(TitlePic)

CEIS Final Project

Kimberly Morrison | CEIS312 | 4/21/2023

Contents

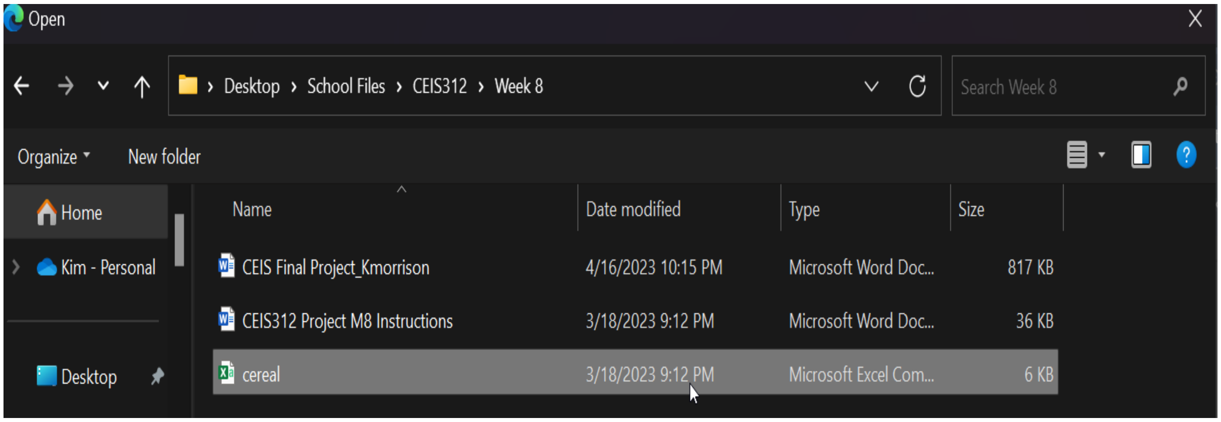
Introduction 2
Uploading dataset 2
Data preparation 3
Selecting features 4
Data Visualization 5
Linear regression model 6
Splitting data 7
Training the model 8
Scoring the model 8
Evaluating the model 9
Web Service 10
Challenges 11
Career skills 11
Conclusion 11
References 11

ıntroduction

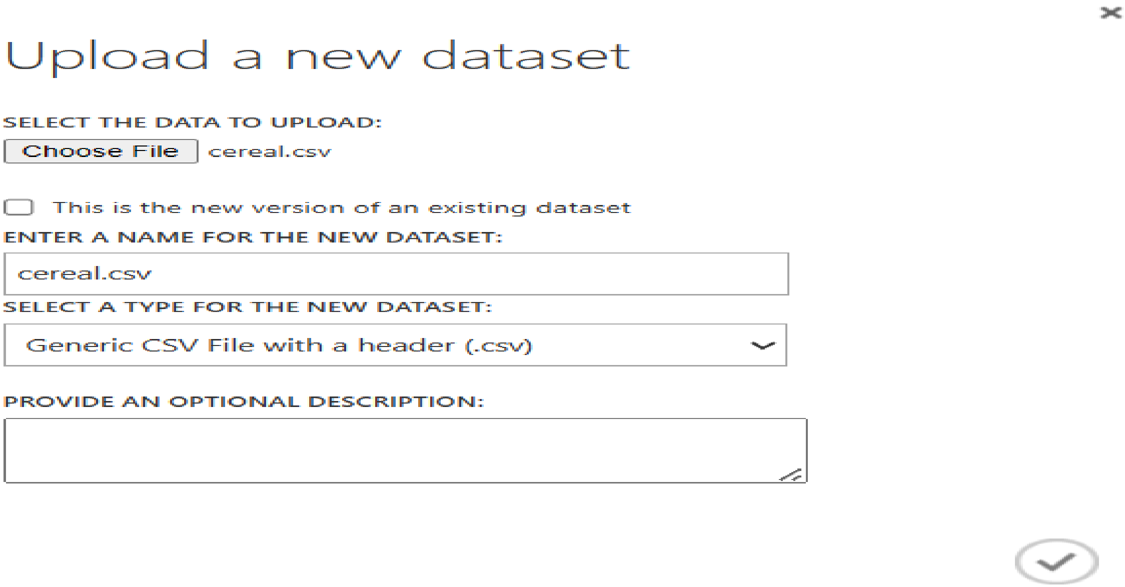
The purpose of this project is to use the data provided about cereals to perform an end-to-end exercise in creating and evaluating a ML model.

The end goal of this project is to use the data provided to predict which features of the cereal dataset are the most important to predict how customers choose their product(s).

Uploading dataset



Name	Date modified	Type	Size
CEIS Final Project_Kmorrison	4/16/2023 10:15 PM	Microsoft Word Doc...	817 KB
CEIS312 Project M8 Instructions	3/18/2023 9:12 PM	Microsoft Word Doc...	36 KB
cereal	3/18/2023 9:12 PM	Microsoft Excel Com...	6 KB



Upload a new dataset

SELECT THE DATA TO UPLOAD:

Choose File cereal.csv

□ This is the new version of an existing dataset

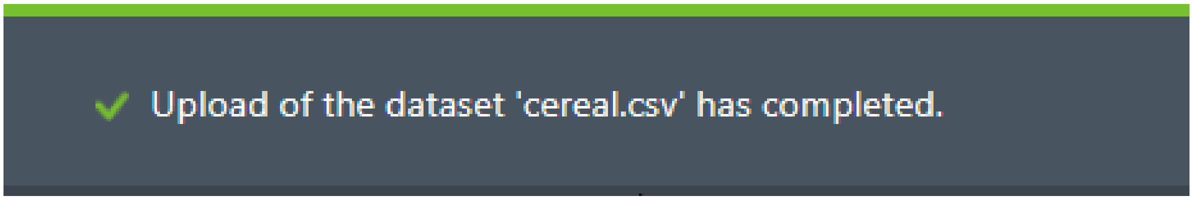
ENTER A NAME FOR THE NEW DATASET:

cereal.csv

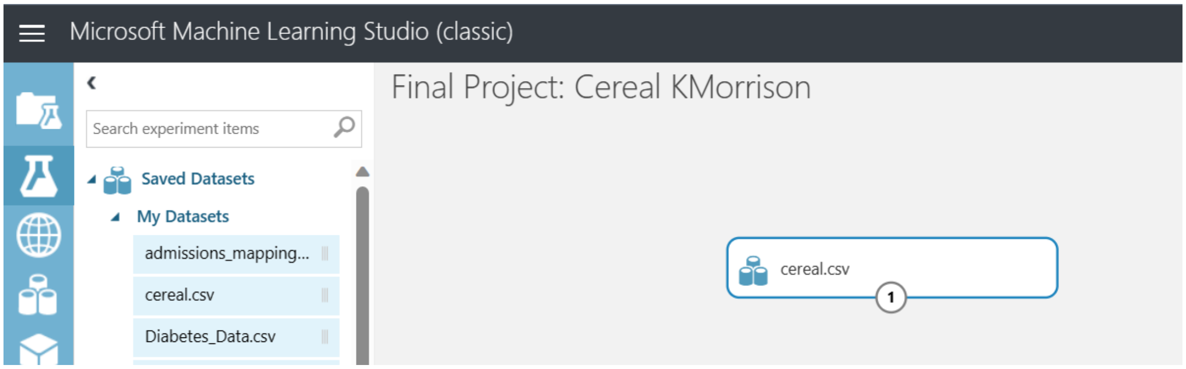
SELECT A TYPE FOR THE NEW DATASET:

Generic CSV File with a header (.csv)

PROVIDE AN OPTIONAL DESCRIPTION:

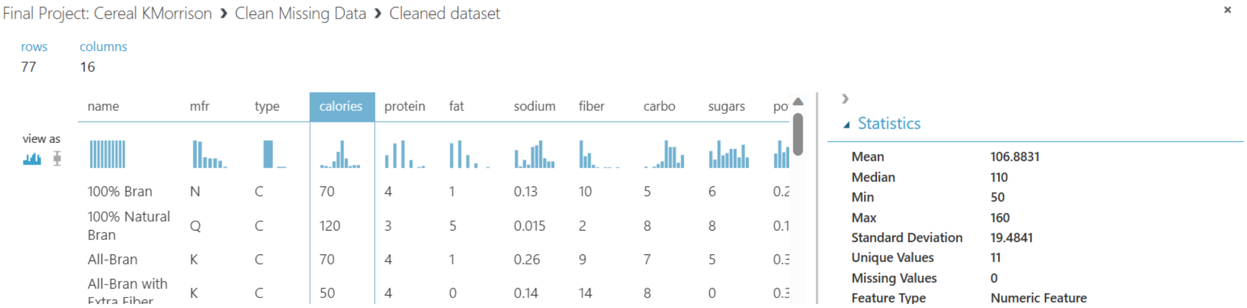


✓ Upload of the dataset 'cereal.csv' has completed.



Data preparation

- Remove missing values
  - There were none found



Final Project: Cereal KMorrison › Clean Missing Data › Cleaned dataset

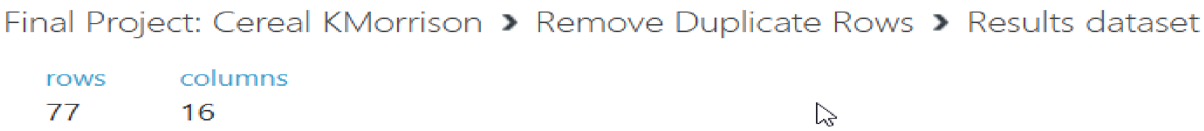
rows	columns
77	16

name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	po
100% Bran	N	C	70	4	1	0.13	10	5	6	0.2
100% Natural Bran	Q	C	120	3	5	0.015	2	8	8	0.1
All-Bran	K	C	70	4	1	0.26	9	7	5	0.3
All-Bran with Extra Fiber	K	C	50	4	0	0.14	14	8	0	0.3

Statistics

Mean	106.8831
Median	110
Min	50
Max	160
Standard Deviation	19.4841
Unique Values	11
Missing Values	0
Feature Type	Numeric Feature

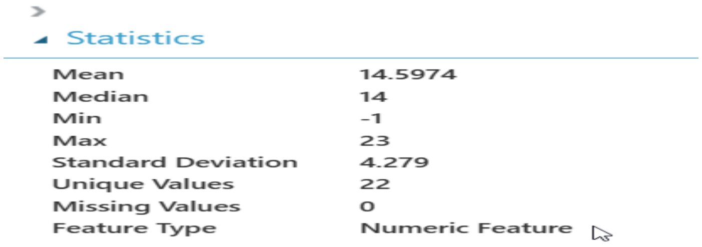
- Remove duplicate rows
  - There were none found.



Final Project: Cereal KMorrison › Remove Duplicate Rows › Results dataset

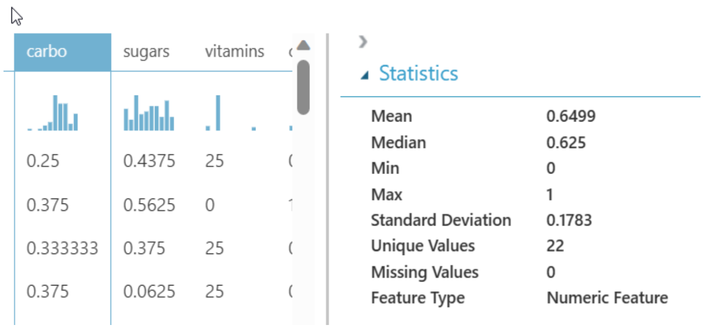
rows	columns
77	16

- Normalize Data
  - There are negative numbers in carbo and sugars



Statistics

Mean	14.5974
Median	14
Min	-1
Max	23
Standard Deviation	4.279
Unique Values	22
Missing Values	0
Feature Type	Numeric Feature

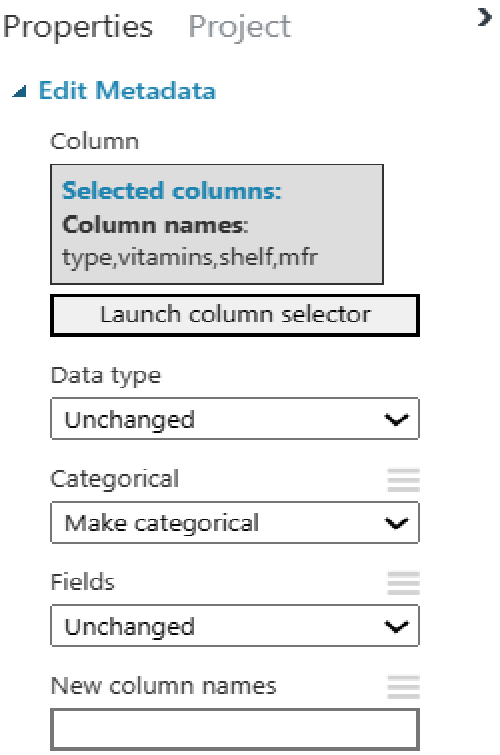


carbo	sugars	vitamins
0.25	0.4375	25
0.375	0.5625	0
0.333333	0.375	25
0.375	0.0625	25

Statistics

Mean	0.6499
Median	0.625
Min	0
Max	1
Standard Deviation	0.1783
Unique Values	22
Missing Values	0
Feature Type	Numeric Feature

- Edit Metadata – need to make the following categorical:
  - Mfr
  - Type
  - Vitamins
  - Shelf



Edit Metadata

Column

Selected columns:
Column names:
type,vitamins,shelf,mfr

Data type: Unchanged

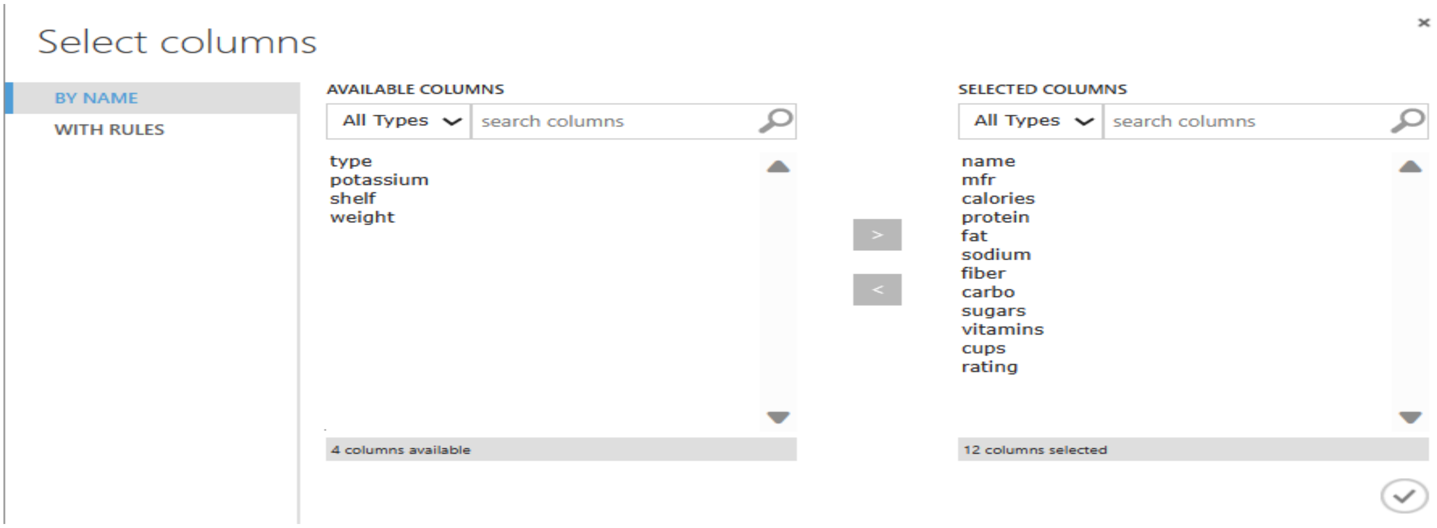
Categorical: Make categorical

Fields: Unchanged

New column names:

Selecting features

- Select columns from dataset.
  - Removed
    - Type – there are only 2 choices of the 77 (or 3% of the data) that are hot vs. cold
    - Potassium – this should be part of the vitamin stat
    - Shelf – this is the shelf the cereals are stocked on in the grocery store
    - Weight – this is the weight of the serving for each cereal. There are very few people, if any, who look at the weight of a serving rather than the serving size.



Select columns

BY NAME

WITH RULES

AVAILABLE COLUMNS

All Types

type
potassium
shelf
weight

4 columns available

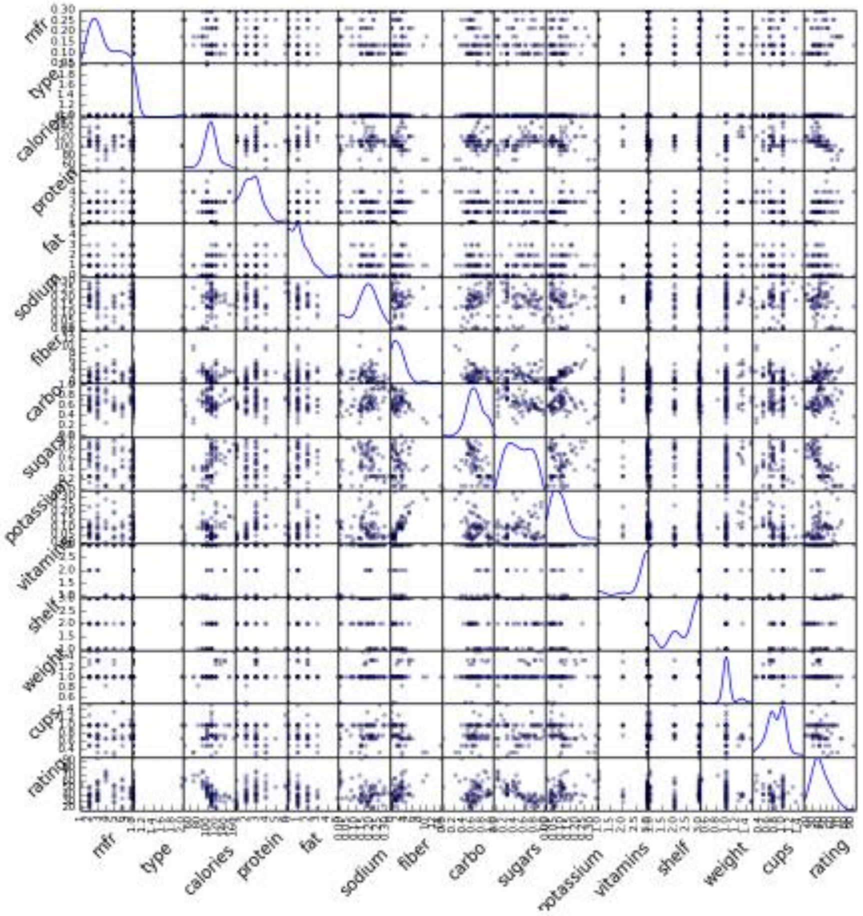
SELECTED COLUMNS

All Types

name
mfr
calories
protein
fat
sodium
fiber
carbo
sugars
vitamins
cups
rating

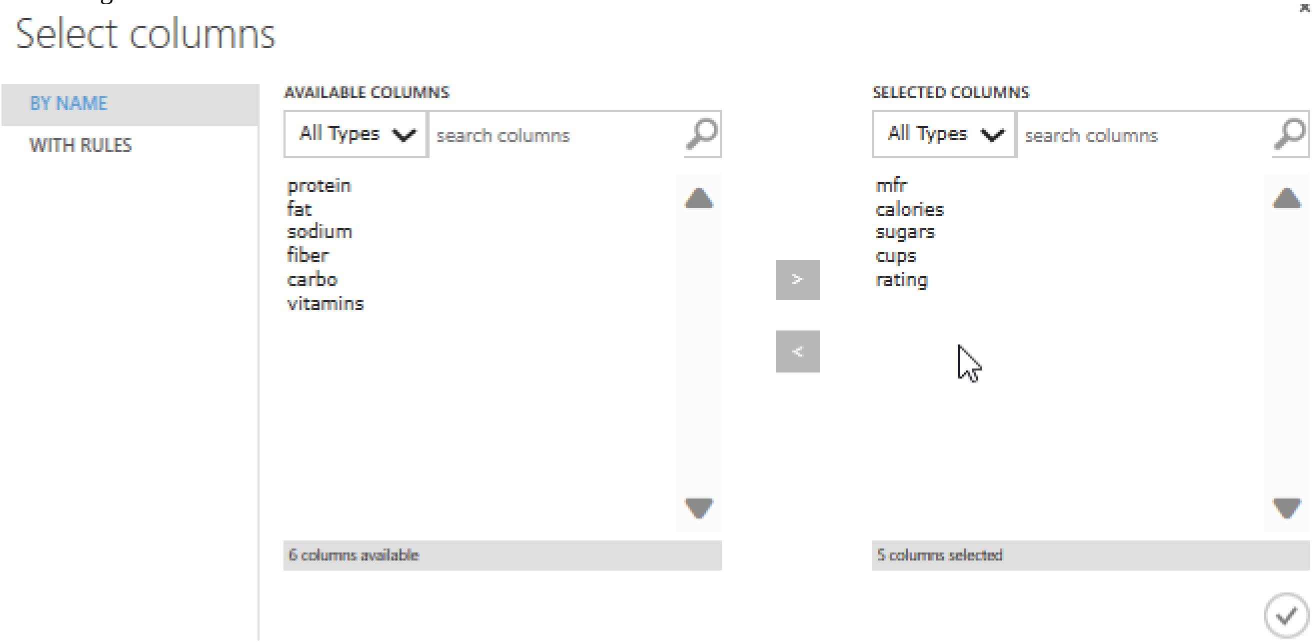
12 columns selected

Data Visualization



Linear regression model

- Linear Regression



Select columns

BY NAME

WITH RULES

AVAILABLE COLUMNS

All Types | search columns

protein
fat
sodium
fiber
carbo
vitamins

6 columns available

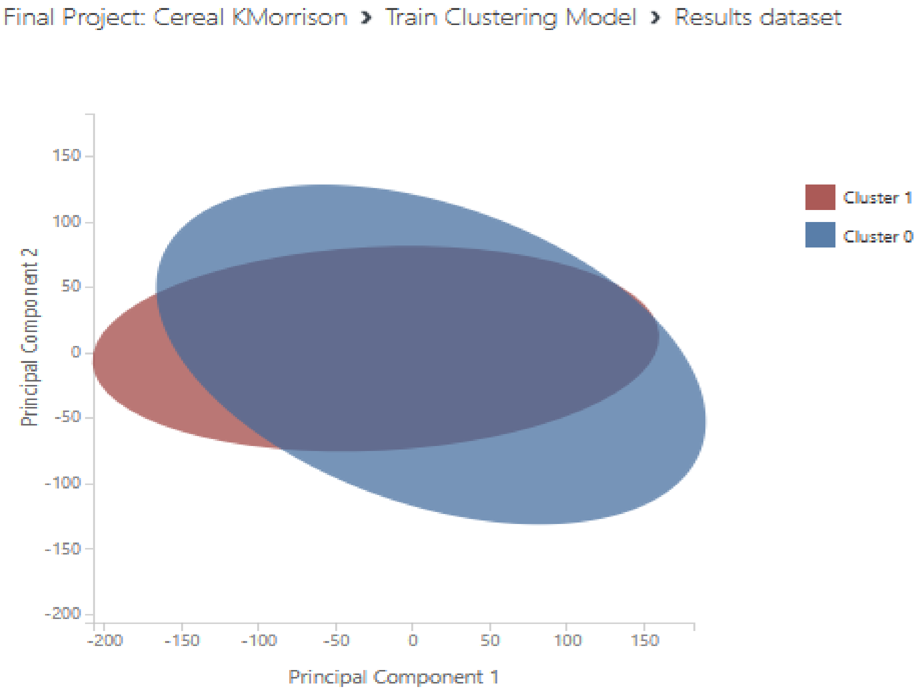
SELECTED COLUMNS

All Types | search columns

mfr
calories
sugars
cups
rating

5 columns selected

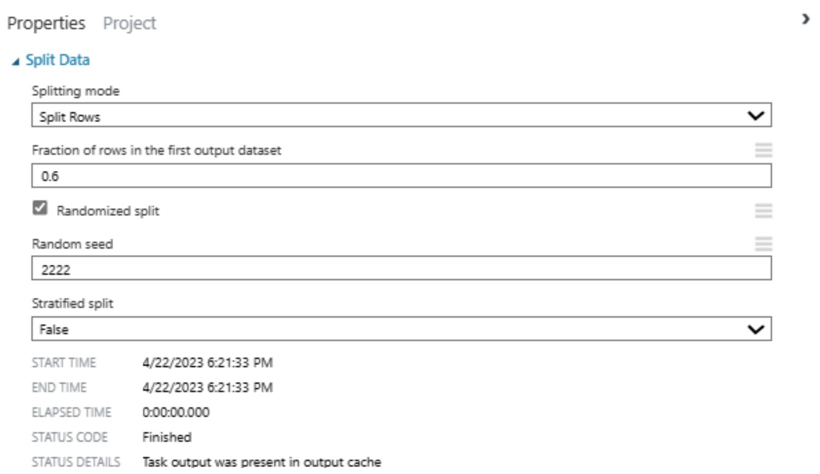
-



Final Project: Cereal KMorrison › Train Clustering Model › Results dataset

- Error in Python script execution when I attempted to review the scatter plots after creating k cluster model

Splitting data



◢ Split Data

Splitting mode

Split Rows

Fraction of rows in the first output dataset

0.6

☑ Randomized split

Random seed

2222

Stratified split

False

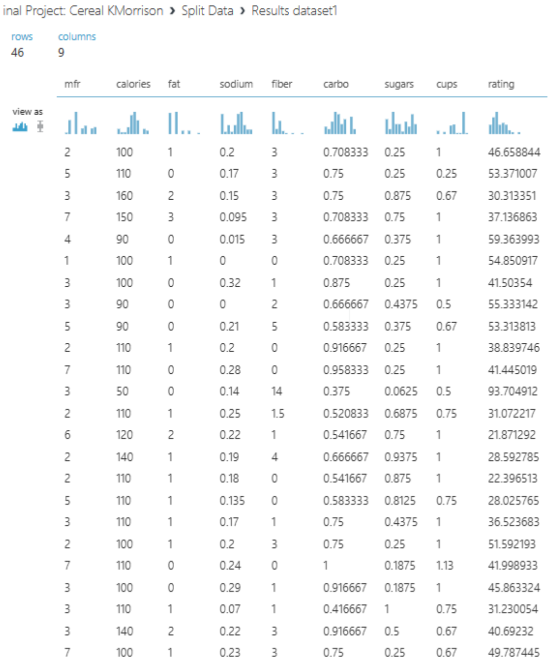
START TIME 4/22/2023 6:21:33 PM

END TIME 4/22/2023 6:21:33 PM

ELAPSED TIME 0:00:00.000

STATUS CODE Finished

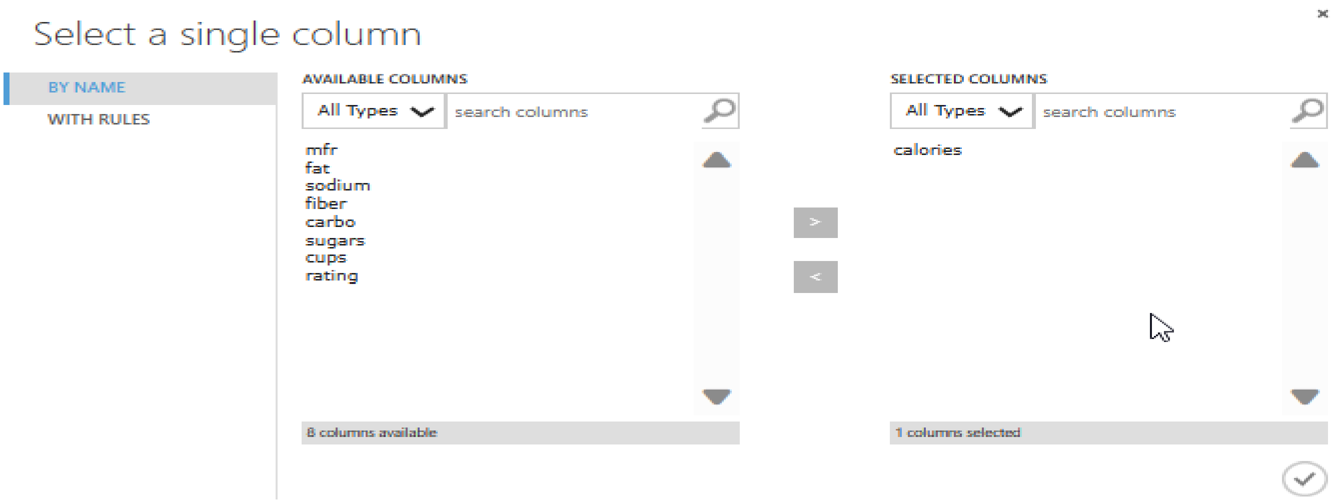
STATUS DETAILS Task output was present in output cache


inal Project: Cereal KMorrison › Split Data › Results dataset1

rows	columns
46	9

mfr	calories	fat	sodium	fiber	carbo	sugars	cups	rating
2	100	1	0.2	3	0.708333	0.25	1	46.658844
5	110	0	0.17	3	0.75	0.25	0.25	53.371007
3	160	2	0.15	3	0.75	0.875	0.67	30.313351
7	150	3	0.095	3	0.708333	0.75	1	37.136863
4	90	0	0.015	3	0.666667	0.375	1	59.363993
1	100	1	0	0	0.708333	0.25	1	54.850917
3	100	0	0.32	1	0.875	0.25	1	41.50354
3	90	0	0	2	0.666667	0.4375	0.5	55.333142
5	90	0	0.21	5	0.583333	0.375	0.67	53.313813
2	110	1	0.2	0	0.916667	0.25	1	38.839746
7	110	0	0.28	0	0.958333	0.25	1	41.445019
3	50	0	0.14	14	0.375	0.0625	0.5	93.704912
2	110	1	0.25	1.5	0.520833	0.6875	0.75	31.072217
6	120	2	0.22	1	0.541667	0.75	1	21.871292
2	140	1	0.19	4	0.666667	0.9375	1	28.592785
2	110	1	0.18	0	0.541667	0.875	1	22.396513
5	110	1	0.135	0	0.583333	0.8125	0.75	28.025765
3	110	1	0.17	1	0.75	0.4375	1	36.523683
2	100	1	0.2	3	0.75	0.25	1	51.592193
7	110	0	0.24	0	1	0.1875	1.13	41.998933
3	100	0	0.29	1	0.916667	0.1875	1	45.863324
3	110	1	0.07	1	0.416667	1	0.75	31.230054
3	140	2	0.22	3	0.916667	0.5	0.67	40.69232
7	100	1	0.23	3	0.75	0.25	0.67	49.787445

Training the model



Select a single column

BY NAME
WITH RULES

AVAILABLE COLUMNS

All Types | search columns

mfr
fat
sodium
fiber
carbo
sugars
cups
rating

8 columns available

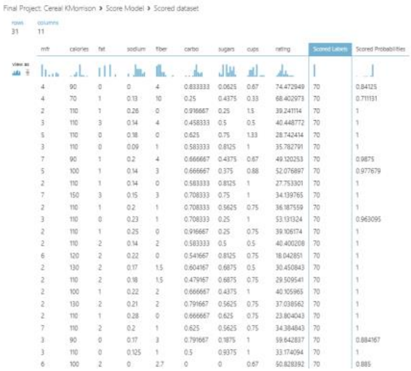
SELECTED COLUMNS

All Types | search columns

calories

1 columns selected

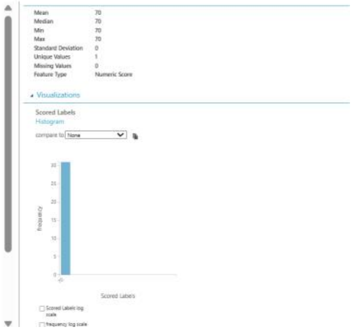
Scoring the model



Final Project: Cereal KMorrison › Score Model › Scored dataset

rows 31 columns 11

mfr	calories	fat	sodium	fiber	carbo	sugars	cups	rating	Scored Labels	Scored Probabilities
4	90	0	0	4	0.833333	0.0625	0.67	74.472949	70	0.84125
4	70	1	0.13	10	0.25	0.4375	0.33	68.402973	70	0.711131
2	110	1	0.26	0	0.916667	0.25	1.5	39.241114	70	1
3	110	3	0.14	4	0.458333	0.5	0.5	40.448772	70	1
5	110	0	0.18	0	0.625	0.75	1.33	28.742414	70	1
3	110	0	0.09	1	0.583333	0.8125	1	35.782791	70	1
7	90	1	0.2	4	0.666667	0.4375	0.67	49.120253	70	0.9875
5	100	1	0.14	3	0.666667	0.375	0.88	52.076897	70	0.977679
2	110	1	0.14	0	0.583333	0.8125	1	27.753301	70	1
7	150	3	0.15	3	0.708333	0.75	1	34.139765	70	1
2	110	1	0.2	1	0.708333	0.5625	0.75	36.187559	70	1
3	110	0	0.23	1	0.708333	0.25	1	53.131324	70	0.963095
2	110	1	0.25	0	0.916667	0.25	0.75	39.106174	70	1
2	110	2	0.14	2	0.583333	0.5	0.5	40.400208	70	1
6	120	2	0.22	0	0.541667	0.8125	0.75	18.042851	70	1
2	130	2	0.17	1.5	0.604167	0.6875	0.5	30.450843	70	1
2	110	2	0.18	1.5	0.479167	0.6875	0.75	29.509541	70	1
2	100	1	0.22	2	0.666667	0.4375	1	40.105965	70	1
2	130	2	0.21	2	0.791667	0.5625	0.75	37.038562	70	1
2	110	1	0.28	0	0.666667	0.625	0.75	23.804043	70	1
7	110	2	0.2	1	0.625	0.5625	0.75	34.384843	70	1
3	90	0	0.17	3	0.791667	0.1875	1	59.642837	70	0.884167
3	110	0	0.125	1	0.5	0.9375	1	33.174094	70	1
6	100	2	0	2.7	0	0	0.67	50.828392	70	0.885



Mean	70
Median	70
Min	70
Max	70
Standard Deviation	0
Unique Values	1
Missing Values	0
Feature Type	Numeric Score

Visualizations

Scored Labels
Histogram

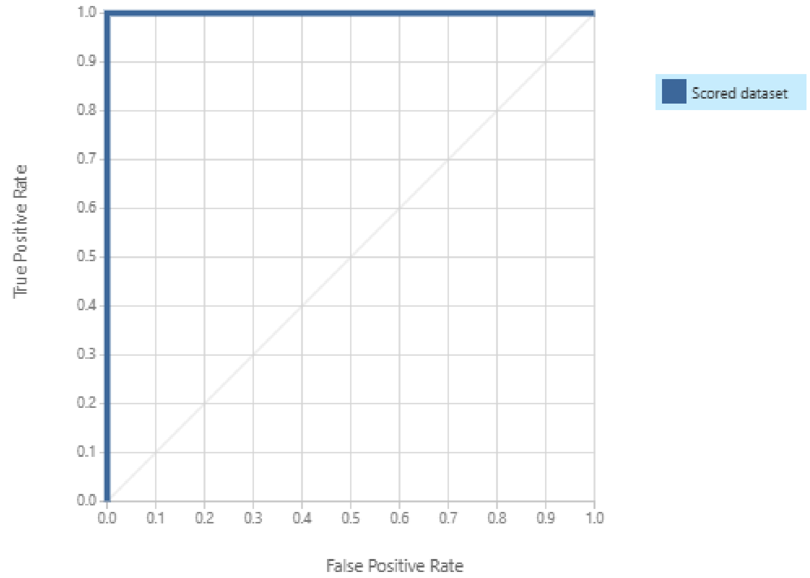
compare to None

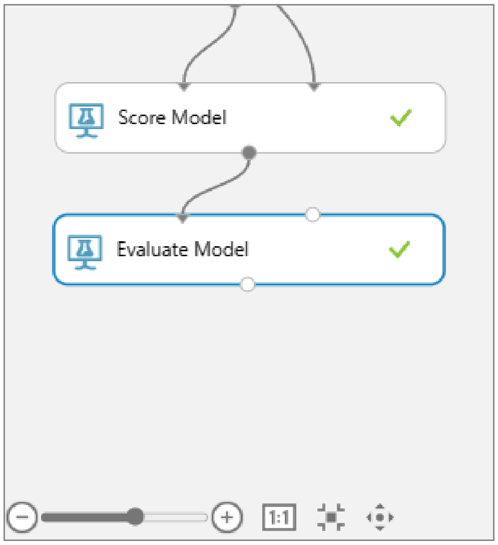
☐ Scored Labels log scale
☐ frequency log scale

Evaluating the model

Final Project: Cereal KMorrison › Evaluate Model › Evaluation results

ROC PRECISION/RECALL LIFT

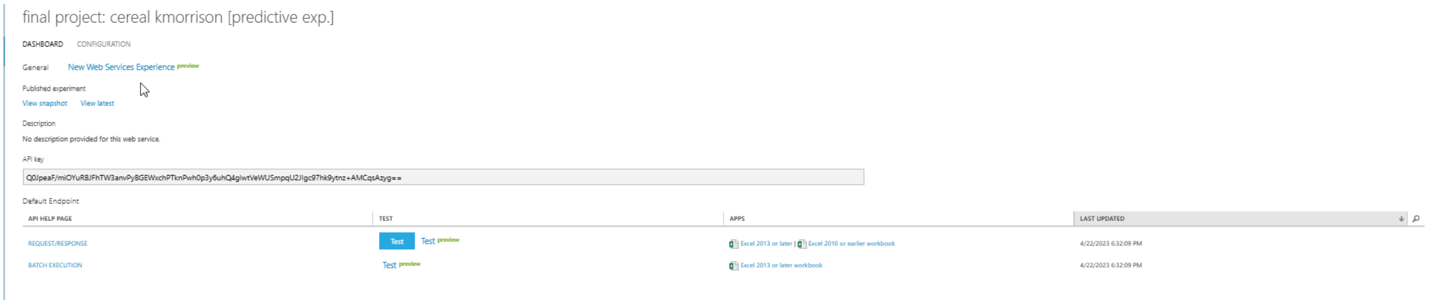




True Positive	False Negative	Accuracy	Precision	Threshold	AUC
30	0	0.968	0.968	0.5	1.000
False Positive	**True Negative**	**Recall**	**F1 Score**		
1	0	1.000	0.984		
Positive Label	**Negative Label**				
90	70				

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	27	0	0.871	0.903	0.947	1.000	0.900	0.250	1.000	0.000
(0.800,0.900]	3	0	0.968	1.000	1.000	1.000	1.000	1.000	1.000	0.000
(0.700,0.800]	0	1	1.000	0.968	0.984	0.968	1.000	1.000	0.000	1.000
(0.600,0.700]	0	0	1.000	0.968	0.984	0.968	1.000	1.000	0.000	1.000
(0.500,0.600]	0	0	1.000	0.968	0.984	0.968	1.000	1.000	0.000	1.000
(0.400,0.500]	0	0	1.000	0.968	0.984	0.968	1.000	1.000	0.000	1.000
(0.300,0.400]	0	0	1.000	0.968	0.984	0.968	1.000	1.000	0.000	1.000
(0.200,0.300]	0	0	1.000	0.968	0.984	0.968	1.000	1.000	0.000	1.000
(0.100,0.200]	0	0	1.000	0.968	0.984	0.968	1.000	1.000	0.000	1.000
(0.000,0.100]	0	0	1.000	0.968	0.984	0.968	1.000	1.000	0.000	1.000

Web Service



final project: cereal kmorrison [predictive exp.]

DASHBOARD CONFIGURATION

General New Web Services Experience preview

Published experiment

View snapshot View latest

Description

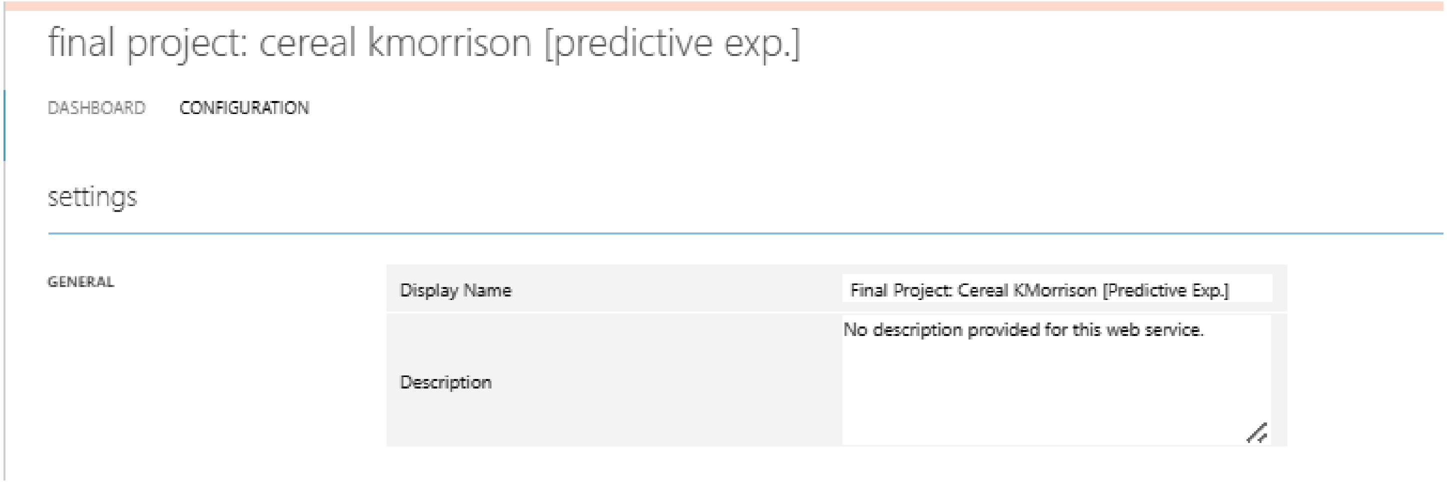
No description provided for this web service.

API key

Q0JpeaF/miOYuRBJFhTW3anvPyBGEWxchPTknPwh0p3y6uhQ4gIwtVeWUSmpqU2JIgc97hk9ytnz+AMCqsAzyg==

Default Endpoint

API HELP PAGE	TEST	APPS	LAST UPDATED
REQUEST/RESPONSE	Test Test preview	Excel 2013 or later | Excel 2010 or earlier workbook	4/22/2023 6:32:09 PM
BATCH EXECUTION	Test preview	Excel 2013 or later workbook	4/22/2023 6:32:09 PM



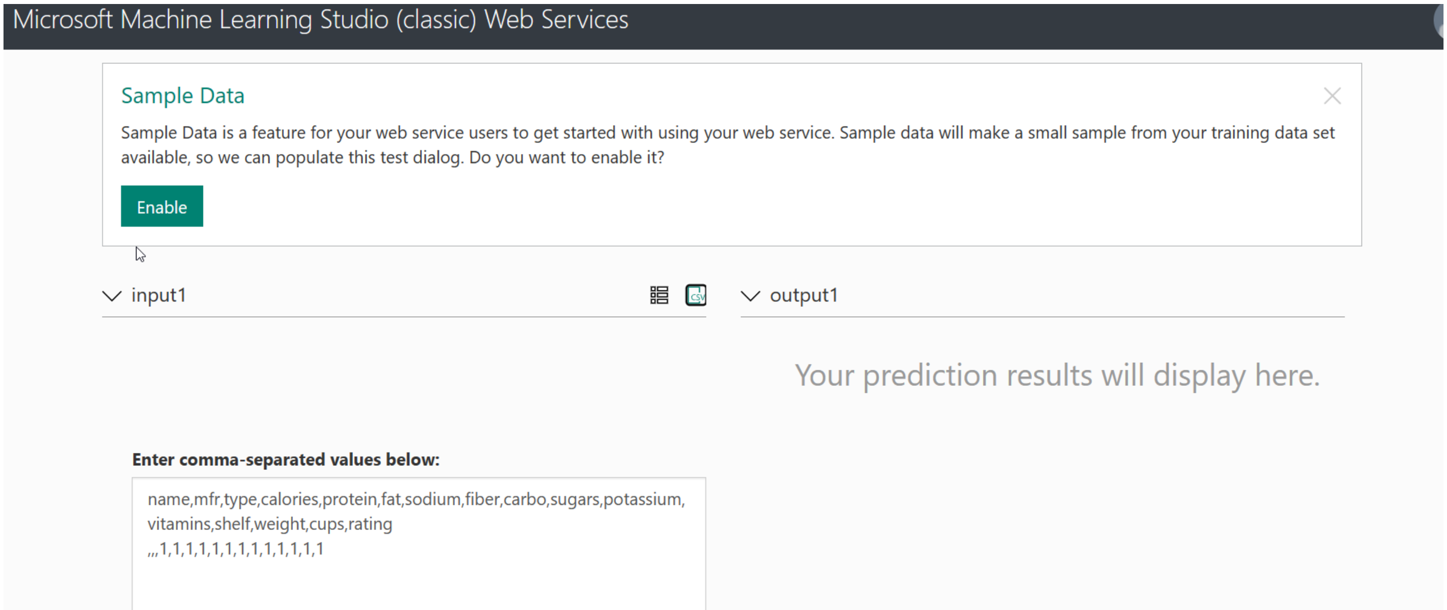
final project: cereal kmorrison [predictive exp.]

DASHBOARD CONFIGURATION

settings

GENERAL

Display Name	Final Project: Cereal KMorrison [Predictive Exp.]
Description	No description provided for this web service.



Microsoft Machine Learning Studio (classic) Web Services

Sample Data

Sample Data is a feature for your web service users to get started with using your web service. Sample data will make a small sample from your training data set available, so we can populate this test dialog. Do you want to enable it?

Your prediction results will display here.

Enter comma-separated values below:

```
name,mfr,type,calories,protein,fat,sodium,fiber,carbo,sugars,potassium,
vitamins,shelf,weight,cups,rating
,,,1,1,1,1,1,1,1,1,1,1,1,1,1
```

URL:
https://ussouthcentral.services.azureml.net/workspaces/439b6c9d90104635abccfe1580df8a3b/services/964cfd4145b442c68b854b99244c90d0/execute?api-version=2.0&details=true

API key:
Q0JpeaF/miOYuRBJFhTW3anvPyBGEWxchPTknPwh0p3y6uhQ4gIwtVeWUSmpqU2JIgc97hk9ytnz+AMCqsAzyg==

Challenges

- Link to the area of issue: The first time I ran though to the training and scoring the model, I did not remove the name of the cereal. There was an error that said Error 1000 Internal library exception. I had to look up the error to determine how to resolve (ErrAzure)
    - Resolution: Remove the cereal name from the dataset with column selector
- Not knowing Python was a large hurdle, we had to use the scripts provided by our Professor. I think some of the visualizations would have been easier to troubleshoot and/or script had I been proficient with Python

Career skills

- Learning how to research errors when running into roadblocks.
- Understanding AI and ML learning and how it is benefitting the tech space for many companies
- Ability to articulate what AI and ML are in order to understand how it is being used in your business

Conclusion

This class should require more than eight weeks to be able to complete a model independently. I think that having experience with Python to be able to understand what it is that we are executing.

References

(n.d.). Retrieved from https://learn.microsoft.com/en-us/previous-versions/azure/machine-learning/studio-module-reference/errors/machine-learning-module-error-codes?redirectedfrom=MSDN

(n.d.). Retrieved from https://bernardmarr.com/what-is-an-artificial-neural-networks/