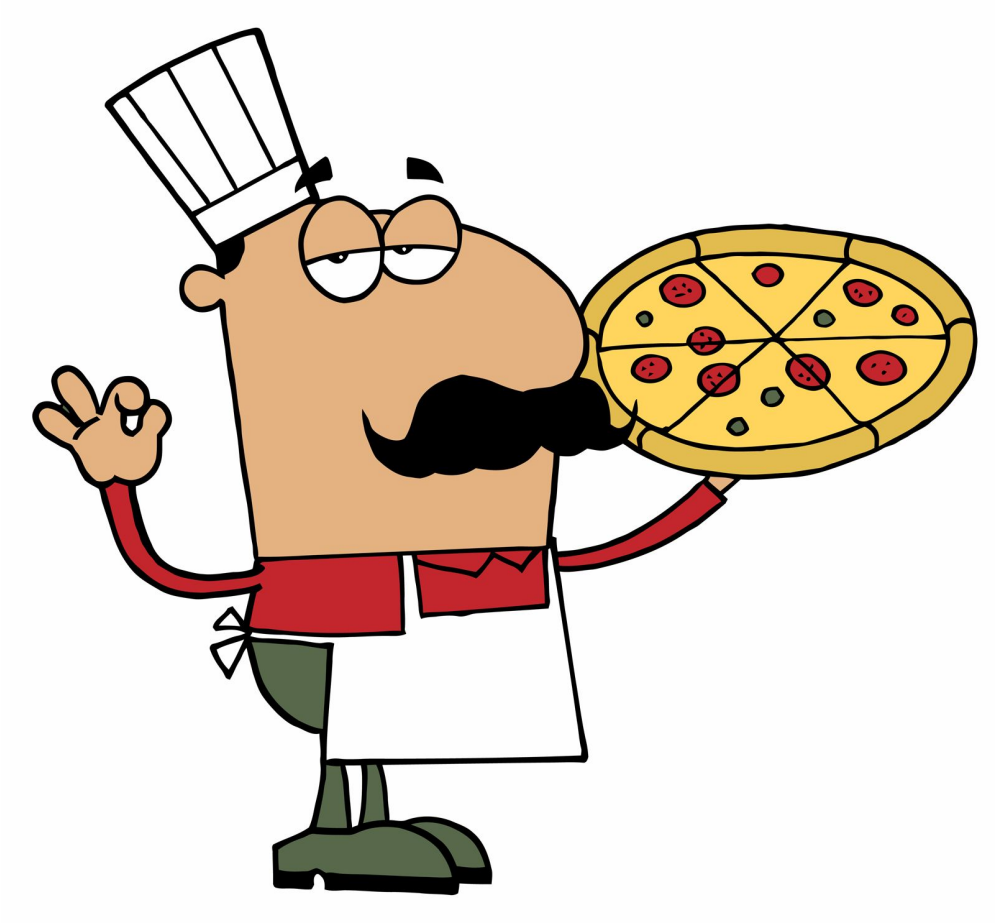
INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

Understanding the trends in Pizza restaurants and the pizza they sell



Submitted By:
Ashish Jain (MT18052)
Sarosh Hasan (MT18084)
Shubham Gupta (MT18055)

1. Articulation of the problem

Problem Statement

The problem is to identify the trends of various Pizza Restaurants and their products across geographies.
There are various kinds of inferences that we had drawn to explore our problem like how prices of a pizza vary across geographies, most famous Pizza in a different type of North America, etc.

Methods Used

- **Association Pattern Mining: -** Finding interesting association rules between different pizzas.
- **Classification:-** It is done between two classes i.e. Veg and Non-Veg based on the nutrients present in them.
- **Regression:-** Predicting calories by nutrients present in a particular pizza.

Dashboard

Interactive Map:- The dashboard provided help in analyzing the pizza distribution over North America through an interactive map.
HeatMap:- The heatmap helps in finding the regions or cities where the concentration of pizza outlets is more.

2. The complexity of data acquisition

The data is collected from “data.world”. The data consist of two dataset -

- Pizzas data from multiple restaurants in the US, provided by Datafiniti's Business Database.
- USDA National Nutrient DB.

The first dataset contains the different pizzas data and the second dataset contains the nutrient data of the pizzas and other products so we curated the nutrients data of pizzas from nutrients Db and the first dataset. The curated data help us in inferring different fact and predicting things like calories e.t.c.

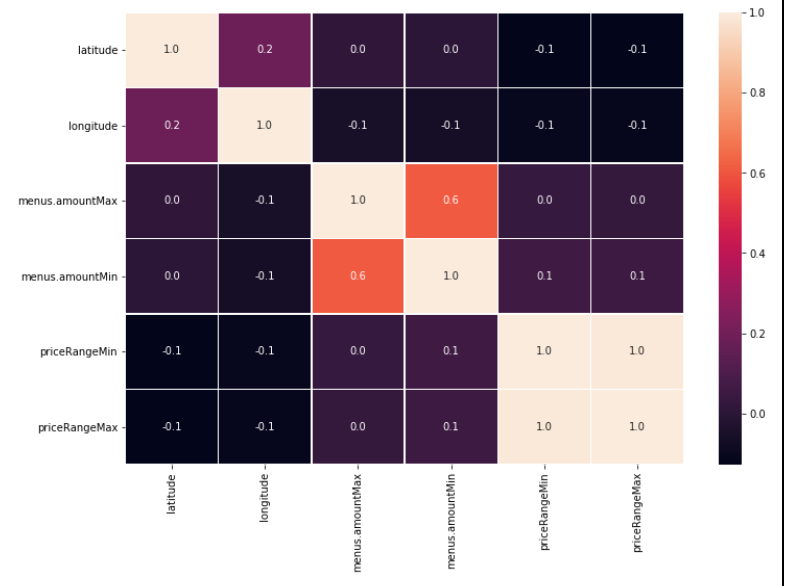
Dataset Descriptions -
This dataset “Datafiniti” is a list of 10,000 samples. The total no of features is 24.

keys	latitude	...	menus.currency	menus.dateSeen	menus.description	menus.name	name	postalCode	priceRangeCurrency	priceRangeMin	p
!91616	34.832300	...	USD	2018-05-01T04:25:37.197Z,2018-04-16T04:36:02.3...	NaN	Cheese Pizza	Shotgun Dans Pizza	72120	USD	0	
!22936	33.509266	...	USD	2018-03-03T02:38:06.381Z,2018-01-18T20:18:10.0...	NaN	Pizza Cookie	Sauce Pizza Wine	85012	USD	0	
'97122	39.144883	...	USD	2018-04-10T07:58:34.585Z,2018-04-21T05:43:21.4...	a saucelessampcomma double cheese pizza with a...	Pizza Blanca	Mios Pizzeria	45209	USD	0	
!63116	42.516669	...	USD	2016-10-20T21:50:02Z,2016-03-29T05:08:59Z	NaN	Small Pizza	Hungry Howies Pizza	48071	USD	25	
!65359	39.286630	...	USD	2016-03-31T02:34:04Z	NaN	Pizza Sub	Spartan Pizzeria	21224	USD	0	

This dataset “National Nutrient DB” is a list of 7229 samples with 42 features.

FoodGroup	ShortDescrip	Descrip	Energy_kcal	Protein_g	Fat_g	Carb_g	Sugar_g	Fiber_g	...	Folate_USRDA	Niacin_USRDA
Non-Veg	BUTTER,WITH SALT	Butter, salted	717.0	0.85	81.11	0.06	0.06	0.0	...	0.0075	0.002625
Non-Veg	BUTTER,WHIPPED,WITH SALT	Butter, whipped, with salt	717.0	0.85	81.11	0.06	0.06	0.0	...	0.0075	0.002625
Non-Veg	BUTTER OIL,ANHYDROUS	Butter oil, anhydrous	876.0	0.28	99.48	0.00	0.00	0.0	...	0.0000	0.000188
Non-Veg	CHEESE,BLUE	Cheese, blue	353.0	21.40	28.74	2.34	0.50	0.0	...	0.0900	0.063500
Non-Veg	CHEESE,BRICK	Cheese, brick	371.0	23.24	29.68	2.79	0.51	0.0	...	0.0500	0.007375

Correlation Heatmap and data feature description



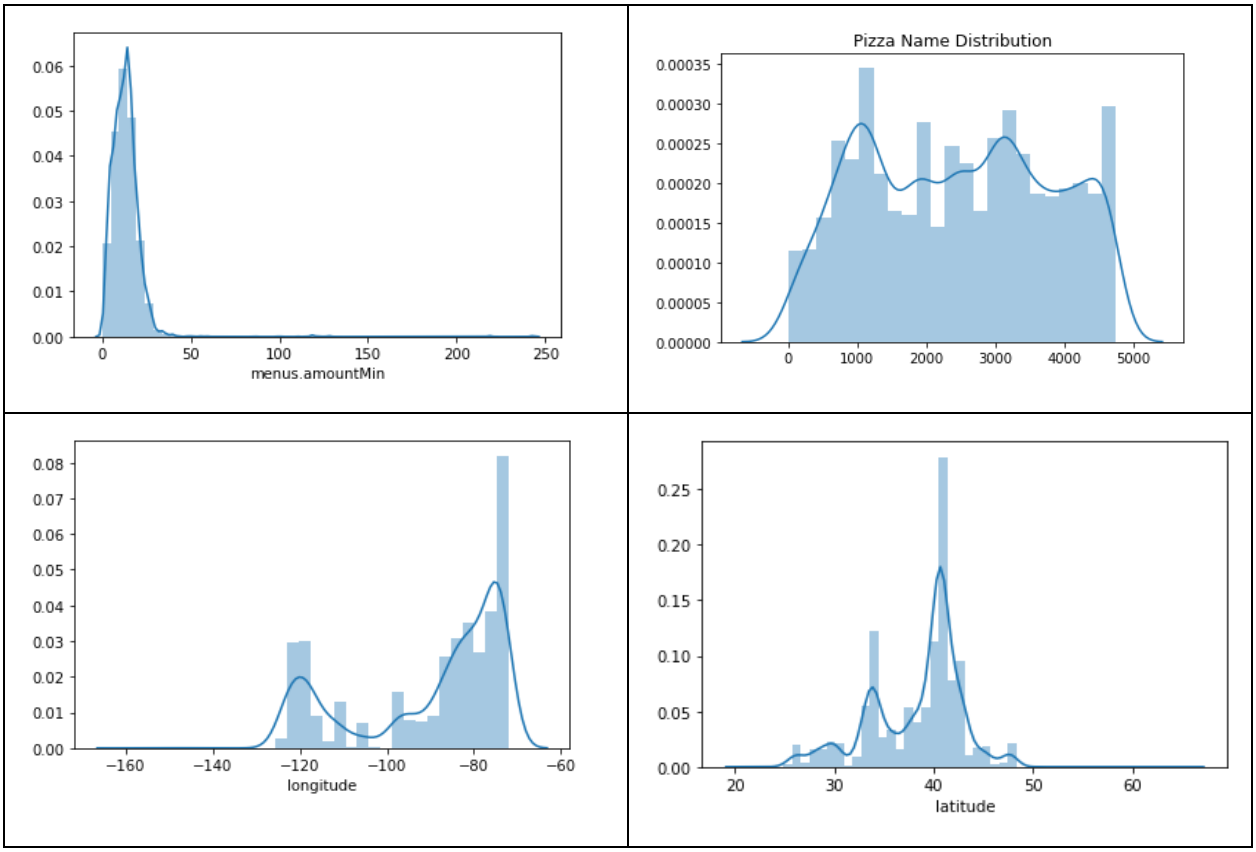

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 24 columns):
id                    10000 non-null object
dateAdded             10000 non-null object
dateUpdated           10000 non-null object
address               10000 non-null object
categories            10000 non-null object
primaryCategories     10000 non-null object
city                  10000 non-null object
country               10000 non-null object
keys                  10000 non-null object
latitude              10000 non-null float64
longitude             10000 non-null float64
menus.amountMax       10000 non-null float64
menus.amountMin       10000 non-null float64
menus.currency        10000 non-null object
menus.dateSeen        10000 non-null object
menus.description     3718 non-null object
menus.name            10000 non-null object
name                  10000 non-null object
postalCode            9996 non-null object
priceRangeCurrency    10000 non-null object
priceRangeMin         10000 non-null int64
priceRangeMax         10000 non-null int64
province              10000 non-null object
geometry              10000 non-null object
dtypes: float64(4), int64(2), object(18)
```

3. Hypotheses

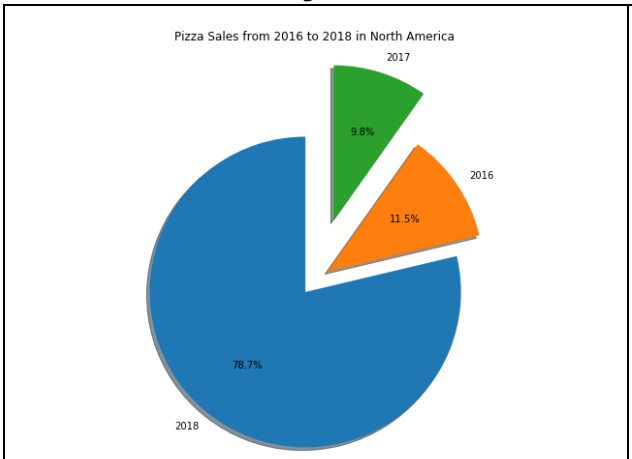
- The demand and rate of a product vary according to geography and type of Pizzas.
- The calories of a pizza depend upon its nutrients.
- Nutrients can be used for the classification of pizza as Veg and Non-Veg.

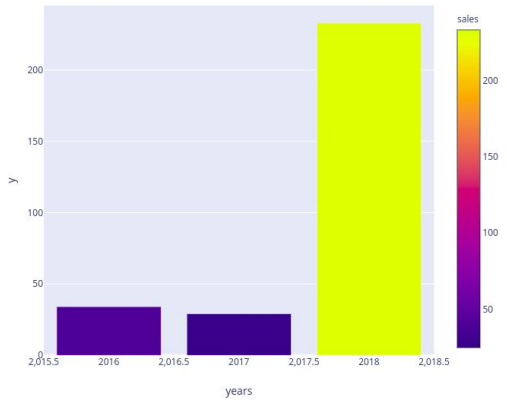
4. Visualization and statistics to support your hypothesis.

Data Distributions -



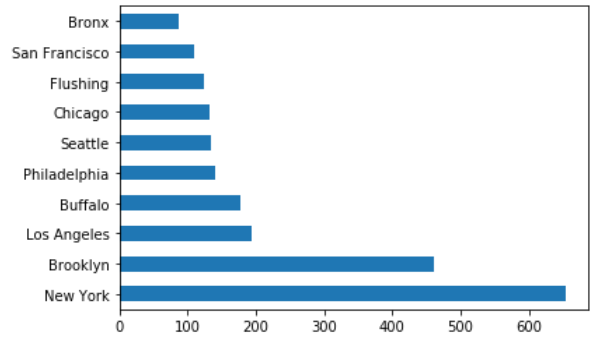
Pizza sale analysis from 2016 to 2018



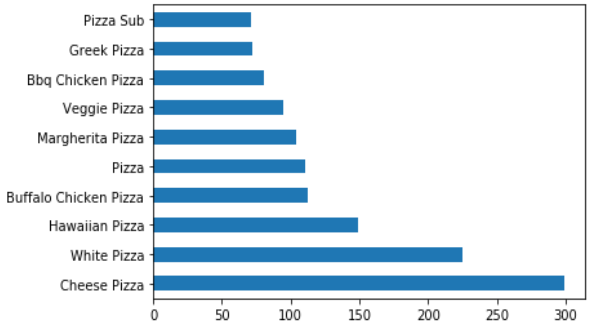


Inferences Drawn from Data -

Which are the top 10 cities with the most pizza restaurant?



Which are the top 10 pizzas in North America?



Which are the cheapest and the most expensive pizza and its pizza restaurant?

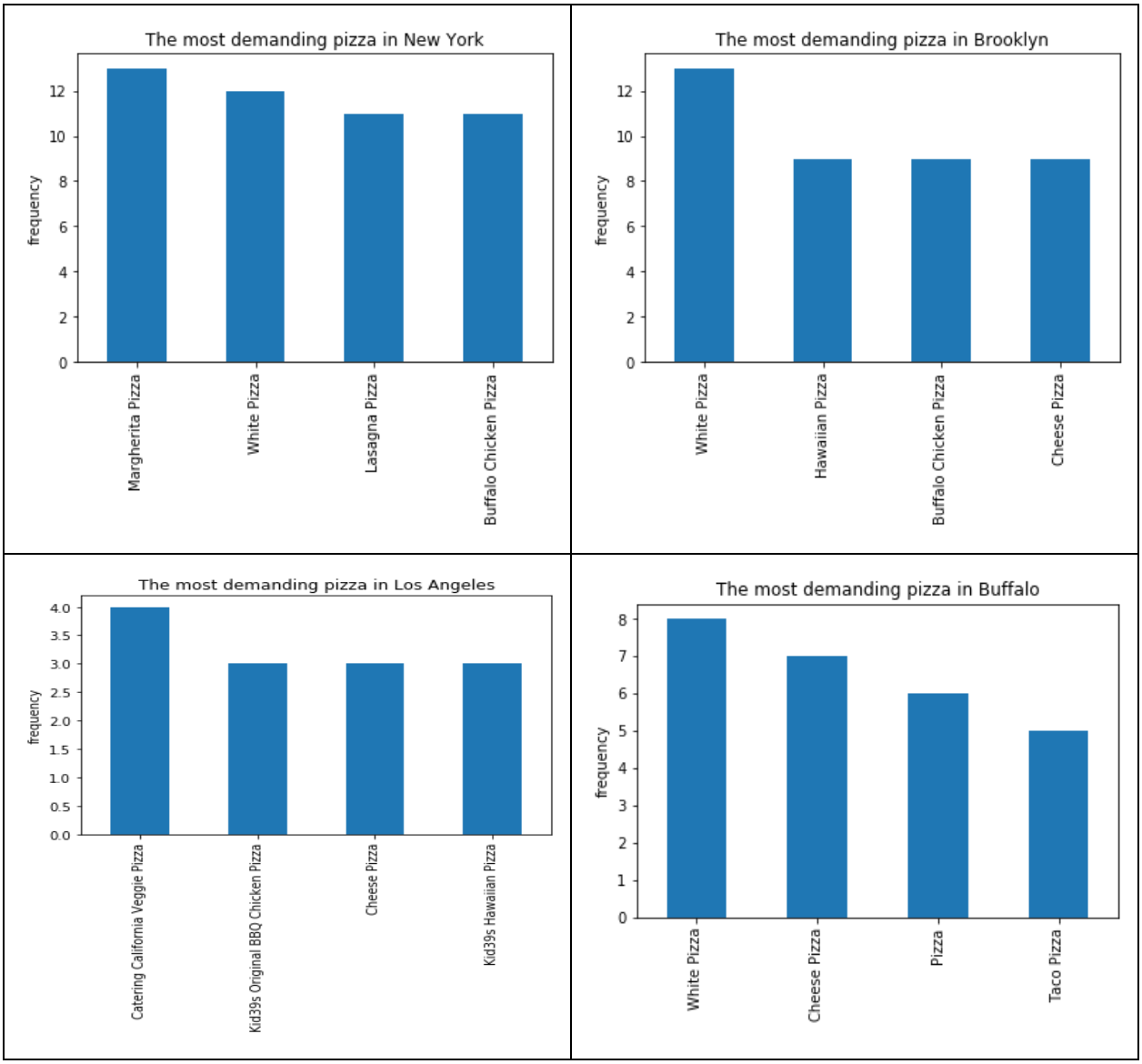
Most expensive

	name	menus.name	menus.amountMax
9337	Rocco's	Taco Pizza	1395.0

Cheapest pizzas

	name	menus.name	menus.amountMin
804	Fratellis Pizzeria	Pizza By the Slice	0.25
2777	DiAngelos	6" Pizza Sub	0.25
2778	DiAngelos	French Bread Pizza	0.25
7827	Stacia's Gourmet Pizza and Pasta	Garlic Herb Pizza Crust	0.25

Which are the most demanding pizzas in the top 4 cities?

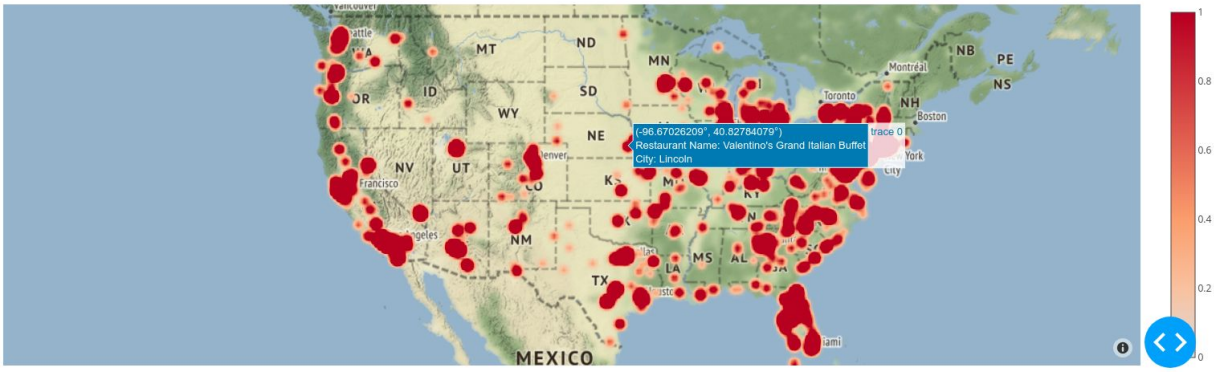


Dashboard Visualization

Interactive Data Analysis

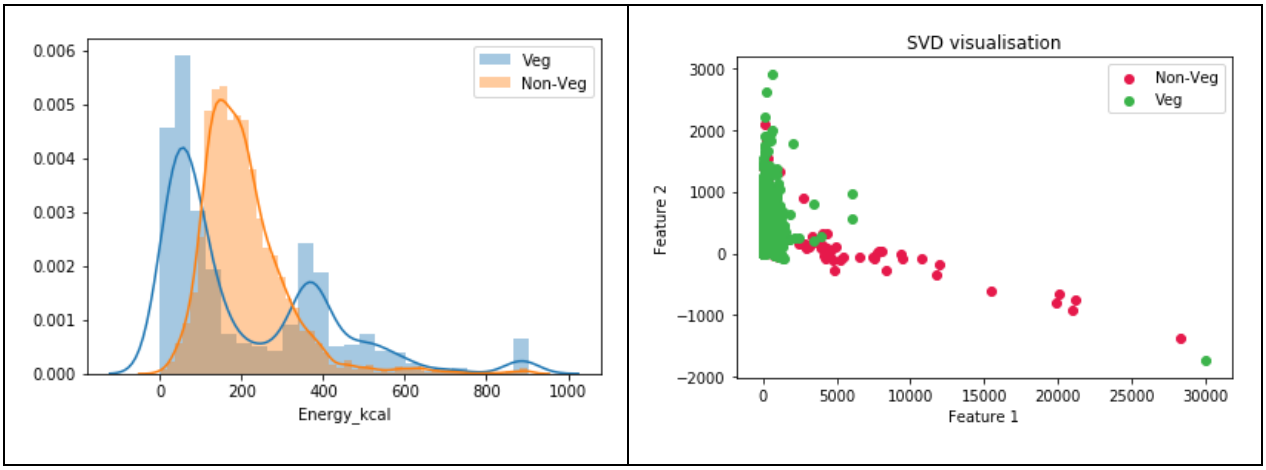
Restaurants at different locations in USA

Heatmap of Demand of Restaurants



Darker the region indicates higher demand

Visualization of nutrients dataset -



Visualization of Decision tree during classification -

5. Use of predictive modeling

- ## Association Rule Mining using Apriori Algorithm

  The association rule algorithm used to find out if the customer orders one type of pizza, then he will order which other pizza with it, using the support and confidence level as thresholds.
  We have applied the Apriori algorithm on the pizza orders from the dataset. The pizza orders consider the pizzas in one order as one list and we pass the list of these orders through the algorithm.
  We got a total of 56 rules in output. Some of the results are shown below in the table.

Rule	Support	Confidence	Lift
BBQ Chicken Pizza -> Buffalo Chicken Pizza	0.0048	0.239	5.57
Baked Ziti Pizza -> Lasagna Pizza	0.0035	0.296	15.38
White Pizza -> Baked Ziti Pizza	0.0048	0.407	4.97
Hawaiian Pizza Baking Required -> Big Murphy39s Stuffed Pizza Baking Required	0.0035	0.727	118.7
Buffalo Chicken Pizza -> Blt Pizza	0.003	0.5	11.65
Chicken Parmigiana Pizza -> Buffalo Chicken Pizza	0.0039	0.428	9.992
White Pizza -> Chicken Pizza	0.0052	0.5	6.109
Salad Pizza -> Lasagna Pizza	0.0052	0.272	21.48
Pizza Burger -> Pizza Fries	0.0056	0.309	10.88

Inference from the table for the 1st entry

The confidence level for the rule is 0.239 which shows that out of all the orders that contain BBQ Chicken Pizza, 23.9% of the orders also contain Buffalo Chicken Pizza.

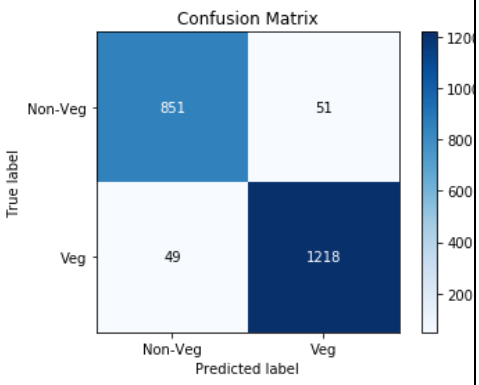
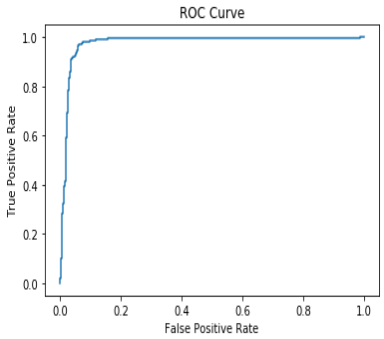
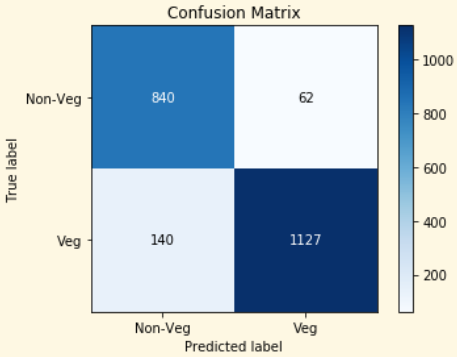
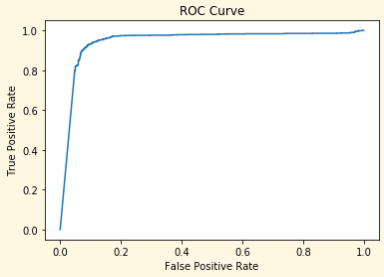
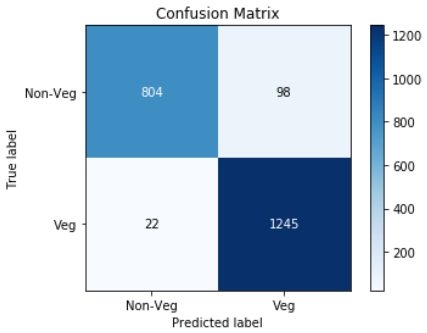
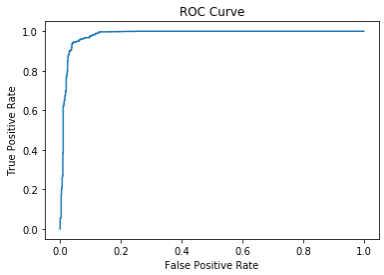
The lift of 5.57 tells us that BBQ Chicken Pizza is 5.57 times more likely to be ordered by the customers who order Buffalo Chicken Pizza.

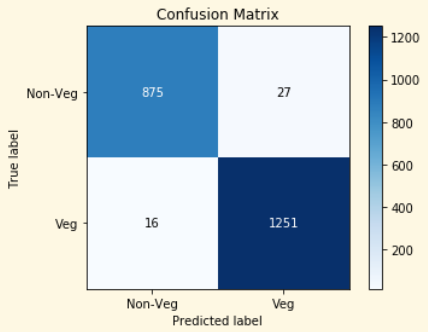
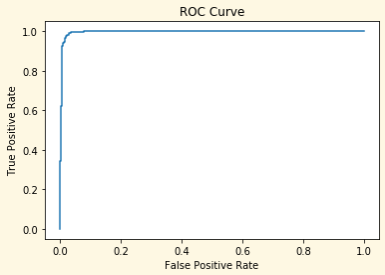
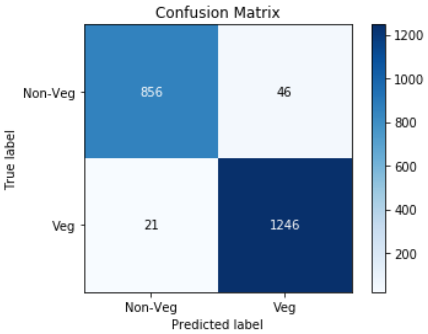
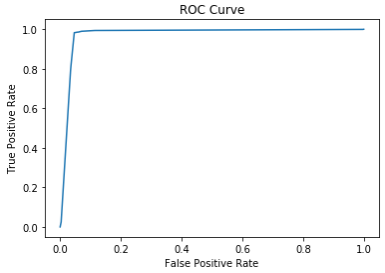
Similar inferences can be generated for other rules.

- **Classification**

  We have nutrients related to different pizza types. The pizza can be categorized as Veg Pizza and Non-Veg Pizza.

  This is a binary classification problem. We employ different machine learning classifiers for classification.

Classifier	Confusion Matrix	ROC
Logistic Regression		
Multinomial Naive Bayes		
Random Forest		

XG Boost		
Decision Tree		

Classifier	Accuracy	ROC AUC	F1-Score Macro Average	Macro Average Precision	Macro Average Recall
Logistic Regression	95.38	0.97	0.95	0.95	0.95
Multinomial Naive Bayes	90.6	0.94	0.91	0.9	0.91
Random Forest	94.46	0.98	0.94	0.95	0.94
XG Boost	98.01	0.99	0.98	0.98	0.98
Decision Tree	96.91	0.97	0.97	0.97	0.97

- **<u>Regression</u>**

  We have nutrients for different Pizza types. We are predicting Calories in a particular pizza. The calories are the estimated values of our model. A deep neural network with three hidden layers and the Relu activation function is used.

Model	MSE	RMSE	MAE	R - SQUARE
Linear Regression	286.09	16.91	6.67	0.98
Lasso Regression	282.03	16.79	6.76	0.99
Neural Network	413.08	20.38	6.51	0.985

6. Conclusion

The sale of pizza in North America increases in 2018. The Gradient Boosting classifier(XGBoost) outperforms others in terms of accuracy and AUC-ROC. From various association rules, we can see that there is a high chance of buying stuffed pizza if someone bought Hawaiian Pizza. The calories of a particular pizza are best estimated by the Lasso Regression model which has the largest R-Square score. The dashboard can be used to visualize all the interactive plots. This problem can be extended to any food product.

7. Reproducibility

Repository link: https://github.com/ajain0395/pizza-trend-analysis
Clone repository
See the readme file in the repository for further instructions

1. Interactive Plots (Dashboard)

Installation

1. pip install dash
2. pip install plotly

Execution

1. python Pizza_trends.py
2. Open browser and type URL "localhost:4253"

Script Running

Open Python jupyter-notebook in pizza-trend-analysis directory and open files

1. For Classification

a. Nutrition_Analysis_Classification.ipynb

2. For Regression

a. Nutrition_Analysis_Regression.ipynb

3. Association Rules

a. Pizza.ipynb

8. References -

1. https://data.world/sdhilip/pizza-datasets/workspace/file?filename=Pizza.csv
2. https://data.world/datafiniti/pizza-restaurants-and-pizzas-on-their-menus/workspace/file?filename=Datafiniti_Pizza_Restaurants_and_the_Pizza_They_Sell_Jun19.csv
3. https://data.world/craigkelly/usda-national-nutrient-db
4. https://towardsdatascience.com/probability-distributions-in-data-science-cce6e64873a7
5. https://www.kaggle.com/datafiniti/pizza-restaurants-and-the-pizza-they-sell
6. https://towardsdatascience.com/deep-neural-networks-for-regression-problems-81321897ca33
7. https://stackabuse.com/association-rule-mining-via-apriori-algorithm-in-python/