A Project Report on

Cereal Analysis Based on Ratings by using machine learning techniques

Submitted to Smart Bridge

By:

Anumukonda Naga Pavan Chandu	18481A0204
Choragudi Sandeep	17481A0213
Balijepalli kaushik venkat sai	18481A0207
Gangu Mahesh	18481A0227
Gummadi Chandu	18481A0230

Under the guidance of

Mrs. Pradeepthi

Lead – Artificial Intelligence at SmartBridge Educational Services Pvt. Ltd

CONTENTS

INTRODUCTION

This data analysis report explores the various relationships and ideas that can be concluded from using statistical methods and packages from the R programming language, concerning a dataset of seventy-seven different types of cereals.

The data was found on the Mercer Blackhawk server, and contains twelve different data fields, as follows:

QUANTITATIVE DATA NOMINAL DATA

NAME	Calories
MANUFACTURER	Protein
TYPE	Fat
	Sodium
	Fiber
	Carbohydrates
	Sugars
	Potassium
	Vitamins
	Shelf
	Weight
	Cups
	Rating*

There are seventy-four cold cereals and three hot cereals.

According to a FoodDive report[1] in 2017, nine out of ten consumers eat cereal for breakfast.
Unhealthy dietary habits are some of the most discussed problems concerning teens and children. Therefore, it is important to know the relationships between things like calories, sugar, vitamins, and how these factors compare to the rating of a cereal. After reading this report, it will be clear where the ratings are coming from and how they relate to the nutritional value of the cereal.

[1] https://www.fooddive.com/news/9-out-of-10-consumers-eat-cereal-for-breakfast-but-just-under-half-like- it/507552/

Asking the right questions determines the effectiveness of the data analysis. Some of the questions (concerning cereal data) that will be answered in this report are:

- Is there a relation between sugars, calories, carbs, and fat?
- How are calories and potassium distributed?
- Which manufacturers produce cereal with highest calories?
- How does rating compare to calorie count?
- Which nutrients are essential for a good rating for a cereal?
- Is there a relation between manufacturer and rating?

- Is there a relation between shelf number and rating?
- Can we use machine learning models to predict the rating of a cereal based on its nutritional values?

Methods

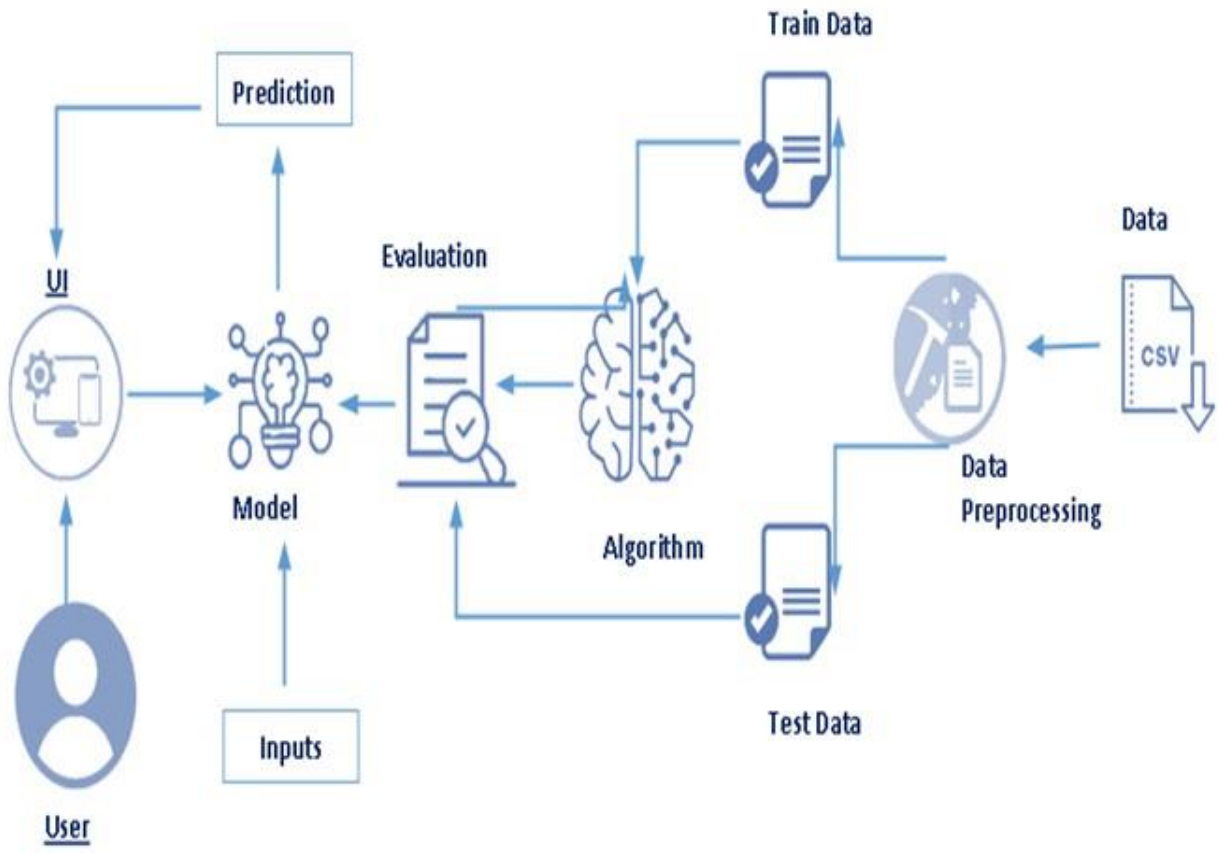
R is a robust platform for data manipulation that comes with many tools to visualize, summarize, and compute data.

The tools that were utilized in this analysis include:

Tool Purpose	*R function*	
Scatter-plot Matrix	Creates a correlation matrix plot between parameters in a dataframe.	pairs()
Pie Chart	Creates a pie chart where data is divided into “slices” to illustrate proportions.	
Histogram	Creates a histogram, which is a plot to show the frequency distribution of a variable in a dataframe.	hist()
Bar Chart	Creates a bar chart that represents categorical data with bars with heights	barplot()

Box-and-whisker Plot	proportional to the value they represent.	
Scatterplot	Creates a convenient illustration of the quartiles of a dataset, which is helpful for understanding the spread.	boxplot()
3D Scatterplot	Displays values for two variables of data on a Cartesian plane, which is helpful for understanding relationships between variables.	plot()
Linear Regression Model	A type of scatterplot that shows the relationship between three variables.	scatterplot3d()
	Fits a linear equation to the relationship between two variables. Very helpful in making predictions about future data.	lm()

ARCHITECTURE:



Software Installation:

Anaconda Navigator:

1. Anaconda Navigator is a free and open-source distribution of the Python and R programming languages for data science and machine learning related applications.
2. It can be installed on Windows, Linux, and Mac OS.
3. Conda is an open-source, cross-platform, package management system.

4. Anaconda comes with so very nice tools like JupyterLab, Jupyter Notebook,
5. QtConsole, Spyder, Glueviz, Orange, Rstudio, Visual Studio Code. For this project, we will be using Jupyter notebook and Spyder.

Data Set:

For this article we will be using the dataset provided by IBM which available at the UCI Machine Learning repository we using**00294** data set.

https://www.kaggle.com/crawford/80-cereals

Importing Libraries :

The first step is usually importing the libraries that will be needed in the program.

The required libraries to be imported to Python script are:

- Numpy: It is an open-source numerical Python library. It contains a multi-dimensional array and matrix data structures.
- It can be used to perform mathematical operations on arrays such as trigonometric, statistical, and algebraic routines.

- Pandas: It is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.

- Matplotlib: Visualisation with python. It is a comprehensive library for creating static, animated, and interactive visualizations in Python.

- Seaborn: Seaborn is a library for making statistical graphics in Python. ● Seaborn helps you explore and understand your data.
- Its plotting functions operate on data frames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plot.

- Pickle: The pickle module implements serialization protocol, which provides an ability to save and later load Python objects using special binary format.

Data Preprocessing

Data Pre-processing includes the following main tasks

- Import the Libraries.
- Importing the dataset.
- Checking for Null Values.
- Data Visualization.
- Label Encoding.
- OneHot Encoding.
- Splitting Data into Train and Test.

Importing The Dataset

- You might have your data in .csv files, .excel files
- Let's load a .csv data file into pandas using read_csv() function.We will need to locate the directory of the CSV file at first (it's more efficient to keep the dataset in the same directory as your program).
- If your dataset is in some other location ,Then

  Data=pd.read_csv(r"File_location")

  Note:r stands for "raw" and will cause backslashes in the string to be interpreted as actual backslashes rather than special characters.

- If the dataset is in the same directory of your program, you can directly read it, without giving raw as r.
- Our Dataset cereal contains following Columns

    1. name
    2. mfr
    3. type
    4. calories
    5. protein
    6. fat
    7. sodium
    8. fiber
    9. carbo
    10. sugars
    11. potass
    12. vitamins
    13. shelf
    14. weight
    15. cups
    16. rating

The output column to be predicted is rating .Based on the input variables we predict the food with a high beneficiary diet.

1. After loading it is important to check the complete information of data as it can indication many of the hidden information such as null values in a column or a row

2.Check whether any null values are there or not. if it is present then following can be done,

a. Imputing data using Imputation method in sklearn

b. Filling NaN values with mean, median and mode using fillna() method.

Handling Missing Values

1.After loading it is important to check the complete information of data as it can indication many of the hidden information such as null values in a column or a row

2.Check whether any null values are there or not. if it is present then following can be done,

a.Imputing data using Imputation method in sklearn

b.Filling NaN values with mean, median and mode using fillna() method.

```
data.isnull().any()
```

```
name        False
mfr         False
type        False
calories    False
protein     False
fat         False
sodium      False
fiber       False
carbo       False
sugars      False
potass      False
vitamins    False
shelf       False
weight      False
cups        False
rating      False
dtype: bool
```

3. Heatmap:It is way of representing the data in 2-D form.It gives coloured visual summary of the data.

```
plt.figure(figsize = (14, 8))
sns.heatmap(data.corr(), annot=True)
```

```
<AxesSubplot:>
```

From the heatmap, we see that there are no missing values in the dataset

Data Visualization:

- Data visualization is where a given data set is presented in a graphical format. It helps the detection of patterns, trends and correlations that might go undetected in text-based data.
- Understanding your data and the relationship present within it is just as important as any algorithm used to train your machine learning model. In fact, even the most sophisticated machine learning models will perform poorly on data that wasn't visualized and understood properly.
- To visualize the dataset we need libraries called Matplotlib and Seaborn.
- The Matplotlib library is a Python 2D plotting library which allows you to generate plots, scatter plots, histograms, bar charts etc.

Let's visualize our data using Matplotlib and seaborn library.

Before diving into the code, let's look at some of the basic properties we will be using when plotting.

xlabel: Set the label for the x-axis.

ylabel: Set the label for the y-axis.

title: Set a title for the axes.

Legend: Place a legend on the axes.

1. data.corr() gives the correlation between the columns

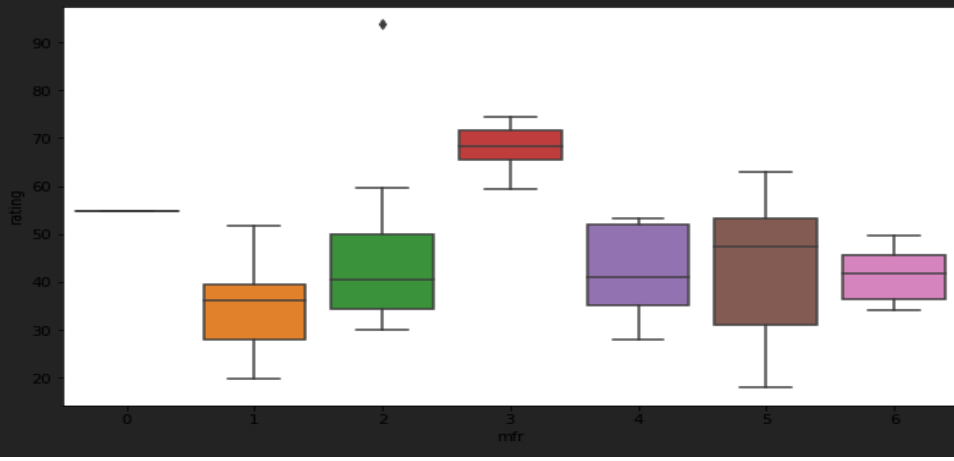
```
data.corr()
```

	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf
mfr	1.000000	-0.007103	-0.076328	0.017059	0.077661	-0.175791	0.056159	-0.063045	-0.132900	-0.003241	-0.274766	0.003323
type	-0.007103	1.000000	-0.071596	0.269265	-0.002615	-0.321552	-0.078114	-0.123023	-0.285219	-0.079825	-0.180633	-0.131730
calories	-0.076328	-0.071596	1.000000	0.019066	0.498610	0.300649	-0.293413	0.250681	0.562340	-0.066609	0.265356	0.097234
protein	0.017059	0.269265	0.019066	1.000000	0.208431	-0.054674	0.500330	-0.130864	-0.329142	0.549407	0.007335	0.133865
fat	0.077661	-0.002615	0.498610	0.208431	1.000000	-0.005407	0.016719	-0.318043	0.270819	0.193279	-0.031156	0.263691
sodium	-0.175791	-0.321552	0.300649	-0.054674	-0.005407	1.000000	-0.070675	0.355983	0.101451	-0.032603	0.361477	-0.069719
fiber	0.056159	-0.078114	-0.293413	0.500330	0.016719	-0.070675	1.000000	-0.356083	-0.141205	0.903374	-0.032243	0.297539
carbo	-0.063045	-0.123023	0.250681	-0.130864	-0.318043	0.355983	-0.356083	1.000000	-0.331665	-0.349685	0.258148	-0.101790
sugars	-0.132900	-0.285219	0.562340	-0.329142	0.270819	0.101451	-0.141205	-0.331665	1.000000	0.021696	0.125137	0.100438
potass	-0.003241	-0.079825	-0.066609	0.549407	0.193279	-0.032603	0.903374	-0.349685	0.021696	1.000000	0.020699	0.360663
vitamins	-0.274766	-0.180633	0.265356	0.007335	-0.031156	0.361477	-0.032243	0.258148	0.125137	0.020699	1.000000	0.299262
shelf	0.003323	-0.131730	0.097234	0.133865	0.263691	-0.069719	0.297539	-0.101790	0.100438	0.360663	0.299262	1.000000
weight	-0.240092	-0.039880	0.696091	0.216158	0.214625	0.308576	0.247226	0.135136	0.450648	0.416303	0.320324	0.190762
cups	-0.066967	0.060057	0.087200	-0.244469	-0.175892	0.119665	-0.513061	0.363932	-0.032358	-0.495195	0.128405	-0.335269
rating	0.140942	0.203024	-0.689376	0.470618	-0.409284	-0.401295	0.584160	0.052055	-0.759675	0.380165	-0.240544	0.025159

2. Box plot

```
plt.figure(figsize = (10, 6))
sns.boxplot(data = data, x = "mfr", y = "rating")
```

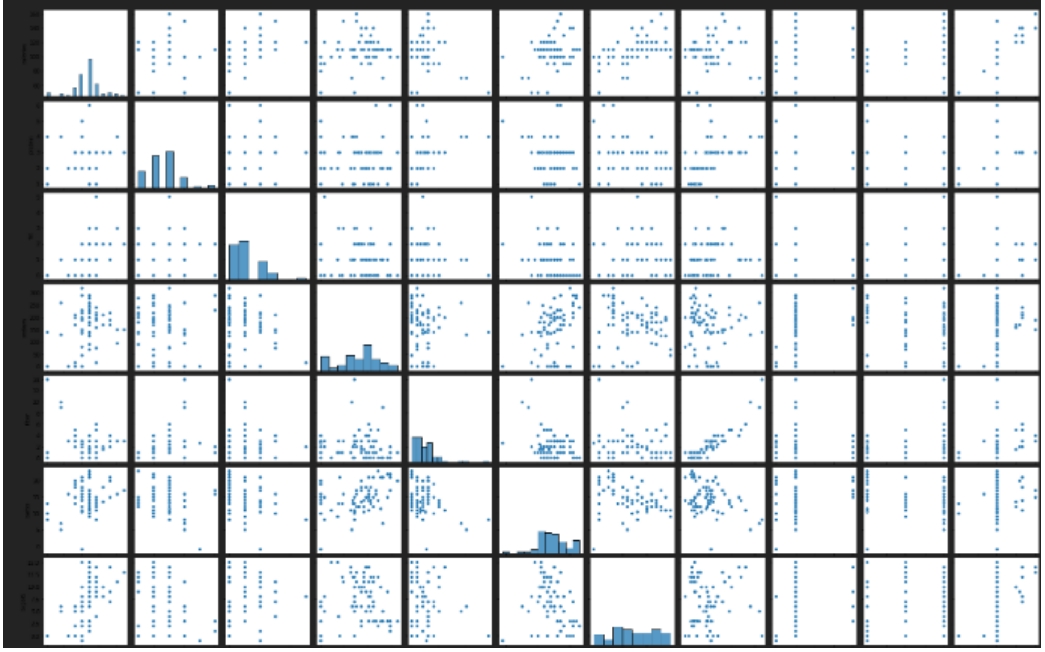
```
<AxesSubplot:xlabel='mfr', ylabel='rating'>
```



3. Pairplot

```
sns.pairplot(data=data, markers=["^","v"], palette="inferno")
```

```
<seaborn.axisgrid.PairGrid at 0xa2068c8>
```



Splitting The Dataset Into Dependent And Independent Variable

Split The Dependent And Independent Features Into Train Set And

- In machine learning, the concept of dependent variable (y) and independent variables(x) is important to understand. Here, Dependent variable is nothing but output in the dataset and the independent variable is all inputs in the dataset.
- With this in mind, we need to split our dataset into the matrix of independent variables and the vector or dependent variable. Mathematically, Vector is defined as a matrix that has just one column.

To read the columns, we will use **iloc** of pandas (used to fix the indexes for selection) which takes two parameters — [row selection, column selection].

Let's split our dataset into independent and dependent variables.

1. The independent variable in the dataset would be considered as 'x' and name, mfr, type, calories , protein , fat, sodium , fiber, carbo, sugars , potass , vitamins , shelf , weight , cups columns would be considered as independent variable.
2. The dependent variable in the dataset would be considered as 'y' and the ' rating ' column is considered as dependent variable.

Now we will split the data of independent variables,

```
x= data.iloc[:,0:14].values
y= data.iloc[:,14:15].values
```

From the above code ":" indicates that you are considering all the rows in the dataset and "**0:18**" indicates that you are considering columns 0 to 8 such as sex, job and purpose as input values and assigning them to variable x. In the same way in second line ":" indicates you are considering all the rows and "**18:19**" indicates that you are considering only last column as output value and assigning them to variable y.

Application Building:

a. Create HTML file

b. Build Python Code

c. Run the app

Predicting the output using the model

Let us build an app.py flask file which is a web framework written in python for server-side scripting. Let's see the step by step procedure for building the backend application.

In order to develop web api with respect to our model, we basically use the Flask framework which is written in python

RESULT:

Cereal AnalysisBased on Ratings by using Machine Learning Techniques

A customer wants to buy some food items with high dietary benefits so that he wants to know which food item has high dietary benefits. It is so difficult to choose an item .Usually a customer expects to consume dietary cereals with high proteins, fiber and low sugars, fats. Predicting a brand with high dietary cereals became a big issue.

We use machine learning algorithms to predict the food with high beneficiary diet. The model can predict the rating of the food more accurate by giving the inputs which are the cereals and ingredients present in the food. Thus a customer can get high dietary food by the rating of the food given to it from the cereals and ingredients present. The rating is predicted using the neural networks model.

Click me to continue with prediction

Cereal Analysis Prediction

Calories

Protien

Fat

Sodium

Fiber

Carbo

Sugars

Potass

Vitamins

Shelf

Weight

Cups

Predict

Conclusions

• There is a positive correlation between calories and sugars in cereal.
• Most cereals do not have relatively high potassium values.
• Kellogg's offers the most cereals out of any manufacturer that are above the median calorie count (110).
• The more calories that a cereal has, the less likely it is to receive a high rating.
• Manufacturers that want to bring in high ratings should create cereals that are high in fiber, protein, and potassium and avoid creating cereals with high calorie counts or lots of sugar or fat.
• Cereals with high ratings are more likely to be placed on the first or third shelf, because that is generally where the consumers' eyes gravitate.
• Using a linear regression model can allow for accurate predictions of future cereal with less than ten percent error on average.
o For instance, a cereal that has thirteen grams of sugars, one-hundred and ten calories, and two grams of protein is projected to receive a rating of 32.03