationships for machine learning model. While variance is defined as sensitivity to the fluctuations of training datasets. In other words, a model needs to be fit well not underfit or overfit. Decision tree, model that applied in current problem is considered low bias high variance technique. Therefore, adjusting parameters i need to make balance between bias and variance. Furthermore, if need technique like pruning can be used to balance bias and variance.

- Max depth : lower value, higher bias
- Min sample split : higher value, higher bias
- Min sample leaf : higher value, higher bias

5. Recommendation System

Assumption

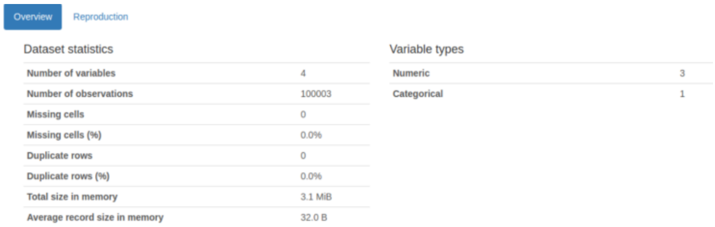
As no dataset being provided or empty link, we assume that we can use any public dataset available as in other question.



5. This link is dataset for recommendations system.

A) Dataset

The dataset come fromhttps://www.kaggle.com/datasets/dev0914sharma/dataset. The dataset consists of movie rating per customer. There are three parameters that are used, namely customer id, movie id, rating. The plan is to create a Collaborative recommendation system from this dataset. Image 5.1. an overview of dataset.



Dataset statistics

Number of variables	4
Number of observations	100003
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	3.1 MiB
Average record size in memory	32.0 B

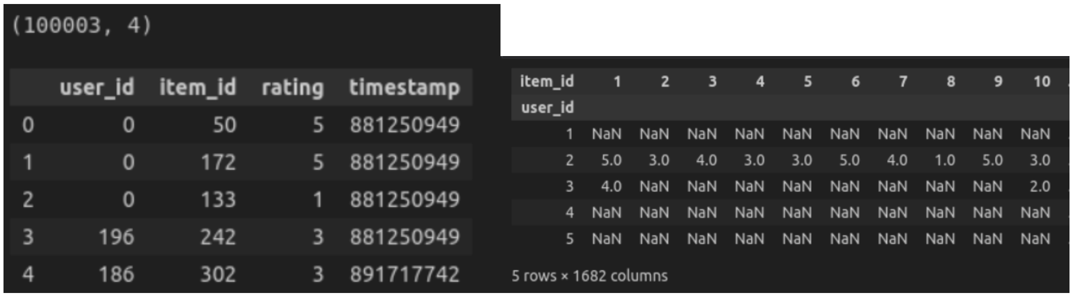
Variable types

Numeric	3
Categorical	1

Image 5.1. Overview of Dataset

Preprocessing

The preprocessing need for collaborative recommendation is to change/pivot the matrix to appropriate format, which is movie id as column, customer id as row and rating as its value. For non-existing rating, it just filled is not available value (Image 5.2).



```
(100003, 4)
```

	user_id	item_id	rating	timestamp
0	0	50	5	881250949
1	0	172	5	881250949
2	0	133	1	881250949
3	196	242	3	881250949
4	186	302	3	891717742

item_id	1	2	3	4	5	6	7	8	9	10
user_id										
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	5.0	3.0	4.0	3.0	3.0	5.0	4.0	1.0	5.0	3.0
3	4.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2.0
4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

5 rows × 1682 columns

Image 5.2. Sample data of before (left) and after (right) preprocess

B) Collaborative recommendation system

First, as other answer, explanatory data analysis carried out after pivoting the data. Simple example is drawn to scatter plot to see angle of vector between user choices (Image 5.3).

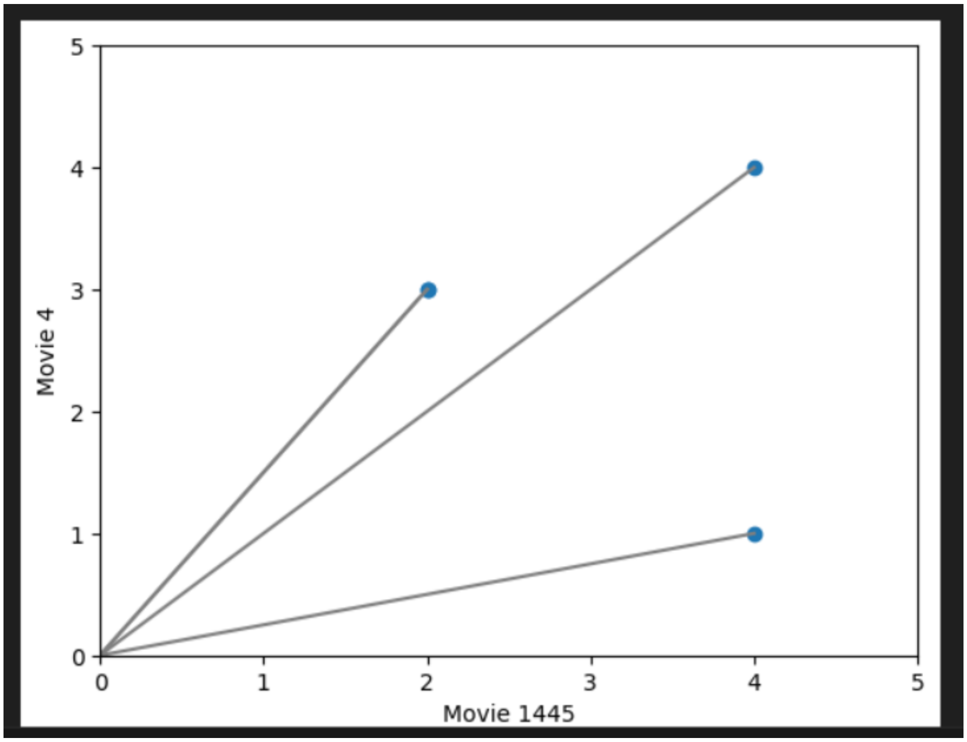


Image 5.3. Movie Rating for Different User

Singular Value Decomposition (SVD) is selected as the algorithm. Subsequently, a Grid Search Cross-Validation (GSCV) is performed to identify the optimal hyperparameters using a grid search over parameters such as n_epoch, lr_all, and reg_all. GSCV is executed with Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) as evaluation metrics. The maximum results are as follows:

```
{'n_epochs': 10, 'lr_all': 0.005, 'reg_all': 0.4}
```

- n_epoch: The number of iterations
- lr_all: The learning rate, how quick model learns
- reg_all: The regularization term during optimization

With that value, once again, SVD id train with dataset that split in half with ration of 8:2 between training and testing dataset. After that, test dataset is used to predict the result than compare it to actual data with RMSE and MAE. The results are 0.9553 and 0.7683 respectively. With higher RMSE and MAE value means worse accuracy. This indicates that this model is not that accurate.