Principal Component Analysis for the Nutrients of pizza

Akash Nandigama
40112092

CONCORDIA UNIVERSITY

DEPARTMENT OF
INFORMATION SYSTEMS AND
SECURITY

***Abstract*— Pizza is one of the most commonly consumed food item among the world. According to a study 83% of consumers eat pizza once per month and with a five year forecast growth rate of 10.7% [1]. This project can be presented as a classification of nutrients and the principal components of pizza. The processed data obtained from the principal component analysis is checked for accuracy and complexity is reduced by using algorithms such as logistic regression, support vector machine and K-nearest neighborhood.**

Keywords—Principal Component Analysis, Pizza nutrient classification, Logistic Regression, Support Vector Machine, KNN.

I. INTRODUCTION

Pizza is known as a famous food item around the globe, it is most widely consumed among customers from different age groups. According to a study 83% of consumers eat pizza in a month [1]. So we can imagine the scale of consumption and business statistics happening every year, billions of dollars of business happens in this business across the globe in a year and yet only few brands reach that top level of handling this business.

The taste of pizza depends on the ingredients used, toppings, sauce and cheese. Each company will have their own recipe of making this ingredients turn into a good tasting pizza. Pizzerias are continuously working on the technological advancements to derive the best ingredients combination to make pizza tastier which in turn increases the sales.

In the domain of machine learning classification is used to provide the insights and necessary information for such kind of data where the end goal is to know which pizza brand works for us and measurements of making the pizza tastier so that it would be a great seller.

Increment in the quantity increases the unpredictability and time required for calculation. In this report PCA is used to reduce dimensions and for evaluation purposes logistics Regression, KNN and support vector machine algorithms are used and their results are compared in terms of fitness time and accuracy.

II. PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis is used to reduce dimensionality of huge number of co-related attributes in a particular dataset. It basically converts data from correlated to un- correlated. It is basically a linear transformation of

Multidimensional features into features which are not correlated called as Principal Components that carry most of the important information of the high dimensional data [4].
PCA Algorithm consists of 5 steps:

1. Ignoring the class labels from the dataset such that it becomes n+1 dimensional to n dimensional dataset and further be easy for classification of principal components.
2. Data standardization is done by computing the mean for each dimension in the dataset i.e. computing the centered data matrix by subtracting column means.

$$Y= HX$$

3. Covariance matrix is calculated by and covariance matrix is a square matrix (p*p).

$$S = \frac{1}{n-1} Y'Y$$

4. Eigen values and Eigen vectors of S are calculated by using Eigen decomposition. After obtaining the eigen values the common approach used is to rank the eigenvectors from highest to lowest corresponding eigen values and choose eigen values with top magnitude.

$$S = A\Lambda A' = \sum_{j=1}^{p} \lambda_j \mathbf{a}_j \mathbf{a}_j'$$

5. Transformed data matrix (n*p) is computed by

$$Z=YA$$

By carrying out the above steps PCA can be performed and dimensionality reduction of the selected data can be achieved.

III. CLASSIFICATION ALGORITHM

Classification is the process of identifying, understanding and grouping ideas together into categories. Using these pre-categorized data sets machine learning can be implemented to estimate future datasets into categories.

The classification algorithms in machine learning use training data as input and estimate the likelihood that subsequent data falls into pre-categorized datasets. As an example most of these algorithms are used to categorize spam and non-spam emails that we receive.

There are several types of classification algorithms in machine learning, most commonly used algorithms are

- Logistic Regression
- K- Nearest Neighborhood
- Support vector machine

A. Logistic Regression

Logistic regression is the calculation used to estimate the binary outcome for example, Yes/No, Pass/Fail. Binary regression logistic model has two outputs which are categorical whereas independent variable can be either categorical or numerical but dependent variables are categorical. It is used to calculate the dependent variable Y when independent variable X is given [5].

$$P(Y=1|X) \text{ or } P(Y=0|X)$$

B. K-Nearest Neighborhood

K-nearest neighborhood (KNN) is one of the most used algorithm in data mining and machine learning, it uses datasets to find the closest distant variables in future data. In classification phase k is user defined constant variable and an unlabeled vector is determined by assigning the label which is most nearest to the query point among the k training samples [6].

Euclidian distance or weighted distance is the commonly used distance metric in KNN algorithm. Whereas the Euclidian distance formula is given as follows.

Using the below formula we can calculate the distance between one point and the other in the dataset and smaller the distance indicates how closer the data is related.

$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\cdots+(q_n-p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

C. Support vector Machine

Support vector machine is used for representation of vector points in space in such a way that are divided into categories as wide as possible. Vector points which are added later are mapped into the space and categories are assigned by prediction on the amount of side gap.

SVM can perform both linear and non-linear classification efficiently. This algorithm will map the inputs into high dimensional feature spaces and an unsupervised learning Approach used for analyzing un- labeled data. There are 2 types of SVM known as simple SVM which is used for linear regression and classification problems and Kernel SVM used for non-linear data classification [3].

IV. EXPERIMENTAL RESULTS OF CLASSIFICATION

D. Data Classification

The data used in this analysis is Pizza nutrient classification. This data is chosen from kaggle [2]. It has 9 columns which includes the class labels that correspond to the classification of brand of pizza. Some of the labels include proteins, fat, sodium levels etc. Below figure depicts the first 25 rows of pizza data set.

	A	B	C	D	E	F	G	H	I
1	brand	id	mois	prot	fat	ash	sodium	carb	cal
2	A	14069	27.82	21.43	44.87	5.11	1.77	0.77	4.93
3	A	14053	28.49	21.26	43.89	5.34	1.79	1.02	4.84
4	A	14025	28.35	19.99	45.78	5.08	1.63	0.8	4.95
5	A	14016	30.55	20.15	43.13	4.79	1.61	1.38	4.74
6	A	14005	30.49	21.28	41.65	4.82	1.64	1.76	4.67
7	A	14075	31.14	20.23	42.31	4.92	1.65	1.4	4.67
8	A	14082	31.21	20.97	41.34	4.71	1.58	1.77	4.63
9	A	14097	28.76	21.41	41.6	5.28	1.75	2.95	4.72
10	A	14117	28.22	20.48	45.1	5.02	1.71	1.18	4.93
11	A	14133	27.72	21.19	45.29	5.16	1.66	0.64	4.95
12	A	14101	27.35	21.2	45.59	4.94	1.65	0.92	4.98
13	A	14108	26.98	21.2	45.03	5.15	1.67	1.64	4.97
14	A	14164	28.7	20	45.12	4.93	1.56	1.25	4.91
15	A	14154	30.91	19.65	42.45	4.81	1.65	2.81	4.72
16	A	24005	30.91	20.77	42.03	4.9	1.61	1.39	4.67
17	A	24026	30.83	17.88	44.33	5.26	1.76	1.7	4.77
18	A	24094	32.73	20.06	39.74	5.24	1.69	2.23	4.47
19	A	24108	34.58	17.53	40.87	5.05	1.61	1.97	4.46
20	A	24102	31.8	20.35	40.44	5.43	1.61	1.98	4.53
21	A	24082	31.02	19.05	42.29	5.27	1.71	2.37	4.66
22	A	34017	27.02	19.56	47.2	4.95	1.65	1.27	5.08
23	A	34020	27.78	20.01	45.59	4.97	1.7	1.65	4.97
24	A	24136	30.88	20.58	42.26	4.96	1.63	1.32	4.68
25	A	24122	32.2	19.25	43.42	4.62	1.5	0.51	4.7

Figure 1: Sample 25 rows of the original data-set.

The parameters in the above dataset are
mois -- Amount of water per 100 grams in the sample
prot -- Amount of protein per 100 grams in the sample
fat -- Amount of fat per 100 grams in the sample
ash -- Amount of ash per 100 grams in the sample
sodium -- Amount of sodium per 100 grams in the sample
carb -- Amount of carbohydrates per 100 grams in the sample
cal -- Amount of calories per 100 grams in the sample

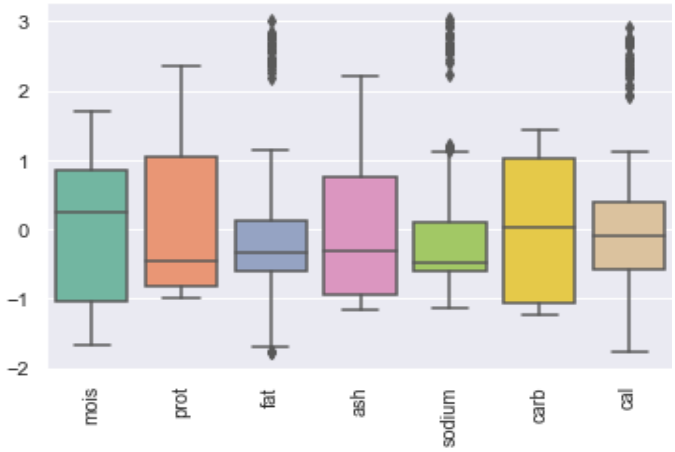

Figure 2: Boxplot of the data-set

Box plot is used for depicting groups of numerical data in form of the quartiles. It is a graphical representation of the information which consists of 3 quartiles such as first, middle and greatest quartile. The shape of box plot is in square with lines reaching out from top and base. Box plot of each feature of the dataset is presented above and we can observe that fat, sodium and cal attributes have outlier points.

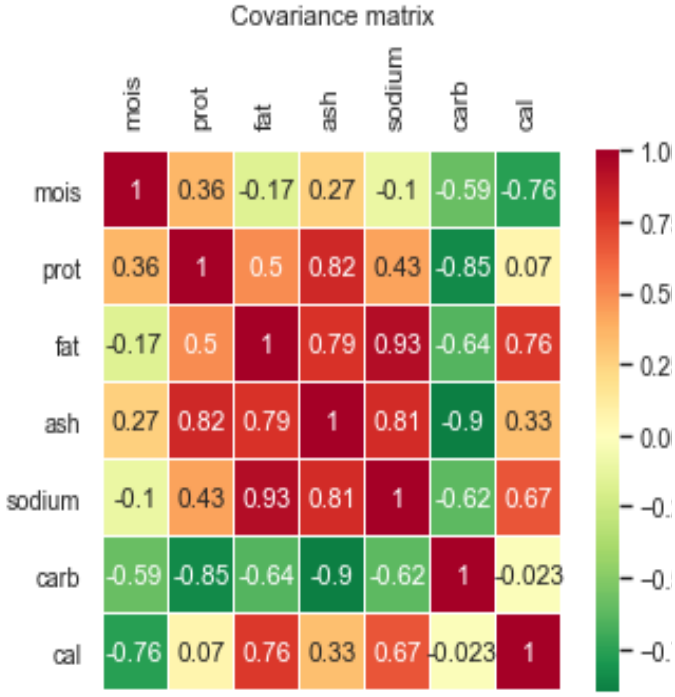

Figure 3: Covariance matrix of the features

Figure 4 shows the covariance matrix of the features and is used to observe how different the features are correlated to each other. It is a matrix whose components in the x,y position depicts the covariance between the equivalent relative components of an vector whereas every component in the vector is a scalar irregular variable.

E. PCA THEORY

Principal Component Analysis is a unique multivariate algorithm applied for dimensionality reduction, feature extraction and data visualization. PCA is designed by conversion of high dimensional vector space into a low dimensional space.

The purpose of PCA data is to reduce the features in the data. There are 9 attributes in this data in which one attribute is class label and the other attribute is id of the record. So, in remaining 7 attributes after applying PCA to the dataset the number of attributes have been reduced to 4 as shown in pareto diagram and the first two attributes are considered as principal components to reduce dimensionality as they contribute most of the variance.

After Computing PCA, the following are key results Eigen vector matrix is described as follows:

Eigen Vector Matrix

$$\begin{pmatrix} 0.064 & -0.628 & -0.421 & -0.220 & -0.006 & 0.446 & 0.481 \\ 0.378 & -0.269 & 0.746 & -0.010 & -0.387 & -0.0001 & 0.276 \\ 0.044 & 0.234 & -0.199 & -0.507 & 0.173 & -0.525 & 0.377 \\ 0.471 & -0.110 & 0.056 & 0.552 & 0.670 & 0.058 & 0.056 \\ 0.435 & 0.201 & -0.455 & 0.446 & -0.602 & 0.003 & -0.0005 \\ -0.42 & 0.320 & 0.052 & 0.334 & 0.007 & 0.0005 & 0.776 \\ 0.244 & 0.567 & 0.113 & -0.279 & 0.078 & 0.721 & 0.012 \end{pmatrix}$$

The column values of the data are considered as principal components.

Z1 = 0.064(mois) – 0.625(prot) – 0.421(fat) – 0.220(ash) – 0.006(sodium) + 0.446(carb) + 0.481(cal)

Z2 = 0.378(mois) – 0.269(prot) + 0.746(fat) – 0.010(ash) – 0.387(sodium) – 0.0001(carb) + 0.276(cal)

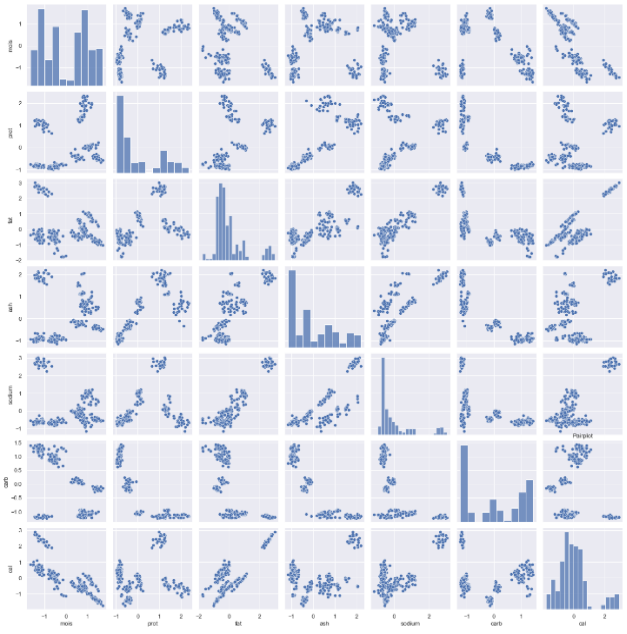

Figure 4: Scatter plot

It is also known as pair plot where one variable in similar information is coordinated with another variable. The diagonal of this plot is represented in histogram which helps us to observe the approximation of a single variable and the scatter plots above and below this diagonal is used to depict the relationship between two factors.

The below figure shows the scatter plot of PC2 coefficients to PC1 coefficients. From the figure we can observe that carb variables are plotted on the left side and all the other variables

are plotted on the right side.

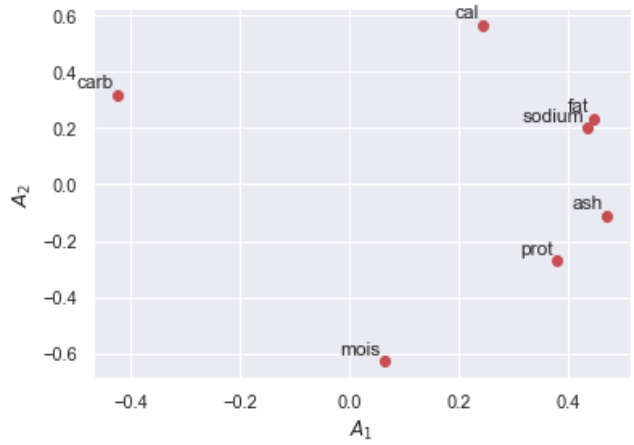

Figure 5: Scatter plot of PC2 Coefficients vs PC1 Coefficients

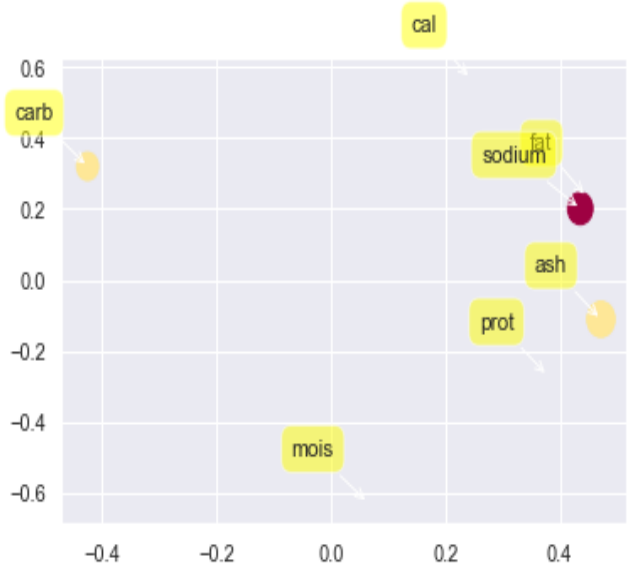

Figure 6: PC2 Coefficients vs PC1 Coefficients

Below Pareto chart explains the variance of the features. It is plotted between number of components and explained variance. If we consider the first two components, the cumulative variance will be around 90%. So, in order to obtain more than 80% accuracy, we can consider the first two components.

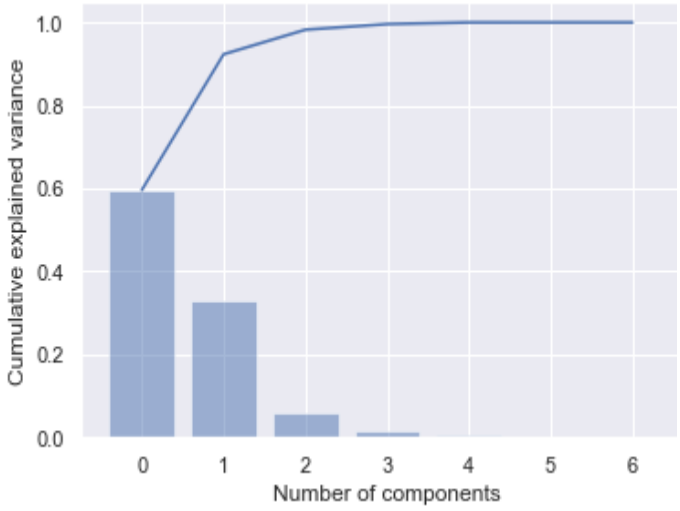

Figure 7: Pareto plot

Scree plot is used to observe the division of total variance in the information as represented by every principal component. In the figure shown below Component 1 has higher variance and from the component 2 the graph starts decreasing at an exponential rate resulting the least value component for 4, 5, 6 and 7.

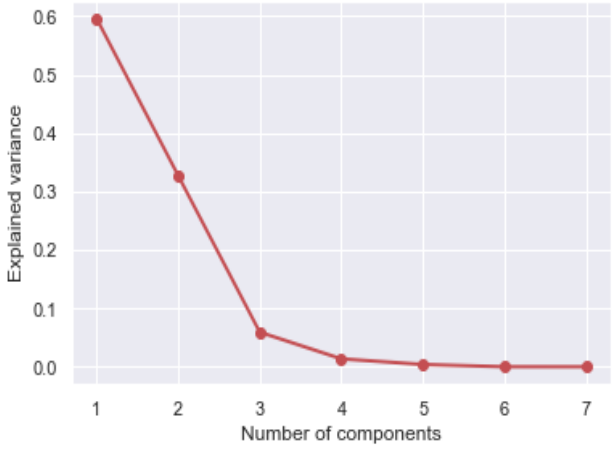

Figure 8: Scree plot

An auto encoder is used to learn information coding's in a unique way. The objective of an auto encoder is to implement in encoding for a strategy of information, dimensionality decrease. On the decay side, a reproducing side is recognized which is used by auto encoder to make the reduced portrayal as close as possible.

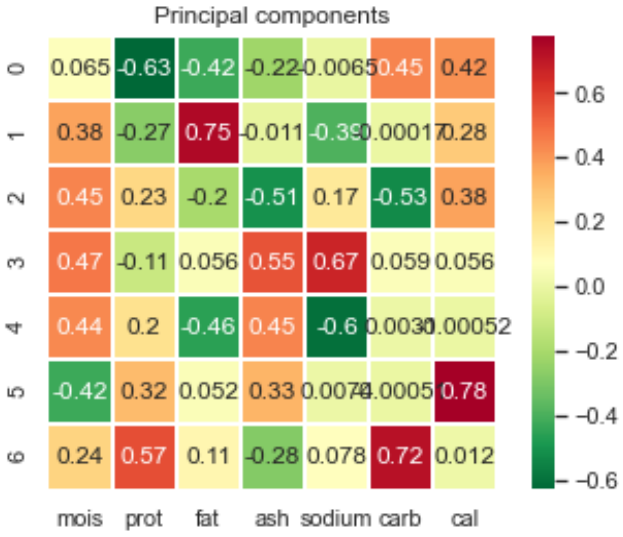

Figure 9: Covariance matrix of Principal Components.

Bi plot is a graphical representation of displaying factors and test units at the same time by multivariate information matrix. A PCA bi plot is used to display segment scores and variable loadings acquired by PCA in various dimensions. In the below graph A, B, C, D, E, F, G, H, I and J are class labels.

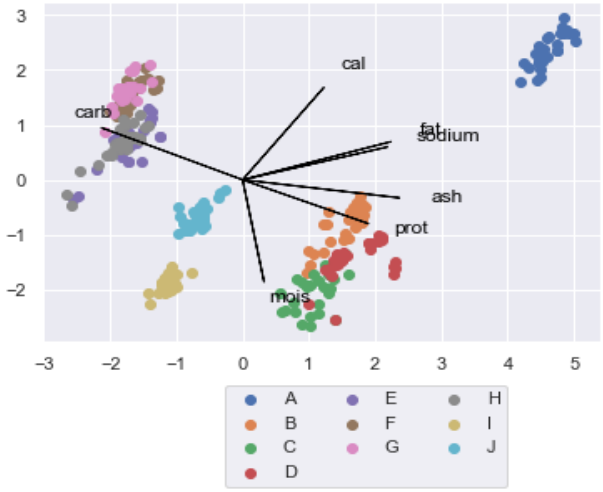

Figure 10: Bi plot

V. PCA CONTROL CHARTS

Multivariate control graphs are widely utilized in lot of businesses to analyze forms described by large number of attributes. Control charts in PCA are used to portray issues presented by high connections by varying the arrangement of related factors to uncorrelated arrangement of factors. Advancements of non-parametric PCA control outlines that it doesn't require any distributional suppositions.

The Hotelling control chart is an enhancement of X bar chart. It provides feasibility for observations to be plotted in scatter plot [7]. The shifts in mean of more than a variable can be observed in this chart. The statistic for a single variable is as follows:

$$T^2 = (\mathbf{x} - \mathbf{x})' S^{-1} (\mathbf{x} - \mathbf{x})$$

We know that

$S^{-1} = A\Lambda^{-1}A'$ and

$z_i = A'xi$

So, $T_i^2 = Z_i' \Lambda^{-1} Z_i$

Upper control limits of phase1 and phase 2 are as follows:

$$UCL = \frac{(m-1)^2}{m} \beta_{\alpha, \frac{p}{2}, (q-p-1)/2}$$

UCL = $P(n+1)(n-1)/\, n(n-P)\, F\alpha$,p,n-p

The below figure predicts Hotelling's T2 control chart

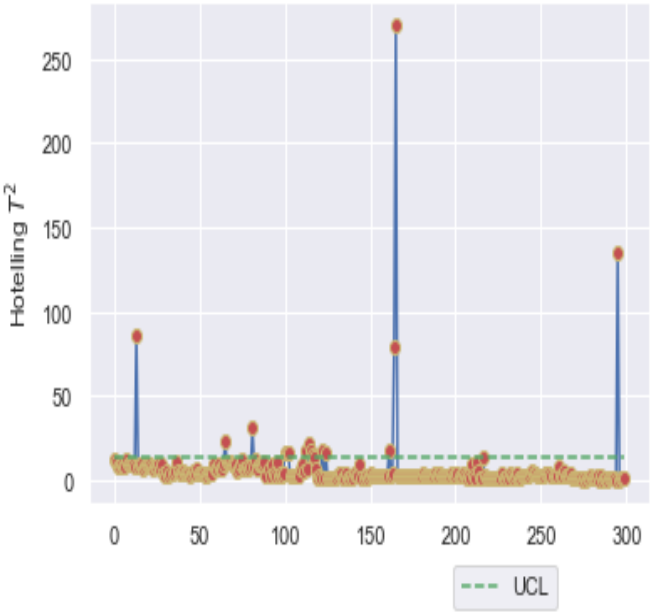

Figure 11: Hotelling's T2control chart

From the figures 11 and 12 both the charts have respective control limits. In figure 12 upper control limit is indicated by a green line, which states that any value which crosses that line is considered as outliers. In figure 12 we can know that 16 samples are out of control. In figure 13 any values which is between UCL and LCL are considered as samples under control.

PCA control chart for the first component is as follows:

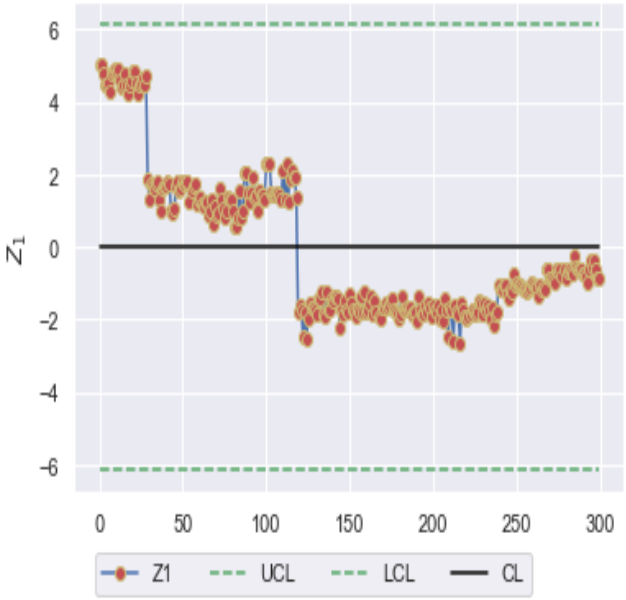

Figure 12: PC Control chart

In the above chart control limits are calculated by using below formulae.

UCL = $3\sqrt{\lambda j}$

CL = 0

LCL = $-3\sqrt{\lambda_j}$

VI. CLASSIFICATION USING MACHINE LEARNING ALGORITHM

Original data is considered and after the dimensionality reduction is done, two principal components are extracted. Logistic regression, Support vector machine and K-nearest neighborhood machine learning algorithms are used as classifiers. These algorithms are used to conduct experiments on both original dataset and two principal components. Fit time and accuracy results are presented for each algorithm.

Five - Fold Accuracy table for Logistic Regression

Fold	Original Dataset	Two principal components
1	0.85	0.78
2	0.80	0.78
3	0.88	0.76
4	0.83	0.76
5	0.93	0.85
Average	0.85	78.6

Five- fold Accuracy table for K-Nearest Neighborhood

Fold	Original Dataset	Two principal components
1	0.85	0.83
2	0.83	0.81
3	0.91	0.91
4	0.86	0.85
5	0.88	0.88
Average	0.86	0.85

From the above table we can observe that accuracy is almost equal on the original dataset and the two principal components.

Five - fold Accuracy table for Support Vector Machine

Fold	Original Dataset	Two principal components
1	0.8	0.75
2	0.8	0.78
3	0.81	0.71
4	0.81	0.76
5	0.78	0.75
Average	0.8	0.75

In terms of accuracy K-Nearest Neighborhood algorithm is 86% accurate on the original dataset and 85% accurate when applied by considering principal components. Accuracy of KNN is higher when compared to logistic regression and SVM.

Five - Fold Fit time table for Logistic Regression

Fold	Original Dataset	Two principal components
1	0.046	0.031
2	0.046	0.046
3	0.031	0.046
4	0.046	0.046
5	0.046	0.031
Average	0.04	0.04

From the above table we can conclude that fit time for original dataset and two principal components is equal.

Five – Fold fit time for K-Nearest Neighborhood

Fold	Original Dataset	Two principal components
1	0.002	0.002
2	0.002	0.003
3	0.002	0.001
4	0.004	0.003
5	0.003	0.001
Average	0.026	0.002

Five – Fold fit time for Support Vector machine

Fold	Original Dataset	Two principal components
1	0.004	0.002
2	0.003	0.003
3	0.004	0.001
4	0.004	0.001
5	0.003	0.002
Average	0.036	0.0016

In terms of fit time for both the algorithms KNN and SVM fit time for original dataset is higher than the fit time for two principal components.

VII. CONCLUSION

This project is carried out by applying PCA on pizza dataset to reduce the dimensionality and deriving principal components. Classification algorithms Logistic Regression, K-Nearest Neighborhood and Support Vector Machine are used for calculating accuracy and fit times in which among these three algorithms it is derived that KNN has highest accuracy (85%) when principal components are considered and in terms of fitness time KNN has lowest fit time on the original data and SVM has the lowest fit time on the principal components. Various charts and static measures are used to summarize the

findings. It is also observed that score the data before applying PCA is 76.6% and after applying the PCA is 83.3% for logistic regression which shows that the components which we have considered are contributing most to the data and our principal components are valid.

VIII. REFERENCES

[1] https://www.pmq.com/the-2019-pizza-power-report-a-state-of-the-industry-analysis/#:~:text=According%20to%20a%20Technomic%20study,forecasted%20growth%20rate%20of%2010.7%25.

[2].https://www.kaggle.com/shishir349/can-pizza-be-healthy

[3]. https://en.wikipedia.org/wiki/Support-vector_machine

[4].https://towardsdatascience.com/the-mathematics-behind-principal-component-analysis-fff2d7f4b643

[5].https://monkeylearn.com/blog/classification-algorithms/

[6] https://towardsdatascience.com/knn-k-nearest-neighbors-1-a4707b24bd1d

[7] https://www.spcforexcel.com/knowledge/variable-control-charts/hotelling-t2-control-chart