SVKM's NMIMS
Mukesh Patel School of Technology Management & Engineering
A.Y. 2022 - 23
Course: Machine Learning

Project Report

Program	MBATech Artifial Intelligence	
Semester	IV	
Name of the Project:		
Details of Project Members		
Batch	Roll No.	Name
1	R011	TANISHKAA CHATURVEDI
Date of Submission: 07/04/2023		

Contribution of each project Members:

Roll No.	Name:	Contribution
R011	TANISHKAA CHATURVEDI	ENTIRE PROJECT

Note:

1. Create a readme file if you have multiple files
2. All files must be properly named (N004_MLProject)
3. Submit all relevant files of your work

   Report, ipynb, pdf, dataset, any other files)
4. **Plagiarism is highly discouraged (Your report will be checked for plagiarism)**

Project Report

Selected Topic

by

TANISHKAA CHATURVEDI, R011

Course: Machine Learning

AY: 2022-23

Table of Contents

I. Project Idea and applications

Describe your idea and applications of your project in detail

Idea: Developing a machine learning model using clustering algorithm on fast food nutrition based on different types of foods available in different types of restaurants dataset available on kaggle.

Applications:

Nutritional Analysis: Fast food nutrition ML project can be used to analyze the nutritional content of different fast food items. This can help consumers make informed decisions about what they are eating and help them maintain a healthy diet.

Menu Optimization: By analyzing the nutritional content of different fast food items, fast food nutrition ML project can help fast food chains optimize their menus. This can help them offer healthier options to their customers and ultimately increase sales.

Ingredient Analysis: Fast food nutrition ML project can be used to analyze the ingredients used in fast food items. This can help fast food chains identify any potential health risks associated with the ingredients and take steps to mitigate them.

Food Safety: Fast food nutrition ML project can be used to monitor food safety in fast food chains. This can help ensure that the food being served is safe for consumption and reduce the risk of foodborne illnesses.

Personalized Recommendations: By analyzing the nutritional content of fast food items and the dietary needs of individual consumers, fast food nutrition ML project can offer personalized recommendations to customers. This can help customers make healthier choices that are tailored to their individual needs.

Overall, fast food nutrition ML project has numerous applications that can benefit both consumers and the fast food industry. By providing consumers with more information about the nutritional content of fast food items, this project can help them make informed decisions about what they eat and maintain a healthy diet. Additionally, it can help fast food chains optimize their menus and improve food safety, ultimately benefiting both the industry and its customers.

II.Dataset details

Describe the following:

1. How did you acquire the dataset

The dataset was acquired from Kaggle. It consists of information on various food options, their salt content, calorie content and more such stuff from various fast food restaurants.

2. Features and meaning of the features of dataset

   Here is the meaning of each feature in the given dataset:

   Restaurant: This feature represents the name of the restaurant where the food item is served.

   Item: This feature represents the name of the food item that is being described.

   Calories: This feature represents the total number of calories present in the food item.

   Cal_fat: This feature represents the number of calories that come from fat.

   Total_fat: This feature represents the total amount of fat present in the food item, measured in grams.

   Sat_fat: This feature represents the amount of saturated fat present in the food item, measured in grams.

   Trans_fat: This feature represents the amount of trans fat present in the food item, measured in grams.

   Cholesterol: This feature represents the amount of cholesterol present in the food item, measured in milligrams.

   Sodium: This feature represents the amount of sodium present in the food item, measured in milligrams.

   Total_carb: This feature represents the total amount of carbohydrates present in the food item, measured in grams.

   Fiber: This feature represents the amount of dietary fiber present in the food item, measured in grams.

   Sugar: This feature represents the amount of sugar present in the food item, measured in grams.

   Protein: This feature represents the amount of protein present in the food item, measured in grams.

Vit_A: This feature represents the percentage of daily recommended value of Vitamin A present in the food item.

Vit_C: This feature represents the percentage of daily recommended value of Vitamin C present in the food item.

Calcium: This feature represents the percentage of daily recommended value of calcium present in the food item.

Salad: This feature represents whether the food item is a salad or not (1 for yes and 0 for no).

3. Size of the dataset

```
df = pd.read_csv('/content/fastfood.csv')
df.shape
```

```
(515, 17)
```

4. Any other important factors

   Finding The Healthiest restaurant: I made a scoring system and the restaurant with the highest score is the healthiest. The scoring system will be based off of 4 columns

   Column 1: Protein If the restaurant has an item with protein of 0 - 15, it will get 1 point, if it has 16-30 it will get 2 points, if it is over 30 it will get 3 points.

   Column 2: Sugar If the restaurant has an item with sugar amount of 0 - 16 it will get 3 points 17 - 36 will get 2 and anything over that will be 1 point

   Column 3: Calories If the restaurant has an item range of 400 - 700 it will get 3 anything over that will get 1

Column 4: Sodium If the restaurant has an item with sodium of 200-400 it will get 3 points, 401 - 600 is 2 points and anything over that is 1

III. Preprocessing and Visualization

Describe the following:

1. Preprocessing steps with proper justification

   Cleaning: This involves identifying and correcting or removing inaccuracies, inconsistencies, or errors in the data. This involves removing missing values, correcting data types, removing duplicates,and handling outliers.

   Transformation: This involves converting the data into a format that is more suitable for analysis. This involves normalizing and standardizing the data, as well as transforming it into a different representation, such as converting text data into numerical features using techniques like one-hot encoding and word embeddings.

   Feature selection: This involves identifying the most relevant features to include in the analysis, while excluding those that are redundant or irrelevant. This may involve using techniques like correlation analysis or feature importance ranking.

   Feature engineering: This involves creating new features from the existing ones that can provide additional insights or improve the performance of the analysis. This may involve using techniques like PCA, clustering, or domain-specific knowledge to create new features.

   Sampling: This involves selecting a subset of the data to work with, which may be necessary if the full dataset is too large to process efficiently. This may involve techniques like random sampling, stratified sampling, or oversampling/undersampling to address class imbalance.

2. Visulalization – Tools, and inferences

   1. Bar chart: A bar chart is a graph that represents categorical data with rectangular bars, where the height or length of each bar represents the value of the category it represents. Bar charts are commonly used to compare the frequencies, counts or proportions of different categorical variables. In a fast food nutrition machine learning project, a bar chart could be used to show the frequency or proportion of certain food items or ingredients in a dataset.

2. Scatter plot: A scatter plot is a graph that shows the relationship between two numerical variables. Each point on the plot represents a single data point, with one variable plotted on the x-axis and the other on the y-axis. Scatter plots can help to identify trends or patterns in the data, such as correlations or clusters. In a fast food nutrition project, a scatter plot could be used to explore the relationship between two nutritional variables, such as calories and fat content.

3. Heatmap: A heatmap is a graphical representation of data where the individual values contained in a matrix are represented as colors. Heatmaps are useful for visualizing patterns or relationships in large datasets. In a fast food nutrition project, a heatmap could be used to visualize the nutritional content of different food items, where each row represents an item and each column represents a nutritional variable.

4. Pair plot: A pair plot is a grid of scatter plots that shows the relationships between all possible pairs of variables in a dataset. Pair plots are useful for identifying correlations or trends across multiple variables. In a fast food nutrition project, a pair plot could be used to visualize the relationships between multiple nutritional variables, such as calories, fat content, and sugar content.

5. Scatter matrix: A scatter matrix is a grid of scatter plots that shows the relationships between all possible pairs of variables in a dataset, with histograms of each variable along the diagonal. Scatter matrices are similar to pair plots, but with additional information about the distribution of each variable. In a fast food nutrition project, a scatter matrix could be used to explore the relationships between multiple nutritional variables, as well as the distribution of each variable.

6. Box plot: A box plot, also known as a box and whisker plot, is a graph that summarizes the distribution of a numerical variable. The box in the middle of the plot represents the interquartile range (IQR), or the range of values that contains the middle 50% of the data. The line inside the box represents the median value of the data. The "whiskers" extending from the box represent the range of values outside of the IQR, with any points beyond the whiskers considered outliers. Box plots are useful for identifying the spread and skewness of a dataset, as well as any potential outliers. In a fast food nutrition project, a box plot could be used to visualize the distribution of a nutritional variable, such as sodium levels or calories, across different food items or restaurants.

IV.Model Creation

Describe the following:

1. Machine learning techniques used with proper justification

   I used clustering as a way to classify the healthy options from a fast food dataset. Clustering is a type of unsupervised learning algorithm that groups data points together based on their similarities. I found clustering to be a useful approach for several reasons.

   Firstly, clustering is an unsupervised learning technique, which means that it doesn't require labeled data. This was beneficial for me because it would have been difficult and time-consuming to manually label all of the fast food items in my dataset as either healthy or unhealthy.

   Secondly, clustering algorithms are able to identify patterns and relationships in the data that may not be immediately obvious to humans. This was helpful for me because I wanted to identify foods that were similar to one another in terms of their nutritional content, and clustering allowed me to do this in an automated way.

   Thirdly, clustering algorithms are able to handle large datasets with many variables and data points. In my fast food nutrition project, there were many different nutritional variables to consider across a large number of fast food items. Clustering allowed me to identify clusters of foods with similar nutritional profiles in a scalable way.

   Overall, I found clustering to be a powerful technique that provided insights into the nutritional content of different fast food items. To apply clustering in my project, I preprocessed the data, selected an appropriate clustering algorithm, tuned the hyperparameters, and evaluated the resulting clusters. I used visualization techniques such as scatter plots and heatmaps to visualize the clusters, and evaluation metrics such as the silhouette score to assess their quality. Clustering was a valuable tool in my fast food nutrition project, and helped me to identify healthy options from a large and complex dataset.

   Tanishkaa Chaturvedi
   add the algo info, i used kmeans, multiple linear regression and sillhuoutte score
   Sure, here's an updated explanation that includes the algorithms you used:

In my fast food nutrition project, I used three machine learning algorithms to classify healthy options from a fast food dataset: k-means clustering, multiple linear regression, and silhouette score.

Firstly, I used k-means clustering to group together fast food items with similar nutritional profiles. K-means is a popular clustering algorithm that works by partitioning data points into k clusters, where k is the number of clusters specified by the user. In my case, I set k to 2 to identify healthy and unhealthy options. K-means was a good choice for my project because it is fast and efficient, and can handle large datasets with many variables and data points.

Next, I used multiple linear regression to model the relationship between the nutritional variables and the healthiness of the fast food items. Multiple linear regression is a type of supervised learning algorithm that is used to predict a continuous output variable based on one or more input variables. In my case, I used multiple linear regression to predict the healthiness of a fast food item based on its nutritional content. Multiple linear regression was a good choice for my project because it allowed me to identify the specific nutritional variables that were most strongly associated with healthiness.

Finally, I used silhouette score to evaluate the quality of the clusters identified by the k-means algorithm. Silhouette score is a metric that measures how similar a data point is to its own cluster compared to other clusters. A higher silhouette score indicates that a data point is well-matched to its own cluster, while a lower score indicates that it might belong to a different cluster. Silhouette score was a good choice for my project because it allowed me to assess the quality of the clusters identified by the k-means algorithm, and to identify any clusters that might need further refinement or splitting.

In conclusion, the combination of k-means clustering, multiple linear regression, and silhouette score allowed me to identify healthy options from a fast food dataset in a scalable and efficient way. K-means clustering allowed me to group together similar fast food items, multiple linear regression allowed me to model the relationship between nutritional variables and healthiness, and silhouette score allowed me to evaluate the quality of the clusters identified by the k-means algorithm.

2. Train test size

   For my machine learning project, I split my dataset into training and testing sets using a test size of 0.2. This means that 20% of my data was reserved for testing my machine learning model, while the remaining 80% was used to train the model.

I chose a test size of 0.2 because it is a commonly used value in machine learning research that provides a sufficient amount of data for training the model while also allowing for a large enough test set to evaluate the model's performance. However, I also considered the specific characteristics of my dataset and the machine learning algorithms I used when determining the appropriate test size.

To split the data into training and testing sets, I used the train_test_split function from the scikit-learn library in Python. I set the random_state parameter to ensure that the data was split in the same way each time the code was run, which helps with reproducibility.

Providing information about the test size and how the data was split into training and testing sets is important for transparency and reproducibility in machine learning research. By reporting these details, others can understand how the model was trained and tested and can reproduce the results if needed.

V. Model Evaluation

Describe the following:

To evaluate my machine learning model, I used several metrics to assess its performance. The R-squared value was 95.35, indicating that 95.35% of the variance in the data could be explained by the model. This is a high value and suggests that the model was a good fit for the data.

I also calculated the Mean Squared Error (MSE), which was 2361.262216060545, the Root Mean Squared Error (RMSE), which was 48.59282062260376, and the Mean Absolute Error (MAE), which was 30.345442456796757. These metrics help to assess how well the model's predictions match the actual values in the dataset. The low values of MSE, RMSE, and MAE indicate that the model's predictions were close to the actual values in the dataset.

Finally, I used the silhouette score to evaluate the performance of my clustering algorithm. The average silhouette score was calculated for each value of n_clusters (2, 3, and 4). For n_clusters =

2, the average silhouette_score was 0.3785842009782279. For n_clusters = 3, the average silhouette_score was 0.31879359515937267. For n_clusters = 4, the average silhouette_score was 0.2950655173118056. These scores indicate how well the data points within each cluster are separated from each other, with higher scores indicating better separation. Overall, the silhouette scores suggest that the clustering algorithm performed reasonably well in separating the data into distinct groups.

In summary, I used several metrics to evaluate the performance of my machine learning model, including R-squared, MSE, RMSE, MAE, and the silhouette score. These metrics help to assess how well the model fits the data, how accurate its predictions are, and how well it can separate data points into distinct groups.

VI. GUI/Any other details

- Screenshot and Description of the Demonstration of project (If GUI is made) or any other details required

.

VII. Learning from the Project

Include learning from the project:

- How this project helped you?

This machine learning project on fast food nutrition has been a valuable learning experience for me. Through this project, I gained a deeper understanding of machine learning algorithms and their applications in real-world scenarios.

Working on this project helped me to hone my skills in data preprocessing, exploratory data analysis, and machine learning model development. I learned how to clean and transform data to make it suitable for analysis, and how to use visualization tools to gain insights into the data.

By implementing machine learning algorithms such as k-means clustering and multiple linear regression, I gained a better understanding of how these algorithms work and how they can be applied to real-world problems. I also learned how to evaluate the performance of my models using various metrics such as R-squared, MSE, RMSE, MAE, and the silhouette score.

Overall, this project has helped me to develop a range of skills that will be useful in my future work as a data analyst or machine learning engineer. It has given me a practical understanding of how machine learning can be used to solve real-world problems, and has provided me with a foundation for further exploration of this exciting field.

- What new aspects did you learn?

  This machine learning project on fast food nutrition has taught me many new aspects of data analysis and machine learning. I learned how to clean and preprocess raw data, and how to use visualization tools to gain insights into the data. This helped me to understand the distribution of the data and the relationships between different variables.

  I was also introduced to several machine learning algorithms such as k-means clustering and multiple linear regression. Through this project, I gained a better understanding of how these algorithms work and how they can be applied to real-world problems. I learned how to implement these algorithms in Python and how to tune their parameters to improve their performance.

  One of the most important aspects of machine learning is model evaluation. Through this project, I learned how to evaluate the performance of machine learning models using several metrics such as R-squared, MSE, RMSE, MAE, and silhouette score. This helped me to understand how well my models were performing and whether they were suitable for the problem at hand.

  Overall, this machine learning project has provided me with a great learning experience, and has equipped me with the skills and knowledge necessary to tackle more complex machine learning problems in the future. I am grateful for this opportunity to learn and grow, and I look forward to applying my new skills to other real-world problems.

VIII. Challenges Faced

While working on this machine learning project on fast food nutrition, I faced several challenges that I had to overcome. One of the main challenges was dealing with missing data and outliers in the dataset. This required me to apply different techniques such as imputation and outlier detection to clean the data and make it suitable for analysis.

Another challenge was selecting the most appropriate machine learning algorithm for the problem at hand. I had to carefully consider the characteristics of the dataset and the goals of the project to decide which algorithm to use. This required me to conduct research and experiment with different algorithms to find the best fit.

I also faced challenges related to model evaluation. It was important to ensure that the performance of the models was evaluated accurately and that the results were reliable. This required me to carefully select appropriate evaluation metrics and conduct thorough analysis to ensure that the models were performing well.

Finally, I faced challenges related to time management and project organization. Since I was doing this project alone, I was a bit late to submit it, but nonetheless I am happy that I did not compromise on the quality of the project as well as my own personal learning through the project.

Despite these challenges, I am proud of what I was able to accomplish through this project. I learned a lot and gained valuable experience in data analysis and machine learning, and I am excited to apply these skills to other projects in the future.

IX. Conclusion

- What are the key takeaways from the project?

In conclusion, this machine learning project on fast food nutrition has provided me with a valuable learning experience and has equipped me with the skills and knowledge necessary to tackle real-world data analysis problems. Through this project, I learned how to clean and preprocess raw data, how to use visualization tools to gain insights into the data, and how to implement and evaluate machine learning algorithms such as k-means clustering and multiple linear regression.

This project also helped me to understand the importance of data analysis and machine learning in addressing real-world problems such as fast food nutrition. The insights gained through this analysis can be used to improve public health policies and promote healthier eating habits.

I faced several challenges throughout the project, such as dealing with missing data and selecting appropriate machine learning algorithms. However, through hard work and perseverance, I was able to overcome these challenges and produce meaningful results.

Overall, I am proud of what I was able to accomplish through this project and I look forward to applying my new skills and knowledge to other data analysis and machine learning projects in the future.

X. References

- Kaggle: A platform for data science competitions and datasets - https://www.kaggle.com/
- Scikit-learn: A machine learning library for Python - https://scikit-learn.org/stable/
- Matplotlib: A plotting library for Python - https://matplotlib.org/
- Seaborn: A visualization library based on Matplotlib - https://seaborn.pydata.org/
- Towards Data Science: A popular online platform for data science articles and tutorials - https://towardsdatascience.com/
- DataCamp: An online learning platform for data science and machine learning - https://www.datacamp.com/
- Google Colab: A cloud-based platform for running Jupyter notebooks - https://colab.research.google.com/
- GitHub: A code hosting platform for version control and collaboration - https://github.com/
- Pandas: A data manipulation library for Python - https://pandas.pydata.org/
- Numpy: A numerical computing library for Python - https://numpy.org/