Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/HK3SF

COMPARATIVE CASE STUDY: AN EVALUATION OF PERFORMANCE COMPUTATION BETWEEN SUPPORT VECTOR MACHINE, K-NEAREST NEIGHBORS, K-MEAN, AND PRINCIPAL COMPONENT ANALYSIS

MUHAMMAD ZOHAIB KHAN
Department of Computer Science, Sindh Madressatul Islam University, Karachi, Sindh, Pakistan.
ABDULLAH AYUB KHAN*
Department of Computer Science, Sindh Madressatul Islam University, Karachi, Sindh, Pakistan.
Faculty of Computing Science and Information Technology, Benazir Bhutto Shaheed University Lyari, Karachi, Sindh, Pakistan.
***Corresponding Author**
ASIF ALI LAGHARI
Department of Computer Science, Sindh Madressatul Islam University, Karachi, Sindh, Pakistan.
ZAFFAR AHMED SHAIKH
Faculty of Computing Science and Information Technology, Benazir Bhutto Shaheed University Lyari, Karachi, Sindh, Pakistan.
MUHAMMAD ADNAN KAIM KHANI
Department of Computer Science, Sindh Madressatul Islam University, Karachi, Sindh, Pakistan.
DMITRY MORKOVKIN
Financial University under the Government of the Russian Federation, Moscow, Russia.
OLGA GAVEL
Financial University under the Government of the Russian Federation, Moscow, Russia.
SERGEY SHKODINSKY
Financial Research Institute of the Ministry of Finance of the Russian Federation, Moscow; Russia.
RUDN University, Moscow, Russia.
SVETLANA MAKAR
Financial University under the Government of the Russian Federation, Moscow, Russia.
National Research Mordovia State University, Moscow, Russia.
DENIS TABUROV
Financial University under the Government of the Russian Federation, Moscow, Russia.

Abstract

The rapid advancement in Information Technology (IT) and the development of Artificial Intelligence (AI) makes systems more efficient and effective in performing day-to-day tasks, such as identification, extraction, detection, and recognition-related problems. These pose a serious indication towards the concept of Machine learning (ML) and the proposed efficient techniques for distinct purposes, which are either performed artificially supervised, semi-supervised or unsupervised ML. However, the ML systems have the potential to self-learn and adapt themselves by explicitly programmed existence based on earlier experience. ML and data mining approaches are justified in this research. It is featured in the Principal Component Analysis (PCA) method of supervised and unsupervised learning. Both machine learning algorithm approaches use a variety of methodologies. For this purpose, we use the same datasets for classification, regression, and clustering procedures, as well as the PCA data mining technique, to evaluate and perform on the K-Means, K-nearest neighbor (KNN), and support vector machine (SVM) algorithms. Finally, we present simulations that show the estimated criteria, parameters, and efficiency with effective algorithms' performance, and state-of-the-art results.

Keywords: Machine Learning, Supervised Learning, Unsupervised Learning, Support vector machine (SVM), K-Means, Principal component analysis (PCA).

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/HK3SF

1. Introduction

Nowadays, mostly in the industry, several machine learning systems for artificial intelligence are used. Some of them are supervised learning, while some are un-supervised. In this era, the trend of Artificial Intelligence is turning the world to artificial intelligence [1]-[11]. The subbranch of machine learning. We have presented two different categories: circle and square. Supervised learning works on structured or labelled data, and we can provide different data to the machine. Then, supervised learning has learned from prior data and applies it to current data (new data). We provide model training knowledge, then the machine will predict that it is a circle or square. This supervised learning method is used in the classification and regression of subcategories [12]-[18]. Unsupervised learning works on unstructured or hidden data where we provide the machine with different data. Unsupervised learning observes the data, then classifies the data, as a result, once classification, we acquire different data in each group, and each group is distinct from the others.

1.1. Used of artificial intelligence and machine learning

The computer science industry is Artificial Intelligence (AI), and Machine Learning is a sub-sector of AI [19]-[23]. This emphasizes the creation of smart machines, problem-solving, learning about thinking, and human-like activities. In several companies, manufacturing, weather climate change, healthcare, managing an account, website, and numerous other industries Artificial intelligence (AI) and machine learning are increasingly widely used technologies. Face and voice recognition systems, biometrics, robotic surgery, remote sensing applications, and others, all are the examples.

1.2. Comparison supervised and un-supervised learning

We evaluate both supervised and un-supervised learning methods with the application of these techniques followed by PCA, SVM, KNN and K-Mean algorithms are all learning methods used [24]-[28]. These algorithms are used in the techniques of classification, regression, and clustering. We justify the outcomes of quick, accurate, and better accuracy, recall, accuracy, classifier, prediction, and clusters.

- The major contributions of this paper are discussed as follows:
- Initially, we studied various research related to machine learning techniques and the procedure of executions. And so, design a comparative analysis mechanism that evaluates the computational difference between support vector machine, K-nearest neighbors, K-mean algorithm, and principal component analysis.
- In this paper, we present a comparative analysis of different ML algorithms and report the evaluation criteria, such as linear, non-linear with distinct kernel functions. It also includes the measurement of algorithm matrixes with the help of Euclidean, Manhattan, and Minkowski.
- Finally, we compared the proposed comparative analysis with other state-of-the-art studies.

The reminder of this paper is organized as follows. In Section 2, we study various related literatures and evaluate the analysis matrixes for comparative studies. Section 3 discusses the measurement of performance analysis of principal component

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/UWJK5

analysis, support vector machine, K-nearest neighbors and K-mean algorithm. The experimental results are presented and discussed the comparison of the proposed study with other state-of-the-art methods in Section 4. Finally, we conclude this paper in Section 5.

2. Background

2.1. Classification, regression and clustering in machine learning

Classification is the systematic method of arranging mechanisms into different categories and sub-categories based on their similarities. There are different types of classification in machine learning (Linear Regression, Nave Bayes Classifier, and Linear Classifier). Classifications include structured and labelled data. For example, in the following figure 1, which shows different components of classification which are used for various activities. Because different regression models differ based on the type of relationship between dependent and independent variables, linear and non-linear regression algorithms are based on supervised and unsupervised learning methods, respectively. It performs regression tasks. There are many regression features used in machine algorithms (linear and non-linear), including structured and unstructured data as shown in figure 3. Regression Model features one and two are both features used in linear and nonlinear regression. Clustering is the primary form of unsupervised learning, and clustering has a multitude of uses with common implementations in a multitude of companies. Clustering is that the action of separating and executing information on behalf of the information machine makes a set of data called clustering and we assign special (ID) to each cluster. Figure 3 shows the different clusters [29]-[33].

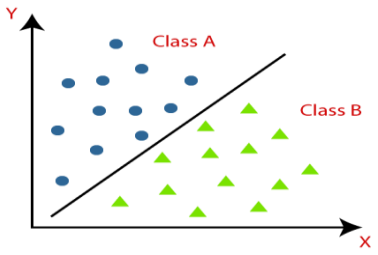

Figure 1. Classification [17]

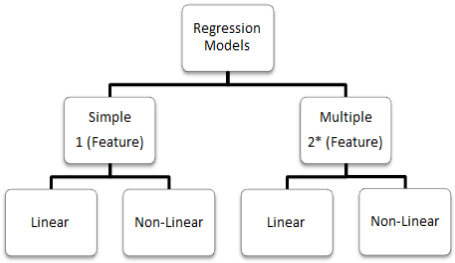

Figure 2. Regression Models

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/UWJK5

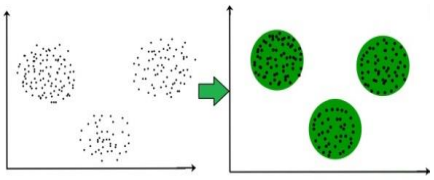

Figure 3. Clustering [18]

2.2. Data mining and the role of principal component analysis (PCA)

Data mining is a technique for finding patterns in large databases, and it includes methods for combining machine learning, analytics, and database systems. DM is a growing sub-discipline of machine learning and deep learning with the goal of extracting knowledge (using sophisticated methods) from large data sets and translating it into an understandable context for further growth. The method of knowledge discovery is data collection (KDD It also covers data processing, pre-processing, simulation and induction prerequisites, importance criteria, complexity considerations, post-processing of discovered constructs, simulation and instructional models, and data mining using various approaches. PCA is a dimensional-reduction method that is commonly used to assist with the estimations of huge sets of data by splitting a number of variables into smaller ones and incorporating most of the data from an information machine into a vast collection of data. Attempting to decrease the rise in test bunch parameters lowers the consideration for precision, but the trap in reducing dimensionality is to trade a bit immovable quality for simplicity of usage [34, 35, 36]. Simpler sets of data are easier to inspect and recreate, making data examination easier and faster for machine learning models with a large number of free components to test. Preprocessing, decay measurement reduction, covariance and relationship eigenvalues and eigenvectors methods are all used in PCA.

2.3. Support vector machine (SVM) algorithm

It is based on the principle of supervised learning. SVM is a classification and regression algorithm. This method is utilized in supervised approaches and functions as a partition [36, 37]. for instance, the graph's X and Y components, are mixed classes. We must now divide two classes. Between the two classes, we'll draw a line. The Maximum Margin Hyperplane is what it's named. The support vector is the smallest data point between two classes. Linear, Nonlinear, and Kernel (Polynomial, Radial Base Function (RBF), and Sigmoid) functions are among the mathematical functions that utilize SVM. Figure 4 depicts the X1 and X2 components, as well as two classes: quadratic and circular. The maximum margin hyperplane separates the two classes, and the maximum margin is represented by two dotted parallel lines with hyperplane. Support vectors are the two data points closest to the greatest margin.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/UWJK5

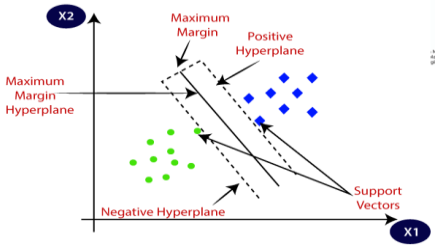

Figure 4. Support Vector Machine (SVM) [19]

The main steps of the Support Vector Machine (SVM) algorithm work as follows:

1. Explain the optimal hyper plane to optimize the margin.
2. Expanding the over-definition with non-linearly distinct problems has a term of penalty for misclassification.
3. Project information to high dimensional space, where it is less demanding to define surfaces with direct preference, reformulates the problem so that information is verifiably translated to this storage.
4. End.

These are briefly defined some mathematical classification and regression function used in (SVM) algorithm. Linear SVM, Non-linear SVM, and Kernel function. These functions are given below:

Linear SVM	***xi.xj***
Non-linear SVM	***ϕ (xi).ϕ (xj)***
Kernel function	***k(xi.xj)***

2.4. K-nearest neighbours (KNN) algorithm

K-Nearest Neighbors (KNN) algorithm is a supervised and unsupervised machine learning algorithm that can be used for both classifications almost along with regression predictive problems the purpose of this method dividing the data into Classes from mixed or unstructured, and structured or labeled data [38, 39, 40]. This algorithm in dataset starts with a first Class of randomly selected nearest value. Which are used in beginning point for every nearest value and then perform repetitive calculation to optimize the position of the center values. KNN is a poor and lazy machine learning algorithm since it does not have a dedicated training process and uses all the training data when classifying and is also a non-parametric learning algorithm since it assumes little about the embedded results. The key step of the KNN method works as follows:

1. We need a dataset to execute any measurement. So, in the middle of the primary step of (KNN), we have to stack both the planning and the test results.
2. After that, we have to pick K's respect, for example, the closest knowledge point. K must be the value of the numbers.
3. At each stage of the test detail, take the following:

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/UWJK5

3.1 Measure the gap between test information and each drive to plan information with the help of each technique to be specific: Euclidean, Manhattan or Hamming separately. Euclidean is the most widely used method to measure the removal.
3.2 Now, on the basis of distance esteem, sort them in rising order.
3.3 After that, the best K columns will be picked from the sorted list.
3.4 Now, the course will be relegated to the test point depending on several of the visits of this line.
4. End.

The following figure 5 specifically described KNN as classified data from category A, and category B.

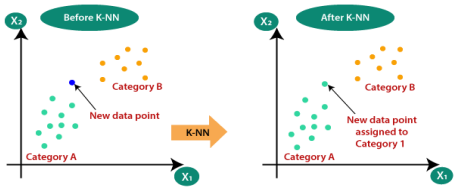

Figure 5. KNN Classification [20]

These are briefly well-defined some mathematical distance methods or functions used in KNN algorithm. Euclidean, Manhattan or Hamming distance and this method work on classification. These methods are given below:

Euclidean $\sqrt{\sum_{i=1}^{k}(xi-yi)2}$

Manhattan $\sum_{i=1}^{k}|xi-yi|$

Minkowski $\left(\sum_{i=1}^{k}(|xi-yi|)q\right)1/q$

2.5. K-mean algorithm

The unsupervised K-mean algorithm is a machine learning approach. This technique divides data into mixed or unstructured categories. The dataset begins with a first group of randomly picked mean values, which serves as a starting point for each subsequent group, and then executes repetitive computations to optimize the position of the center values. The K-means algorithm functions on the following principles:

1. Discover out the number of K clusters.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/UWJK5

2. Compute the centroids by beginning with the reordering of the dataset and then randomly selecting the K information for the centroids without substitution.
3. Identifying the crevices until there is little development inside the centroids. The process of concentrating knowledge on clusters does not shift.
4. Calculate the totality of the squared distance between the focal points of the details and all the centroids.
5. Designate each information Point to the nearest whole number (centroid).
6. Calculate the centroids for the clusters by calculate the cruel of all the information focuses having a place to each cluster.
7. End.

The following Figures 6 are clearly defined K-means or clustering data from sample to cluster.

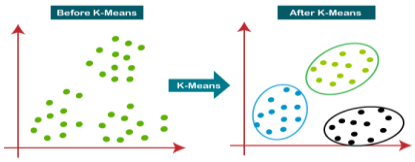

Figure 6. K-means Clustering [21]

3. Performance analysis of principal component analysis (PCA), support vector machine (SVM), K-nearest neighbors (KNN) and K-mean algorithm

Python is a script-specific open source and high-level language used mainly for work and machine learning algorithms, web developments, and databases. We have been used the GUI framework Anaconda Navigator ->Jupiter Notebook. We connect dataset and implemented PCA, SVM, KNN, and K-Means algorithms through python programming language. Basically, dataset work on breast cancer disease. This dataset based on 30 dimensional features (columns) and tuples (rows) on 569. We run three different programs, the first program process PCA along with SVM algorithm, the second program process PCA along with KNN algorithm and finally, the third program process PCA along with K-Means algorithm and all programs same dataset used. Throughout all of these applications are running on the personal computer. The configuration of pc is given below:

- Second Generation Intel (R) Core (TM) i5-2520M CPU @ 2.50 GHz.
- RAM of 4.00 GB.
- system is a 64-bit operating system.
- Windows 10 (Home).
- 500 GB hard disk.

4. Result and discussion

Performance benchmarking the comparison between SVM and KNN algorithms and used technique in supervised learning, and other comparison between K-Means and KNN algorithm and technique used in Un-supervised learning, both comparisons along with PCA implemented. The dataset utilized in this paper breast cancer disease and

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/UWJK5

table 1 are well characterized in this research. This dataset consists of 569 tuples (columns) and 30 dimensional characteristics (columns). The WDBC-Malignant and WDBC-Benign dataset target attributes are grouped into two categories (0,1). We found that the forecast was correct. As a result, highly speedy performance was discovered, which works better and tests regression, classification, and clustering approaches.

4.1. Principal component analysis (PCA) results

All the programs of the breast cancer dataset are shown here. These breast cancer dataset programs are shown in Figures 8, 9 without PCA implementation. The scaled data 30-Dimension dataset is clearly defined in Figure 8, and the mixed data chart is shown in Figure 9 after the PCA approach is applied to minimize dimensions. The PCA technique reduces a 30-dimensional dataset to a 2-dimensional one. Figure 10 depicts the two-dimensional dataset, while Figure 11 depicts two distinct groups. The first PCA colour circle is purple, and the second PCA colour circle is yellow.

Table 1. Breast Cancer Dataset

Mean Radius	Mean Texture	Mean Perimeter	Mean Area	Mean Smoothness	Mean Compactness
17.99	10.38	122.80	1001.0	0.11840	0.27760
20.57	17.77	132.90	1326.0	0.08474	0.07864
19.69	21.25	130.00	1203.0	0.10960	0.15990
11.42	20.38	77.58	386.1	0.14250	0.28390
20.29	14.34	135.10	1297.0	0.10030	0.13280
Mean Concavity	**Mean Concave Points**	**Mean Symmetry**	**Mean Fractal Dimension**	**Radius Error**	**Texture Error**
0.3001	0.14710	0.2419	0.07871	1.0950	0.9053
0.0869	0.07017	0.1812	0.05667	0.5435	0.7339
0.1974	0.12790	0.2069	0.05999	0.7456	0.7869
0.2414	0.10520	0.2597	0.09744	0.4956	1.1560
0.1980	0.10430	0.1809	0.05883	0.7572	0.7813
Perimeter Error	**Area Error**	**Smoothness Error**	**Compactness Error**	**Concavity Error**	**Concave Points Error**
8.589	153.40	0.006399	0.04904	0.05373	0.01587
3.398	74.08	0.005225	0.01308	0.01860	0.01340
4.585	94.03	0.006150	0.04006	0.03832	0.02058
3.445	27.23	0.009110	0.07458	0.05661	0.01867
5.438	94.44	0.011490	0.02461	0.05688	0.01885

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/UWJK5

Symmetry Error	Fractal Dimension Error	Worst Radius	Worst Texture	Worst Perimeter	Worst Area
0.03003	0.006193	25.38	17.33	184.60	2019.0
0.01389	0.003532	24.99	23.41	158.80	1956.0
0.02250	0.004571	23.57	25.53	152.50	1709.0
0.05963	0.009208	14.91	26.50	98.87	567.7
0.01756	0.005115	22.54	16.67	152.20	1575.0
Worst Smoothness	**Worst Compactness**	**Worst Concavity**	**Worst Concave Points**	**Worst Symmetry**	**Worst Fractal Dimension**
0.1622	0.6656	0.7119	0.2654	0.4601	0.11890
0.1238	0.1866	0.2416	0.1860	0.2750	0.08902
0.1444	0.4245	0.4504	0.2430	0.3613	0.08758
0.2098	0.8663	0.6869	0.2575	0.6638	0.17300
0.1374	0.2050	0.4000	0.1625	0.2364	0.07678

```
30-Dimension Data

array([[ 1.09706398, -2.07333501,  1.26993369, ...,  2.29607613,
         2.75062224,  1.93701461],
       [ 1.82982061, -0.35363241,  1.68595471, ...,  1.0870843 ,
        -0.24388967,  0.28118999],
       [ 1.57988811,  0.45618695,  1.56650313, ...,  1.95500035,
         1.152255  ,  0.20139121],
       ...,
       [ 0.70228425,  2.0455738 ,  0.67267578, ...,  0.41406869,
        -1.10454895, -0.31840916],
       [ 1.83834103,  2.33645719,  1.98252415, ...,  2.28998549,
         1.91908301,  2.21963528],
       [-1.80840125,  1.22179204, -1.81438851, ..., -1.74506282,
        -0.04813821, -0.75120669]])
```

Figure 7. 30- Dimension Dataset

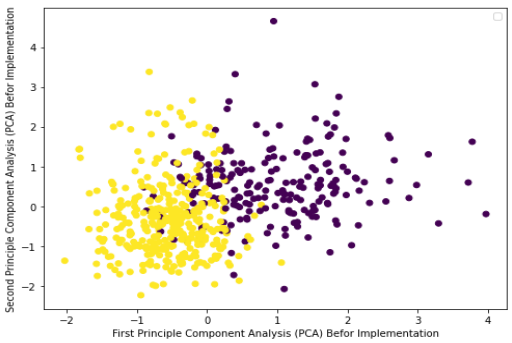

Figure 8. PCA Before Implementation Mixed Data Chart

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/UWJK5

```
2-Dimension Data

array([[ 9.19283683,  1.94858307],
       [ 2.3878018 , -3.76817174],
       [ 5.73389628, -1.0751738 ],
       ...,
       [ 1.25617928, -1.90229671],
       [10.37479406,  1.67201011],
       [-5.4752433 , -0.67063679]])
```

Figure 9. 2-Dimesions Dataset

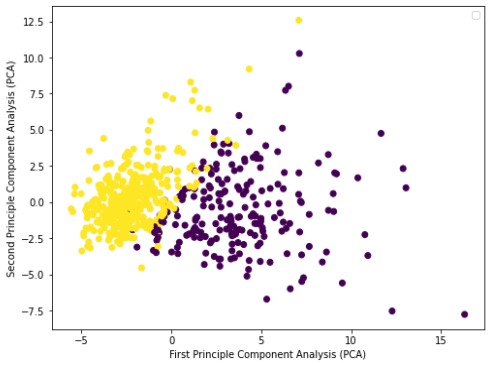

Figure 10. Principal Component Analysis (PCA) Data Chart

4.2. Support vector machine (SVM) and K-nearest neighbors (KNN) algorithms result

The definitive estimate outcome, the support vector and K closest values forecast the ultimate results. Regression, classification, and mathematical methods like kernel linear (radial basis function) and minkowski distance all use these procedures. The comparison of SVM and KNN outcomes is briefly defined in tables 2 and 3, followed by table 4 and figure 12 that clearly defines the comparison graph. This table, as well as the pictures, describe the precision, recall, accuracy, and classifier of both algorithms, as well as figure 13, which depicts the predicted and true labels. Figure 15, 16 depicts precision, recall, accuracy, and classifier in this graph, which is based on two classes (0, 1).

Table 2. SVM Classification Malignant and Benign Outcomes

Results of SVM Class Malignant and Benign				
	Precision	**Recall**	**F1- Score**	**Support**
Class (Malignant 0)	0.88	0.90	0.89	60
Class (Benign 1)	0.94	0.93	0.93	108
Macro Avg.	0.91	0.92	0.91	171
Weighted Avg.	0.92	0.92	0.92	171

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/UWJK5

Table 3. KNN Class Malignant and Benign Result

Results of KNN Class Malignant and Benign				
	Precision	Recall	F1- Score	Support
Class (Malignant 0)	0.00	0.00	0.00	61
Class (Benign 1)	0.64	1.00	0.78	110
Macro Avg.	0.32	0.50	0.39	171
Weighted Avg.	0.41	0.64	0.50	171

Table 4. Results of SVM and KNN Algorithm

	Precision	Recall	Accuracy	Classifier
SVM	0.943396226415094	0.925925925925925 9	0.9181286549707602	0.96
KNN	0.643274853801196	1.0	0.6432748538011696	0.95

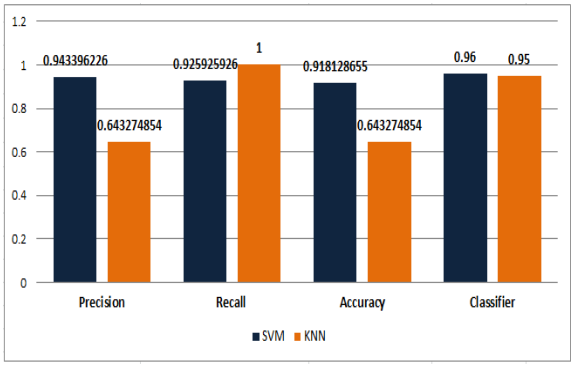

Figure 11. Comparison Graph KNN & SVM Results

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/UWJK5

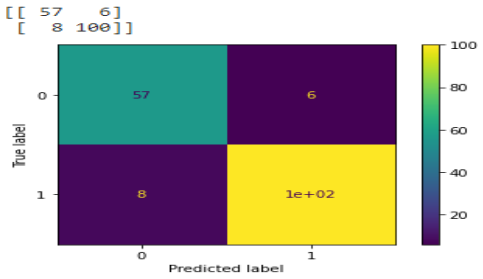

Figure 12. Predicted Label and True Label SVM

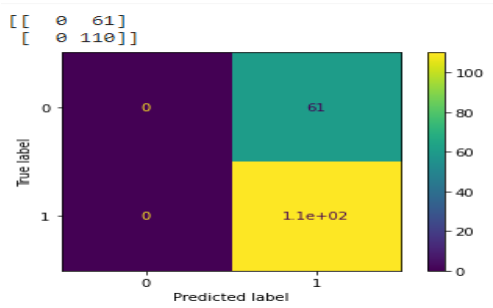

Figure 13. Predicted Label and True Label KNN

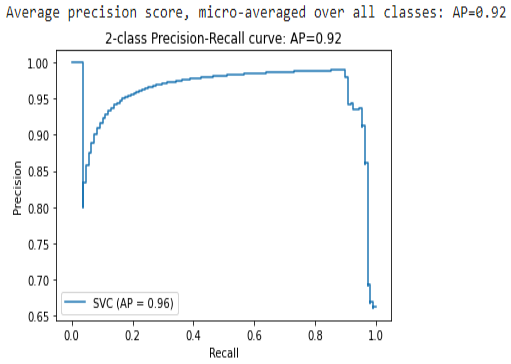

Figure 14. Precision & Recall SVM

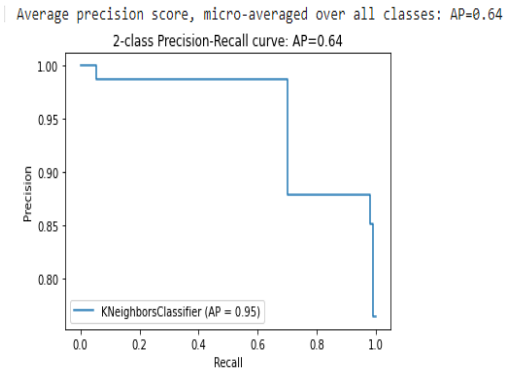

Figure 15. Precision & Recall KNN

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/UWJK5

4.3. K-mean algorithm result

The ultimate outcome the centroid clustering values, the K-mean algorithm, and a three-dimensional dataset in essence, this algorithm utilizes a clustering technique. The breast cancer and clusters dataset are clearly specified as a three-dimensional dataset, with two features based on breast cancer illness values and one feature target attribute number of clusters (see Figure 17). Figure 18 depicts the K-Means Clustering Centroid Value, and figure 19 depicts the three clusters after the unstructured dataset has been turned into structured data. These graphs show three clusters: blue, green, and red, with centroid values on the yellow star (*). Figure 20: Sum of squared error line cha calculated using K-Means.

	1-Dimensional	2-Dimensional	Clusters
0	9.192837	1.948583	1
1	2.387802	-3.768172	1
2	5.733896	-1.075174	1
3	7.122953	10.275589	0
4	3.935302	-1.948072	1
...	...	...	...
564	6.439315	-3.576817	1
565	3.793382	-3.584048	1
566	1.256179	-1.902297	1
567	10.374794	1.672010	1
568	-5.475243	-0.670637	2

569 rows × 3 columns

Figure 16. Breast Cancer & Clusters Dataset

```
km.cluster_centers_
```

```
array([[ 2.18623605,  3.19710261],
       [ 5.05903111, -1.74041615],
       [-2.36453918, -0.25542325]])
```

Figure 17. K-Means Clustering Centroid Value

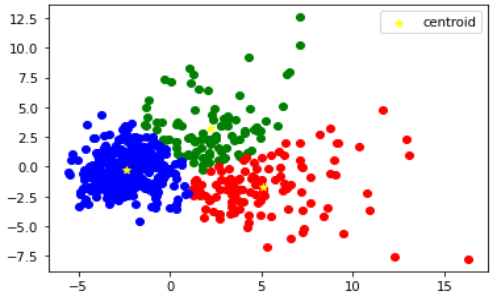

Figure 18. K-Means Three Clusters

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/UWJK5

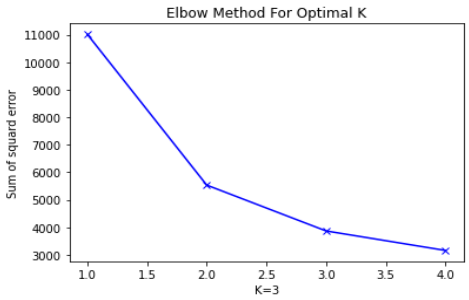

Figure 19. K-Means Sum of Squared Error Line Chart

5. Conclusion

Similar datasets and the data mining technique PCA were used in this study. Due to supervised and unsupervised learning, this data has been labeled and unlabeled, based on two types, WDBC-Malignant and WDBC-Benign (0, 1). First comparison of the algorithm for SVM & KNN and the technique used in supervised learning. Accuracy, recall, precision, classification, and prediction are often different from SVM and KNN outcomes. It was found that the accuracy, recall, precision, classifier, and prediction of KNN outperformed because the accuracy, recall, precision, and prediction of SVM is more effective and faster than that of the algorithm (KNN) and certain other comparisons between the algorithm and technique used in un-supervised learning of K-Means and KNN. KNN accuracy, recall, precision, classifier, could have been observed, While K-means the accurate results of centroid values, clustering, prediction, and sum of square error accurate results, no error is generated. The K-means is more efficient and faster than the KNN algorithm, according to certain results.

References

1) Norsyela Muhammad Noor Mathivanan1, Nor Azura Md.Ghani2, Roziah Mohd Janor3,"Comparative study on dimensionality reduction between principal component analysis and k-means clustering", November 2019.
2) Madhuri Gupta1, Bharat Gupta2,"Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques", International Conference on Computing Methodologies and Communication (ICCMC 2018).
3) Jordi L. Vermeulen, Arne Hillebrand, Roland Geraerts, "Comparative study of k-nearest neighbor techniques in crowd simulation", March 2017.
4) Erwin, M. Fachrurrozi, Ahmad Fiqih, Bahardiansyah Rua Saputra, Rachmad Algani, Anggina Primanita, "Content Based Image Retrieval for Multi-Objects Fruits Recognition using k-Means and k-Nearest Neighbor", 2017 International Conference on Data and Software Engineering (ICoDSE).
5) Alhamza Munther 1, Rozmie Razif 1, Mosleh AbuAlhaj2, Mohammed Anbar3, Shahrul Nizam1, "Preliminary Performance Evaluation of K-means, KNN and EM Unsupervised Machine Learning Methods for Network Flow Classification", April 2016.
6) Iyer Aurobind Venkatkumar, Sanatkumar Jayantibhai Kondhol Shardaben, "Comparative study of Data Mining Clustering algorithms", 2016 IEEE International Conference on Data Science and Engineering (ICDSE).
7) P Thamilselvana, Dr.J.G.R.Sathiaseelanb, "Comparative Study of Data Mining Algorithms for Image Classification", June 2015.
8) Shaikh, Z. A., & Lashari, I. A. (2017). Blockchain technology: The new internet. International Journal of Management Sciences and Business Research, 6(4), 167-177.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/UWJK5

9) Prachi Surlakar, Sufola Araujo, Dr.K.Meenakshi Sundaram, "Comparative Analysis of K-Means and K-Nearest Neighbor Image Segmentation Techniques",2016 IEEE 6th International Conference on Advanced Computing.
10) Shestak, V., Gura, D., Khudyakova, N., Shaikh, Z. A., & Bokov, Y. (2020). Chatbot design issues: building intelligence with the Cartesian paradigm. Evolutionary Intelligence, 1-9.
11) Pooja Mittal1, Nasib Singh Gill2,"Comparative Analysis Of Classification Techniques On Medical Data Sets",Jun-2014, International Journal of Research in Engineering and Technology (IJRET).
12) Shaikh, Z. A. (2018). Keyword Detection Techniques. Engineering, Technology & Applied Science Research, 8(1), 2590-2594.
13) Fei Wang 1,2, Zhao Zhen 1, Bo Wang 3 and Zengqiang Mi 1,"Comparative Study on KNN and SVM BasedWeather Classification Models for Day Ahead Short Term Solar PV Power Forecasting", (Applied Science, 2018).
14) Phan Thanh Noi 1,2, and Martin Kappas 1,"Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery", December 2017
15) Ratna Astuti Nugrahaeni1, Kusprasapta Mutijarsa2, "Comparative Analysis of Machine Learning KNN, SVM, and Random Forests Algorithm for Facial Expression Classification", 2016 International Seminar on Application for Technology of Information and Communication.
16) P Thamilselvana, Dr. J. G. R. Sathiaseelan.b, "Comparative Study of Data Mining Algorithms for Image Classification", June 2015.
17) Tania Pereira1, Joana S. Paiva2, Carlos Correia1, João Cardoso1, "An automatic method for arterial pulse waveform recognition using KNN and SVM classifiers", International Federation for Medical and Biological Engineering 2015.
18) Rajkumar Palaniappan1, Kenneth Sundaraj1 and Sebastian Sundaraj2,"Comparative study of the svm and k-nn machine learning algorithms for the diagnosis of respiratory pathologies using pulmonary acoustic signals", Palaniappan et al. BMC Bioinformatics 2014.
19) Higher Education in Pakistan: An ICT Integration Viewpoint
20) Fereshteh Modaresi, Shahab Araghinejad, "Comparative Assessment of Support Vector Machines, Probabilistic Neural Networks, and K-Nearest Neighbor Algorithms for Water Quality Classification", June 2014.
21) https://static.javatpoint.com/tutorial/machine-learning/images/classification-algorithm-in-machine-learning.png
22) https://media.geeksforgeeks.org/wp-content/uploads/merge3cluster.jpg
23) https://static.javatpoint.com/tutorial/machine-learning/images/support-vector-machine-algorithm.png
24) https://static.javatpoint.com/tutorial/machine-learning/images/k-nearest-neighbor-algorithm-for-machine-learning2.png
25) https://static.javatpoint.com/tutorial/machine-learning/images/k-means-clustering-algorithm-in-machine-learning.png
26) Khan, Abdullah Ayub, and Syed Asif Ali. "Network forensics investigation: behaviour analysis of distinct operating systems to detect and identify the host in IPv6 network." International Journal of Electronic Security and Digital Forensics 13, no. 6 (2021): 600-611.
27) Khan, Abdullah Ayub, Zaffar Ahmed Shaikh, Asif Ali Laghari, Sami Bourouis, Asif Ali Wagan, and Ghulam Ali Alias Atif Ali. "Blockchain-Aware Distributed Dynamic Monitoring: A Smart Contract for Fog-Based Drone Management in Land Surface Changes." Atmosphere 12, no. 11 (2021): 1525.
28) Khan, Abdullah Ayub, Asif Ali Laghari, De-Sheng Liu, Aftab Ahmed Shaikh, Dan-An Ma, Chao-Yang Wang, and Asif Ali Wagan. "EPS-Ledger: Blockchain Hyperledger Sawtooth-Enabled Distributed Power Systems Chain of Operation and Control Node Privacy and Security." Electronics 10, no. 19 (2021): 2395.
29) Shaikh, Zaffar Ahmed, Abdullah Ayub Khan, Laura Baitenova, Gulmira Zambinova, Natalia Yegina, Natalia Ivolgina, Asif Ali Laghari, and Sergey Evgenievich Barykin. "Blockchain Hyperledger with Non-Linear Machine Learning: A Novel and Secure Educational Accreditation Registration and Distributed Ledger Preservation Architecture." Applied Sciences 12, no. 5 (2022): 2534.
30) [26] Khan, Abdullah Ayub, Asif Ali Laghari, Aftab Ahmed Shaikh, Mazhar Ali Dootio, Vania V. Estrela, and Ricardo Tadeu Lopes. "A Blockchain Security Module for Brain-Computer Interface (BCI) with Multimedia Life Cycle Framework (MLCF)." Neuroscience Informatics (2021): 100030.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:04:2022
DOI 10.17605/OSF.IO/UWJK5

31) Khan, Abdullah Ayub, Aftab Ahmed Shaikh, Omar Cheikhrouhou, Asif Ali Laghari, Mamoon Rashid, Muhammad Shafiq, and Habib Hamam. "IMG-forensics: Multimedia-enabled information hiding investigation using convolutional neural network." IET Image Processing (2021).
32) Ayub Khan, Abdullah, Asif Ali Laghari, Aftab Ahmed Shaikh, Sami Bourouis, Amir Madany Mamlouk, and Hammam Alshazly. "Educational Blockchain: A Secure Degree Attestation and Verification Traceability Architecture for Higher Education Commission." Applied Sciences 11, no. 22 (2021): 10917.
33) Khan, Abdullah Ayub, Asif Ali Laghari, and Shafique Ahmed Awan. "Machine learning in computer vision: A review." EAI Transactions on Scalable Information Systems (2021): e4.
34) Khan, Abdullah Ayub, Aftab Ahmed Shaikh, Zaffar Ahmed Shaikh, Asif Ali Laghari, and Shahid Karim. "IPM-Model: AI and metaheuristic-enabled face recognition using image partial matching for multimedia forensics investigation with genetic algorithm." Multimedia Tools and Applications (2022): 1-17.
35) Khan, Abdullah Ayub, Zaffar Ahmed Shaikh, Laura Baitenova, Lyailya Mutaliyeva, Nikita Moiseev, Alexey Mikhaylov, Asif Ali Laghari, Sahar Ahmed Idris, and Hammam Alshazly. "QoS-Ledger: Smart Contracts and Metaheuristic for Secure Quality-of-Service and Cost-Efficient Scheduling of Medical-Data Processing." Electronics 10, no. 24 (2021): 3083.
36) Khan, Abdullah Ayub, Asif Ali Laghari, Shafique Awan, and Awais Khan Jumani. "Fourth Industrial Revolution Application: Network Forensics Cloud Security Issues." Security Issues and Privacy Concerns in Industry 4.0 Applications (2021): 15-33.
37) Laghari, Rashid Ali, Jianguang Li, Asif Ali Laghari, and Shu-qi Wang. "A review on application of soft computing techniques in machining of particle reinforcement metal matrix composites." Archives of Computational Methods in Engineering 27, no. 5 (2020): 1363-1377.
38) Laghari, Asif Ali, Kamran Ali Memon, Muhammad Bux Soomro, Rashid Ali Laghari, and Vishal Kumar. "Quality of experience (QoE) assessment of games on workstations and mobile." Entertainment Computing 34 (2020): 100362.
39) Laghari, Asif Ali, Awais Khan Jumani, and Rashid Ali Laghari. "Review and State of Art of Fog Computing." Archives of Computational Methods in Engineering (2021): 1-13.
40) Huang, I., Yu-Hsuan Lu, Muhammad Shafiq, Asif Ali Laghari, and Rahul Yadav. "A Generative Adversarial Network Model Based on Intelligent Data Analytics for Music Emotion Recognition under IoT." Mobile Information Systems 2021 (2021).
41) Laghari, Asif Ali, and Mureed Ali Laghari. "Quality of experience assessment of calling services in social network." ICT Express 7, no. 2 (2021): 158-161.
42) Junejo, Mehran, Asif Ali Laghari, Awais Khan Jumani, Shahid Karim, and Mansoor Ahmed Khuhro. "Quality of Experience Assessment of Banking Service." International Journal of Information Engineering & Electronic Business 12, no. 6 (2020).
43) Yin, Shoulin, Hang Li, Asif Ali Laghari, Shahid Karim, and Awais Khan Jumani. "A Bagging Strategy-Based Kernel Extreme Learning Machine for Complex Network Intrusion Detection." EAI Endorsed Transactions on Scalable Information Systems 8, no. 33 (2021): e8.
44) Shkodinsky S.V., Dudin M.N., Usmanov D.I. Analysis and evaluation of cyber threats to the national financial system of Russia in the digital economy // Financial Journal. 2021. Vol. 13. No. 3. pp. 38-53.
45) Gura, D., Mikhaylov, A., Glushkov, S., Zaikov, M., & Shaikh, Z. A. (2020). Model for estimating power dissipation along the interconnect length in single on-chip topology. Evolutionary Intelligence.
46) Shaikh, Z. A., Umrani, A. I., Jumani, A. K., & Laghari, A. A. (2019). Technology enhanced learning: a digital timeline learning system for higher educational institutes. International Journal of Computer Science and Network Security, 19(10), 1-5.