Food Recommendation System

Abstract

To determine how nutritious the food is based on its nutritional value and the important chemicals required by humans.

Introduction

Nutrients are substances found in food that carry out biological activity and are important for the human body. They are divided into proteins, fats, carbohydrates, vitamins, and minerals, and perform the following essential functions:

1. Build all the body parts like muscles, bone, teeth, and blood
2. Energy generation (energy and heat)
3. Keeping the body working properly.

Nutrition is becoming a global issue as people today are more aware of the problems caused by it's deficiency. Nutrition is a wall against many diseases and ensures the best possible human health. However, each country has a problem with malnutrition, which is explained by dietary choices and lifestyle. An unhealthy diet can lead to stress, fatigue, and poor performance, as well as to the risk of certain illnesses and other health problems such as: obesity or overweight, tooth decay and high blood pressure.

Hence there was a need to develop a model which involves in determining a value based on the given attributes. This value is called rating and determines how nutritious and healthy the food is. The attributes are mainly nutrients, food chemicals, etc which helped in determining whether it was a nourishing meal or not.

With the usage of this model, nutritious and healthy food can be recommended, and nutrition deficiencies can be avoided in many countries. It also will spread awareness amongst people as to the fact that nutrition is important to have a nourishing life.

Process and Method

The primary objective of the model is to determine a rating, or score based on the given independent variables or attribute values. The other questions that problem will either solve or give more insight are:

1. How much nutrition is required for a particular food on an average for having a well-balanced nutritional and healthy value?
2. What proportion of nutrients is unhealthy i.e., what is the conjunction of attribute values that determines the score of the food which in-turn determines whether the food is healthy or not?
3. Should there be a balanced amount of biochemicals, substances, etc that the human body must intake for a balanced diet?

The dataset obtained has 20000+ records with a good number of attributes and is well spread with respect to nutrients values. However, the problems faced with the dataset are:

1. There are several outliers for some attributes which may seem

irrelevant but may have some value with respect to the model. This serves as a problem since outliers may be misleading.

2. There are a lot of unimportant attributes present in the dataset which provide little to no value with respect to the ratings or score given to the model.
3. The determination of which attributes are important is difficult due to the number of attributes present.

In Spite of all this, these problems can be solved with the help of the following methods:

1. PCA (Principal Component Analysis).
2. Normalization.
3. Scaling of values.
4. Removing outliers which provide little to no value to the model.
5. Removing attributes which provide little to no value to the model.

All these are implemented by the use of python libraries such as sklearn, sklearn.decomposition.PCA,pandas,numpy ,matplotlib(a visual representation of data which helps in determining outliers.)

The outliers were successfully removed and the dimensions were regulated and the necessary attributes are taken into consideration. The graphs helped obtain insights and correlations with all attributes and target value(the rating).

Proposed solution

Pre-processing.

1. Removing of Duplicate Data

We found there were 1801 records found in our dataset. So as a part of pre-processing we removed these duplicate data using drop_duplicates () method.

2. Dropping of Unwanted Columns like title

3. Creating two separate data one for categorical and one for continuous data

4. **Creating New Columns like rating labels to perform a classification algorithm in the later stage**.

As our predictor rating was a continuous variable, we introduced a new column called rating label which performed a range of values to a given label transformation (concept hierarchy).

So, our rating labels was done as follows:

- 0-1 -> Very Bad
- 1-2-> Bad
- 2-3-> Average
- 3-4-> Good
- 4-5 -> Excellent

This new column helped us in visualization of our data.

5. Checking Null values. We found null values in the continuous data hence replaced the null values with the mean.

6. Removal of outliers from the dataset. We used the metric Q1-1.5*IQR>value or value<=Q3+1.5*IQR to find outliers.

7. Finding the correlation between the continuous variables.

8. Standardizing the continuous variables for better model accuracy.

As the categorical variables were dichotomous i.e., values were either 0 or 1.

There wasn’t any requirement of standardization.

But for the continuous variable (sodium, fat, protein, calories) the range of variables had a lot of variation, so we performed a standardization of these variables. The method used for standardization was StandardScaler().

9. PCA analysis to check if dimensionality reduction is helpful for making the model better.

We used a PCA () instance from sklearn.decomposition to decompose into 3 components.These provided us with a low correlation between PCA components.

10. Insights based on visualisation of the dataset.

a. Histogram of rating labels

b. Correlation scatter plot between two continuous variables

c. Histogram on PCA components

d. Correlation scatter plot between the various PCA components

Building A Model

Based on pre-processing and visualization of the dataset we planned on making three different models those are Multiple linear regression, SVM, DecisionTreeRegressor.

1. Multiple Linear Regression (Regression Model)

a. Creating two variables determinant Variable (Independent Variable) and determined variable (Dependent Variable).

Here determinant Variables are Calories, Fat, Protein, Sodium, Ingredients, and the determined variable is Rating(continuous).

b. Dividing the dataset into a training set and test set with a split of 0.2 using train_test_split().

Training set->80%
Test set->20%

c. Creating an instance of LinearRegression() Class of sklearn.linear_model.

d. Fitting the training set on this model using fit() method.

e. Predict the determined variable based on the test dataset using predict() method.

f. Getting the predicted labels on the ratings predicted using the model.

2. SVM(Classification Model)

a. Creating two variables svm_determinant Variable (Independent Variable) and svm_determined variable (Dependent Variable).

Here svm_determinant Variable are Calories, Fat, Protein, Sodium, Ingredients, and the svm_determined variable is RatingLabel (discrete)

b. Using label Encoding in-order to convert svm_determined variable into discrete integer values(As labels are in strings).

c. Dividing the dataset into a training set and test set with a split of 0.2 using train_test_split ().

Training set->80%
Test set->20%

d. Creating an instance of SVC() Class of linear kernel of sklearn.

e. Fitting the training set on this model using fit() method.

f. Predict the determined variable based on the test dataset using predict() method.

3. DecisionTreeRegressor (Regression+Classification)

a. Creating two variables X (Independent Variable) and Y (Dependent Variable).

Here X are Calories, Fat, Protein, Sodium, Ingredients, and the Y is Rating (continuous)

b. Dividing the dataset into a training set and test set with a split of 0.24 using train_test_split ().

Training set->76%
Test set->24%

c. Creating an instance of DecisionTreeRegressor()

d. Fitting the training set on this model using fit() method.

e. Predict the determined variable based on the test dataset using predict() method.

f. Getting the predicted labels on the ratings predicted using the model.

Experimental Results and Insights:

The observations which were concluded after implantation of the project are:

1. Amongst Multiple Linear Regression, SVM and Decision Tree, it is found that SVM is the best model that solves our purpose.

2. The SVM model however at times fails to classify the data in the **average** category.

3. Upon combining all categorical variables, the dimensionality of the model is reduced, thus preventing overfitting.

4. PCA doesn’t benefit towards training the model.

5. Time efficiency of combining all categorical variables is very high which is not ideal.

Evaluation

As a part of evaluation of our model we used various metrics based on the model whether it is classification or regression.

1. For multiple linear regression models the metric used is, root mean squared error, mean squared error.

2. For SVM we used accuracy, precision, recall.

3. For DecisionTreeRegressor we used accuracy, mean absolute error, mean squared error, root mean squared error.

The results for the models:

1. Multiple Linear Regression ->

   RMSE 1.31

2. SVM Classification ->100% accuracy

3. DecisionTreeRegressor -> RMSE 1.72

Limitations

We could have used the concept of Natural Language Processing In order to give weights to each of categorical column instead of assigning the same weights to all columns. We could also try with some things like standardizing the continuous column.

Assumptions: -

Each categorical column in the csv file is given equal importance i.e., all have a same weight in predicting the final rating. This has been ensured by adding all the categorical columns values and obtaining the sum and remove all the 670+ categorical values present in the dataset.

The final dataset is now of size 6 columns which is further used for training.

Conclusion:

After analysing and applying the model on our data we observed that support vector machines or SVM in short was the best to predict the classes of rating. This is because SVM tries to maximize the distance between margins thereby creating a proper classification boundary for each class helping us predict it accurately. Finally, if a restaurant owner would like to know how his food would be liked by customers who in the current era concentrate more on macro and micronutrients could provide details of some of the nutrients and get a rating which can be used to further improvise the quality before launching into the market.

Sno	Name	SRN	Contributions	Time Spent (in Hrs)
1	Venkata Krishnarjun Vuppala	PES2UG19CS451	Implemented the Machine Learning Models	40 hrs
2	SVSC Santosh	PES2UG19CS346	Did Preprocessing and Data Visualization	40 hrs
3	S. Mahammad Aasheesh	PES2UG19CS342	Implemented the Machine Learning Models	40 hrs
4	Kuntal Gorai	PES2UG19CS198	Did Data Visualization and Data Visualization	40 hrs