International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2019): 7.583

An Approach for Detecting Noise in Cotton Fiber Properties Data Using Nearest Neighbor Algorithm

Mona Shalaby

Researcher, Cotton Grading Research Section, Cotton Research Institute, Egypt

Abstract: ***The purpose of the current study is to decrease noise of cotton fiber properties data using one of methods for machine learning such as K Nearest Neighbor (KNN). The present investigation was carried out at Egyptian & International Cotton Classification Center (EICCC). The first data was for Giza 87, Giza 88, Giza 86, Giza 90 and Giza 95 each one separately and the second data was for combination of all previous cotton varieties. A wide range of lint cotton grades used in this work. The studied traits were basic fiber properties; length, strength and micronaire value. The highest classification accuracy were 117.65 % for G 87 and 149.25 % for combined data. The integrated statistics among fiber length, strength and micronaire value concluded that spinning consistency value (SCI) which is the most intrinsic technological value were in acceptable range for Giza 87, Giza 88, Giza 86, Giza 90, Giza 95 separately and combined data. For instance, SCI values were 163.74 and 174.85 for G 87 of data treated without KNN and with KNN, respectively. Therefore, any study of cotton fiber properties plays a crucial role in determining spinning performance.***

Keywords: Egyptian cotton varieties, Length, Strength, Micronaire value, Machine learning, KNN, SCI

1. Introduction

Cotton is the world's most important fiber crop and the second most important oil seed crop. The primary product of cotton plant is the lint that covers the seeds (Freeland et al. 2006). Cotton fibers grow from the epidermis of the seed coat cells and begin to elongate from the day of anthesis flowering (Chaudhry and Guitchounts 2003).

Cotton classing has progressed from subjective human classers to instruments. For grade, classification is based on appearance (sight and touch for human classer) with cotton fiber quality properties using manual and mechanical instruments such as high, medium and low volume instrument. Egyptian cotton may be classed according to official cotton standards of Egyptian cotton, depending on comparison with certified cotton standard types. The quality of cotton fiber depends on a large set of characteristics which includes length, strength, fineness and maturity as seen with details in Bradow et al. (1997) and Ghosh et al. (2015).

Long et al. (2013) exhibited using several techniques to determine cotton fiber attributes in many production regions, and these properties explain much but not all of the variation in yarn attributes, and significant work has been conducted into understanding the relative contribution of fiber properties. A relationship exists between cotton fiber properties, both individually and collectively in terms of cotton fiber properties have been confirmed by means of statistical methods by many investigators according to Fiori and Brown (1951).

Fiber length is one of the most important technological properties of cotton fibers in both marketing and processing. Technological changes in the textile industry show that priorities associated with fiber properties have also changed, so the cotton breeders have to concentrate on the improvement in fiber traits to meet the demands of textile industry. Currently, most cotton programs are focused on breeding for longer fibers alone because the current premium and discount schedule reward this type of cotton. The average length of all fibers in a sample or the average length of a given percentage is related to other cotton fiber characteristics such as strength and micronaire (Braden 2005).

Smith (1947) and Moujalvo (2005), Foulk and McAlister (2002) and Abbott et al. (2010). Explaining micronaie with spinning which is the premier of measurement based on air resistance, originally intended to efficiently determine cotton fiber linear density. High micronaire cotton is considered coarse (large perimeter) by spinners and results in fewer fibers in the yarn cross section, which translates into weaker yarns. Alternatively, while lower micronaire is associated with more desirable, finer fibers, it is also interpreted as being immature and prone to dye uptake problems, breakage and nep (fiber knot) formation.

E C G (2015) and C A T G O (2016) exhibited the aim of blending cottons with different quality characteristics may have an effect on fiber characteristics of the blend and resulting yarn quality in terms of Egyptian cotton industry suffered major reduction in both area and total exported quantities and that was accompanied by an increase in imported different origin cottons. In addition to several reasons contributed to the difficulties facing the cotton industry in Egypt including increase in production expenses, deterioration of cultivated varieties in yield and quality and lower demand for Egyptian cotton in international markets due to its higher prices compared to other cottons of nearly similar quality properties.

Various instruments have been developed for commercial use that attempt to quickly and easily measure cotton fiber properties compared to manual instruments. Zhao et al. (2018).

Relationships among cotton fiber properties have also studied by numerous authors by several mathematical methods. One of the most important methods; neighboring which spatial pattern attempts to determine the underlying processes which lead to such patterns in points, lines, and areas. The search for evidence that some observed pattern

has not arisen from some chance process leads into a range of techniques. Spatial heterogeneity usually causes negative and positive correlation among points. When there is competition among points, neighbors can have also negative and positive effect on the response in adjacent points. The choice will depend on the size, shape and spacing of the points and on the biological and physical mechanisms influencing the correlation between points in accordance with Berman (1986).

Traditional statistics has been around for more than a century. Actually, the term was coined in Germany in 1749. If its connection with probability theory is taken into account, then its history may even go as far back as the 16th century. Nevertheless, the point is that, unlike artificial intelligence (AI) and machine learning (ML), traditional statistics is not a new technology.

Machine learning (ML) is ubiquitous in the industry these days. Organizations around the world are scrambling to integrate machine learning into their functions and new opportunities for aspiring data scientists are growing multifold. But we have noticed a huge gap between what the industry needs and what´s on offer right now. Quite a large number of scientists are not clear about almost all machine learning applications. There are different types of machine learning as follows; Reinforcement learning, Unsupervised learning and supervised learning which its ever-growing list of applications, the latest machine learning developments, the top experts among all advanced science aspects (Zhou et al. (2017), Shakhovska et al. (2018) and Kumar and Bindu (2019)).

Ghosh and Chatterjee (2010) used one of the machine learning sub-type; support vector machine (SVM) regression approach to forecast the properties of cotton yarns which were produced from the fiber properties measured by HVI and AFIS. The investigation indicates that the yarn properties can be predicted with a very high degree of accuracy using SVM models and the prediction performance of SVM models are better than other studied models.

This is great place to begin exploring machine learning with K-Nearest Neighbor (KNN). KNN is one of the most basic yet essential classification algorithms in Machine Learning (ML). It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and instruction detection.

In simple terms, machine learning is the science of developing and making use of specialized statistical learning algorithms that produce a predictive model based on information gathered from input data (Moura and Cordeiro (2020). K-nearest neighbors is an algorithm that helps to assess the properties of a new variable (included both of studied properties) with the help of the properties of existing variables. It is applicable in classification as well as predictive methods. K-NN is a simple non-parametric test. It is a useful technique can be to assign weights to the contributions of the neighbors, so the nearer neighbors contribute more to the average than the more distant ones. It simply takes the voting of majority of variables and accordingly treats new variables.

Rodriguez et al. (2008) illustrated nearest neighbor rule is well known classification method largely studied in different communities; as much as its simplicity and its performance. The definition of the distance function is central for obtaining a good accuracy on a given data set and different distance functions based on the correlation of the studied data set to increase the performance.

The concept of similarity (distance) plays a central role in some pattern recognition methods such as K-Nearest Neighbor (K-NN). Distance function can be categorized into those that can handle ordinal (continuous or discrete), nominal and heterogeneous input data consisting of both ordinal and nominal data. Wilson and Martinez (1997).

In recent years, the cotton breeders in Egypt have made a great effort to improve fiber quality properties from seed cotton yield. Therefore, a nearest neighbor analysis was performed with the aim of assessing the relationship (distance) among the cotton fiber properties with each other characteristics contained within data sets for different cotton varieties.

Herein, using any method of technological value gives the decision maker of good use or not for the influencing of cotton fiber properties on yarn performance.

The present study was carried out to investigate data combining of basic fiber cotton properties for different Egyptian cotton varieties using nearest neighbor analysis and their ability to spinning performance.

2. Materials and Methods

Materials used in this work were Egyptian cotton varieties; Giza 87, Giza 88, Giza 86, Giza 90, Giza 95. The cotton varieties were chosen as their genetic diversity and had a wide variation in their fiber properties. For each variety, seven lint cotton grade namely; G/FG, G, FGF/G, FGF, GF/FGF, GF and FF/GF.

Cotton fiber properties studied include, fiber length (mm), fiber strength (g/tex) and fiber fineness and maturity; maicronaire (unit).

Data of cotton fiber were conducted under standard testing conditions of 20 ± 2 ^{o} C temperature and 65 ± 2 % Rh relative humidity (ASTM 2016) at the premises of Egyptian & International Cotton Classification Center (EICCC), Cotton Research Institute (CRI), Agricultural Research Center (ARC) using Cotton Classifying System (CCS) designed to measure all fiber proprieties either short staple (cotton) or long staple (man made fibers). All used samples were collected from 2019 and 2020 cotton growing seasons.

Descriptive statistics analyses and simple correlation coefficients were calculated and elucidated according to Steel and Torrie (1980).

K-nearest neighbor (KNN) algorithm is type of supervised machine learning technique which can be used for both classification as well as regression predictive models.

Nearest neighbor analysis is a method for classifying cases based on their similarity to other cases. In machine learning technique, it was developed as a way to recognize patterns of data without requiring an exact match to any stored patterns, or cases. Similar cases are near to each other and dissimilar cases are distant from each other. Thus, the distance between two cases is a measure of their dissimilarity.

The value of the number of nearest neighbors is called "K". Where the nearest neighbor is calculated based on the K-value in determining how many closest neighbors should be considered to decide the class of the sample data point. If K is too small useful classification information may not be enough, while large K-values can easily cause outliers including in the closest neighbors of the true class.

KNN is a model that classifies data points based on the points that are most similar to each other. It uses test data to make an "educated guess" on what an unclassified point should be classified also. Cases that are near to each other are said to be "neighbors". When a new case (hold out) is presented, its distance from each of the cases in the model is computed. The classifications of the most similar cases the nearest neighbors are tallied and the new case is placed into the category that contains the greatest number of nearest neighbors.

KNN algorithm passes through steps with details and discussion according to Docey (1960) put the basics of nearest neighbor; where K-nearest neighbor (KNN) algorithm which uses feature similarity to predict the values of new data points which further means that the new data point will be assigned a value based on how it closely matches the points in the training set.

Nearest neighbor analysis measure the linear distance between two or more specified properties (UHM, strength and micronaire with each other; X, Y and Z coordinates. Such points are randomly distributed in terms of the poisson distribution. These relations are subsequently used to detect the presence of non randomness in given patterns. With the poission distribution employed as a standard of comparison, a chi-square (X^2) comparison has also been employed; Thompson (1956). Although it is interested in divergence from randomness along the R-scale. The formula used is as follows:

$Rn = 2d \sqrt{n/A}$

Rn nearest neighbor value describing the point pattern.

d observed average neighbor distance.
n total number of phenomena under study.
A area of the phenomena under study.

The nearest neighbor formula will produce a result between 0 and 2.15, where the following distribution patterns form a continuum: cluster (0), random (1) and regular (2.15). Thus an important point to make is that the nearest neighbor statistic cannot be 0 but must be less than 1.

Determining technological value of cotton as Spinning Consistency Index (SCI) according to Majumdar et al. (2005).

SPSS (2012) software was used for all statistical analyses.

3. Results

Using wide range of cotton varieties is a strategy for merchandising that relies on an impressive range of fiber properties to draw final expected product into the market. Wide cotton varieties can compete with big range of fiber properties in terms of result in Table (1). Studying descriptive statistics such as minimum, maximum, mean, standard deviation and coefficient of variation characteristics; measure obtained from samples of Giza 87, Giza 88, Giza 86, Giza 90, Giza 95 and combined data of the previous varieties.

Steel and Torrie (1980) proved that the correlation coefficient is a simple descriptive statistics that measures the strength of the linear relationship between two interval scale variables; herein cotton fiber attributes such as upper half mean length (UHML), strength, and micronaire (Mike) as might be visualized in a simple relation as seen in (Wright 1921). Correlation analysis indicated that numbers of attributes were highly positively associated with each other for G 87, G 88, G 86, G 90 and G 95. Meanwhile the same trends of data were for combined data except for correlation between UHM and mike as seen in (Table 2 and Table 3).

These results were in agreement with Chaudhry and Guitchounts (2003), Karademir et al. (2010), Wang et al. (2012), and Lokhande and Reddy (2015). Almost the same trend of results was exhibited by Arshad et al. (1993) and Clouvel et al. (1998). However simple correlation matrix has the ability to introduce the strength of relation between two properties but it is premier to detect relation among more than two properties at the same time with each other with the same priority to all.

Applying correlation coefficients give the power ability to improve the performance of nearest neighbor classifier compared to other distance functions that were figured out by Rodriguez et al. 2008). It is acceptable to have the opportunity to apply nearest neighbor analysis (NNA) after using traditional correlation analysis that may lead to efficient results due to Bartlett (1978).

Table 1: Descriptive of cotton varieties

Varieties	Traits	Min.	Max.	Mean	S.D.	C.V.
G 87	UHM	30.18	36.15	33.36	1.75	5.25
	Strength	30.60	50.46	39.97	6.43	16.09
	Mike	2.34	3.98	2.95	0.537	18.20
G 88	UHM	31.49	35.34	33.40	1.02	3.05
	Strength	36.20	48.40	41.14	3.79	9.21
	Mike	2.57	4.32	3.34	0.594	17.78
G 86	UHM	28.76	32.74	31.10	1.21	3.89
	Strength	31.80	48.90	40.29	4.97	12.34
	Mike	3.29	5.06	3.93	0.585	14.89
G 90	UHM	25.98	30.78	28.28	1.54	5.45
	Strength	27.90	43.00	34.38	3.72	10.82
	Mike	2.71	4.72	3.62	0.676	18.67
G 95	UHM	27.92	31.38	29.89	1.02	3.41
	Strength	25.40	42.70	33.81	5.74	16.98
	Mike	2.98	4.73	3.65	0.585	16.03
Combined	UHM	25.98	36.15	31.21	2.39	6.30
	Strength	25.40	50.46	37.92	5.87	15.48
	Mike	2.34	5.06	3.49	0.673	19.28

Min., Max., S.D. and C.V. refer to minimum, maximum, standard deviation and coefficient of variation, respectively

Table 2: Correlation matrix for the investigated attributes

Traits	G 87			G 88			G 86		
	UHM	Strength	Mike	UHM	Strength	Mike	UHM	strength	Mike
UHM	1	0.892**	0.784**	1	0.636**	0.795**	1	0.748**	0.761**
Strength	0.892**	1	0.845**	0.636**	1	0.750**	0.784**	1	0.778**
Mike	0.784**	0.845**	1	0.795**	0.750**	1	0.761**	0.778**	1

** indicates that the correlation coefficient is highly significant at 0.01 probability level

Table 3: Correlation matrix for the investigated attributes

Traits	G 90			G 95			combined		
	UHM	Strength	Mike	UHM	Strength	Mike	UHM	Strength	Mike
UHM	1	0.797**	0.868**	1	0.858**	0.683**	1	0.765**	0.099^{ns}
Strength	0.797**	1	0.731**	0.858**	1	0.842**	0.765**	1	0.476**
Mike	0.868**	0.731**	1	0.683**	0.842**	1	0.099	0.476**	1

** and ns indicate that the correlation coefficient is highly significant at 0.01 probability level and non- significant, respectively

There are certain three properties; length, strength and micronaire. Each property has its own unique attribute associated with yarn properties. Asking for new property which is a combined of length, strength and micronaire and it has the accurate attribute to yarn properties as factorials are the simulation of what happen in real with final product not as done by single one then it would have to use KNN algorithm to determine that.

In terms of K-nearest neighbor; it is an algorithm that helps to assess the properties of a new variable with the help of properties of existing variables (upper half mean (UHM), strength and micronaire (mike). As a result of the selected separated three variables, fourth variable introduced both of them with each other at the same time.

Thus K-nearest neighbor (K-NN) helped in classifying the applicants in two groups training (included points) and holdout (excluded points) based on UHM, strength and micronaire tests. It helped the fiber properties interrelations to easily collect data containing basic properties information and evaluate it accordingly.

As long as, the boundary becomes smoother with increasing value of K. with K increasing to infinity it finally becomes all of points depending on the total majority. KNN depends on feature similarity. Therefore, choosing the right value of K is a process called parameter tuning and is important for better accuracy and selecting the precised K according to slightly stable degree of increasing line distance among points.

As discussed below G 87 was the premier variety to all other cotton varieties; G 88, G 86, G 90 and G 95 as extra long, strong and finest variety. Then to label this variable as existing ones, K-NN can be applied as seen in Table (4) for Giza 87. With focal record points with nearest neighbor distances (K). Considering K= 5 where k-value depends on data trend or data discussion; only it is obvious to be less than used points. Using K= 5, According to nearest neighbor analysis (NNA); there are 7 holdout of focal record in the dataset (2, 3, 7, 10, 11, 16, 20) and 13 training of focal record in K-NN model as seen in Table (4) from (1 to 19 focal record).

The five nearest neighbors of new variable were exhibited. Then according to the chosen dataset K-nearest neighbor test can be used to find the five nearest neighbors distances from shortest followed by longest distances. The K-nearest neighbor is used which is an effective method to be used in classification. The problem is that if the K-value is too small it can cause insufficient information and if a large K-value easily results in outliers according to Saputra et al. (2019).

As discussed above, the K-NN uses the nearest value to predict the target variable. Figure (1) showed the nearest neighbor analysis (NNA) of upper half mean (UHM), strength and micronaire (Mike) with each other according to results in Table (1). The distance between any given point and its nearest neighbor would be small. The smaller the K value the more clustered the spatial pattern.

Table 4: Nearest neighbors and distance in Giza 87

Focal record	Nearest neighbors					Distances				
	1	2	3	4	5	1	2	3	4	5
1	8	4	6	5	10	0.611	0.801	0.969	1.174	1.603
4	1	8	6	5	10	0.801	0.830	1.103	1.161	1.585
5	6	10	8	9	12	0.303	0.450	0.584	0.827	0.894
6	5	8	10	9	12	0.303	0.453	0.656	0.907	0.930
8	6	5	1	4	10	0.453	0.458	0.611	0.830	1.018
9	12	10	5	6	15	0.420	0.686	0.827	0.907	0.931
12	9	10	15	5	6	0.420	0.630	0.698	0.894	0.930
13	17	14	18	19	15	0.163	0.226	0.267	0.606	0.675
14	13	17	18	15	19	0.226	0.366	0.453	0.586	0.615
15	10	14	12	12	17	0.506	0.586	0.675	0.698	0.829
17	13	18	14	19	15	0.163	0.170	0.366	0.494	0.829
18	17	13	19	14	15	0.170	0.267	0.397	0.453	0.860
19	18	17	13	14	15	0.397	0.494	0.606	0.715	1.18

Table (1). Values of neighbor distances for all focal record points ranged from 0.163 (first nearest neighbor for 13 focal record) to 1.603 (fifth nearest neighbor for first focal record). Where almost of focal record points have the closest distance from 0 to 1; that refer to the closest neighboring nearest distances.

Everything must have perspective, a point of view, to be communicated. To communicate the three spatial dimensions, using X, Y and Z coordinates. Each axis is perpendicular to all other axes. These denote height, width and depth. In referring using the same X, Y and Z denotations, but giving them different values or different meanings. In pattern recognition, if properties have different dimensions (such as length vs. strength vs. micronaire), they cannot be expressed in terms of similar units and cannot be compared in quantity (also called incommensurable). They will have the same dimensions on its left and right sides, a property known as dimensional homogeneity. Checking for dimensional homogeneity is a common application of dimensional analysis, serving as a plausibility check on derived equations and computations. It also serves as a guide and constraint in deriving equations that may describe a physical system in the absence of a more rigorous derivation.

There are three coordinates; the X-axis (mike), Y-axis (strength) and Z-axis (UHM). These are the variables for prediction in terms of figure (1).

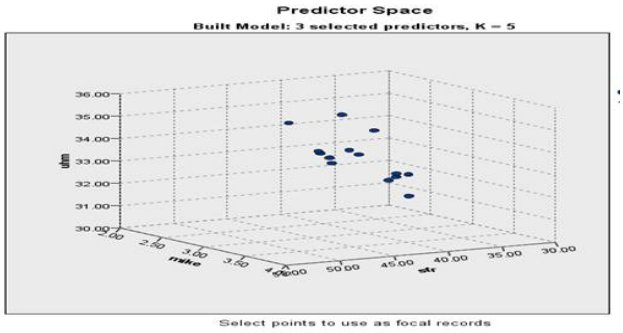

Figure 1: Nearest neighbor of strength, upper half mean and mike in Giza 87

Selecting the second focal record or K-point (4); the value of the point shown in the figure below can be predicted as seen in Figure (2) with the five neighbor distances. The results shows five leading to the five nearest values from the focal record (4); 0.801, 0.830, 1.103, 1.161 and 1.585, respectively. The difference between first and second nearest neighbor distance was 0.029, the differences between second and third distances was 0.273, the difference between (third and fourth) and (fourth and fifth) were 0.058 and 0.424, respectively.

Similarity the peer chart in Figure (3) shows which value is used from which variable (length, strength and micronaire) predict the new variable based on the nearest value. In the peer chart; the nearest values for predicting the new variable whereas dots values are idle.

The numbering within the chart represents the respondents. Where focal record or K-point (4) is the data for the second respondent. Which the algorithm uses to predict values or groups in the response variable. Peer chart also shows the data which is to be used for training the model and left for validation. So far there is no holdout data in this dataset and all the data are used for training the KNN model.

The accuracy of this K-point (4) was almost close to 96.51% and the decrease or increase accuracy In terms of previous;

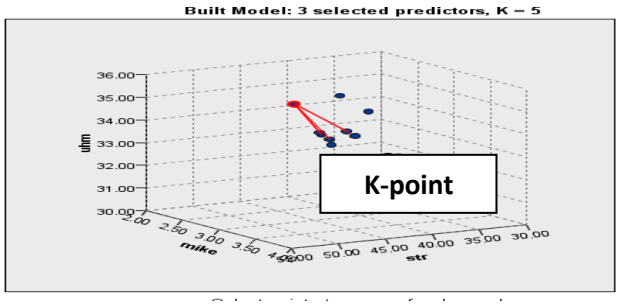

Figure 2: five nearest neighbors for selecting point (4) focal record in G 87

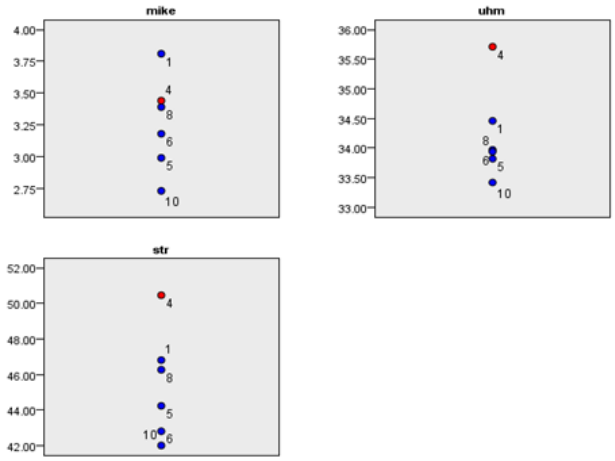

Figure 3: Focal record and nearest neighbor for UHM, strength and mike in G 87

Depend on other following nearest neighbor and select last one depend on the slightly amount of decrease and that depend on the studied data points and the aim of properties of study. The same trend of result was with Wang et al. (2007).

Results of KNN; all focal records of G 87 ranged from 0.163 to 1.174 and that deduced that all distances of cotton fiber properties tend from cluster to random.

Cotton fiber blending is one of the important practices in industry. Spinners purchase cotton bales that have different properties and no two bales can be found that will have the same length, strength, micronaire and all other fiber properties. Therefore, it is necessary to detect the basic cotton fiber properties for data mix of G 87, G 88, G 86, G 90 and G95 to obtain average fiber properties.

Applying the same nearest neighbor analysis for combined data such as done in Giza 87.

The chosen dataset contain various test score of 100 points for each studied cotton fiber properties; length, strength and micronaire. The K-nearest neighbor analysis test uses the nearest value to predict the target variable. Where K= 5 such as in G 87 nearest neighbor Table (4). According to nearest neighbor analysis (NNA); there are 31 holdout of focal record in the dataset and 69 training of focal record in K-NN model as seen in Table (5) from 1 to 99 focal records.

Selecting focal record or K-point (93); the results shows five nearest values; the first (0.229), second (0.434), third (0.448), fourth (0.474) and fifth (0.491).

As long the distance increase from first to fifth nearest neighbor, the related decreased. There is a small degree of increasing in distance according to the applied points (100).

Meanwhile distance increasing in Giza 87 from first to fifth nearest neighbors to large degree in terms of the used data points (20). As long dataset points increase, the distance of nearest neighbors decrease.

According to Table (5); the 5-NN for focal record ranged from 0.055 (81) to 0.642 (80) and that due to the huge focal record compared to focal record in Table (4).

Figure (4) showed the X-axis (mike), Y-axis (UHM) and Z-axis (strength) are the variables for prediction in terms of focal record (93) with the five nearest neighbor distances.

Figure (4) showed the peer chart for UHM, strength and micronaire to predict the new variable based on the nearest value. Peer chart also shows the data which is to be used for the target model. According to the peer chart; there is no holdout data in this dataset and all the data are used for training the KNN model. The accuracy for focal record or K-point (93) ranged from 52.76% to 96.53% from first to fifth neighbors distances, respectively. These results were in accordance with Ma et al. (2014).

All focal records of combined data were very low, narrow and did not pass unit comparing with of G 87. Then it ranged from 0.039 to 0.642 and that illustrated that almost all data points tend to be closer in neighboring than other. As long as focal records increase, neighboring distances decrease where the increase in distances from first to fifth distances in small degree of increasing compared to small numbers of focal record presented in G 87.

Therefore, attributes means how closely are the properties of the new one is related to any of the already known categories. If K is 5 then it will check 5 closest neighbors in order to determine the category. Different properties have different scaling units, like length in mm, strength in m/tex and micronaire in unit. Then how to use them in Euclidean formula. After normalization both will be represented by the value between 0 and 1. KNN is supervised learning algorithm that works on classification problems. Then the aim to strengthen and understanding of studied fiber properties possibilities form and space through developing the final outcome as yarn properties and select elite fiber properties from studied fiber possibilities properties regardless of positive (+) or negative (-) elements where all properties are similar in that they can both be described as visual components with more or less defined boundaries. The data collected is therefore almost always multivariate. It is very hard to analyze process data due to the noises. Then uses three dimension methods is strong and has a very good handling ability.

Table 5: Nearest neighbors and distance in combined data

Focal record	Nearest neighbors					Distances				
	1	2	3	4	5	1	2	3	4	5
1	45	7	2	44	8	0.137	0.274	0.314	0.330	0.331
2	1	45	7	3	4	0.314	0.334	0.356	0.371	0.405
3	4	45	2	1	44	0.298	0.309	0.371	0.418	0.559
4	3	45	2	1	8	0.298	0.375	0.405	0.471	0.486
5	6	11	10	50	8	0.184	0.236	0.274	0.296	0.339
7	44	43	1	2	45	0.196	0.225	0.274	0.358	0.390
8	6	1	5	45	47	0.317	0.331	0.339	0.387	0.412
9	12	51	10	11	56	0.268	0.354	0.379	0.394	0.406
11	10	5	51	50	54	0.083	0.236	0.276	0.295	0.303
12	56	9	57	51	50	0.248	0.268	0.331	0.346	0.424
13	17	14	18	20	53	0.120	0.152	0.157	0.413	0.418
14	13	17	53	18	55	0.152	0.263	0.288	0.291	0.322
17	18	13	14	20	53	0.093	0.120	0.263	0.345	0.528
20	18	17	99	13	92	0.335	0.345	0.371	0.413	0.440
21	24	25	23	84	63	0.212	0.486	0.513	0.528	0.578
23	28	24	41	21	25	0.394	0.400	0.481	0.513	0.580
24	21	25	23	27	84	0.212	0.360	0.400	0.429	0.464
25	27	84	66	24	28	0.170	0.184	0.290	0.360	0.361
27	25	84	66	86	28	0.170	0.293	0.326	0.357	0.371
28	25	27	23	24	41	0.361	0.371	0.394	0.481	0.484
29	36	38	34	32	67	0.106	0.225	0.245	0.261	0.268
31	35	36	68	29	67	0.187	0.290	0.306	0.318	0.323
32	39	38	29	36	60	0.162	0.191	0.261	0.289	0.307
33	89	35	31	88	90	0.265	0.353	0.382	0.384	0.335
34	29	36	60	39	88	0.245	0.268	0.315	0.378	0.382
35	31	98	95	36	90	0.187	0.215	0.222	0.291	0.316
36	29	38	95	34	32	0.106	0.254	0.258	0.268	0.289
38	32	29	36	85	39	0.191	0.225	0.254	0.298	0.299
39	32	60	59	29	38	0.162	0.203	0.267	0.275	0.299
41	44	1	45	7	32	0.245	0.388	0.395	0.413	0.481
44	7	41	43	1	45	0.196	0.245	0.274	0.330	0.396
45	1	3	2	4	7	0.137	0.309	0.334	0.375	0.380
47	48	7	8	59	6	0.039	0.400	0.412	0.415	0.423
48	47	7	59	43	39	0.039	0.395	0.415	0.421	0.428
50	51	6	10	59	11	0.095	0.228	0.244	0.265	0.295
51	50	10	11	56	59	0.095	0.209	0.276	0.288	0.312
54	10	11	55	53	51	0.280	0.303	0.330	0.391	0.426
55	53	56	14	54	10	0.104	0.232	0.322	0.330	0.403
56	55	12	51	53	14	0.232	0.248	0.288	0.311	0.330
57	59	51	12	60	56	0.313	0.325	0.331	0.337	0.339
61	81	63	62	84	65	0.055	0.128	0.338	0.371	0.371
62	61	81	65	63	88	0.338	0.339	0.385	0.456	0.555
63	61	81	84	86	65	0.128	0.163	0.260	0.367	0.388
65	88	68	61	62	63	0.173	0.315	0.371	0.385	0.388
66	86	84	25	27	85	0.241	0.264	0.290	0.326	0.365
67	85	68	29	86	38	0.156	0.200	0.268	0.282	0.309
68	67	31	65	86	85	0.200	0.306	0.315	0.320	0.339
69	72	70	74	98	75	0.074	0.132	0.225	0.270	0.309
70	69	72	74	75	98	0.132	0.197	0.244	0.357	0.371
72	69	70	74	98	75	0.074	0.197	0.203	0.247	0.308
73	79	75	99	74	92	0.177	0.319	0.363	0.376	0.382
74	72	69	70	79	73	0.203	0.225	0.244	0.325	0.376
75	98	90	79	97	72	0.175	0.182	0.287	0.303	0.308
79	73	75	74	80	97	0.177	0.287	0.325	0.372	0.399
80	79	73	97	74	75	0.372	0.444	0.615	0.626	0.642
81	61	63	62	84	65	0.055	0.163	0.339	0.394	0.420

Continue Table 5:

Focal record	Nearest neighbors					Distances				
	1	2	3	4	5	1	2	3	4	5
84	25	63	66	86	27	0.184	0.260	0.264	0.267	0.293
85	67	86	29	38	39	0.156	0.253	0.294	0.298	0.337
86	66	85	84	67	68	0.241	0.253	0.267	0.282	0.352
88	65	67	68	31	34	0.173	0.312	0.320	0.372	0.382
89	94	90	33	92	98	0.162	0.188	0.265	0.282	0.331
90	98	75	89	92	94	0.179	0.182	0.188	0.246	0.263
92	99	94	90	89	95	0.135	0.218	0.246	0.282	0.289
93	97	90	75	89	94	0.229	0.434	0.448	0.474	0.491
94	89	92	99	90	97	0.162	0.218	0.220	0.263	0.360
95	35	36	92	29	98	0.222	0.258	0.289	0.363	0.364
97	93	90	75	99	94	0.229	0.302	0.303	0.328	0.360
98	75	90	35	72	69	0.175	0.179	0.215	0.247	0.270
99	92	94	90	97	89	0.135	0.220	0.303	0.328	0.335

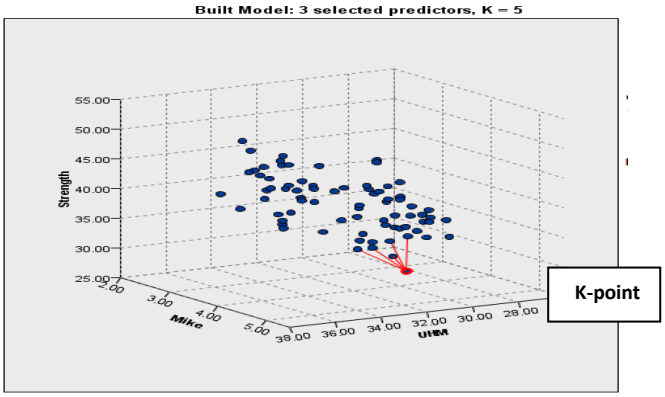

Figure 3: Nearest neighbor of strength, upper half mean and mike in combined data

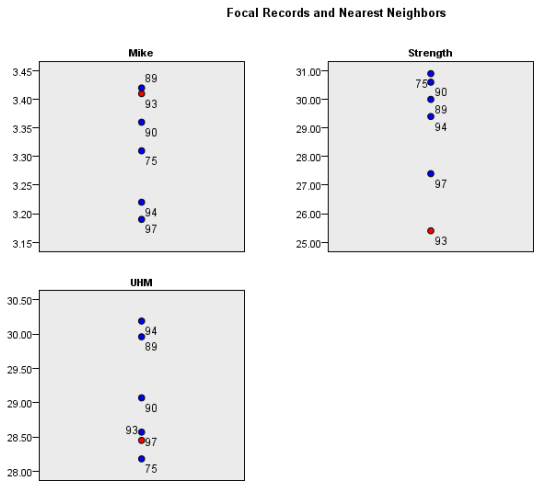

Figure 4: focal record and nearest neighbor for UHM, strength and mike in data mix

The technological value of cotton fiber properties derived by several methods; Table (6) illustrates that spinning consistency index (SCI) which is one of the most used method which ranged from low to high values for Egyptian cotton varieties according to treated data without using KNN and with KNN, respectively. The results in accordance with Screenivasa and Samanta (2000).

Table 6: Spinning consistency index (SCI) for Giza 87, Giza 88, Giza 86, Giza 90, Giza 95 and combined data

Cotton varieties	SCI	
	Without KNN	With KNN
Giza 87	166.63	185.97
Giza 88	164.74	179.49
Giza 86	163.74	174.85
Giza 90	156.44	179.49
Giza 95	159.23	177.27
Combined data	152.73	170.80

In terms of the basics of spinning consistency index (SCI) calculation. For all studied cotton varieties, namely; Giza 87, Giza 88, Giza 86, Giza 90, Giza 95 and combined data which both data treated without and with KNN were in acceptable range of technological value. Moreover, all previous cotton varieties were in a good use for determining technological value of cotton according to KNN application for all previous cotton varieties. A similar trend of results was detected by Majumdar et al. (2005).

Thus KNN helped in classifying studied data in two groups namely; training data and holdout data based on their acquired cotton length, strength and micronaire. It helped cotton researcher to easily collect data containing cotton properties information and evaluate it accordingly to tend for the other step needed analyses such as SCI.

4. Conclusion

K Nearest Neighbor (KNN) is easy to use, quick calculation time, does not need restricted assumptions about the data. Furthermore, accuracy depends on the quality of the data. KNN is widely used in real-life scenarios since it is non-parametric, meaning, it does not make underlying assumptions about the distribution of data. KNN is an example of a supervised learning machine learning method, which means we need first to feed it data so it is able to make a classification based on that data. KNN is a non parametric technique that stores all available cases and classifies new cases based on a similarity measure (distance function). KNN is often used in simple recommendation system, image recognition technology and decision making models.

In terms of using one of classification method such as NNA. It can be applied in research to determine various arrangements of features (central place) which will be useful in making useful and informed decisions about the environment. KNN may enrich understanding of cotton fiber properties as regard to distance. Where grade fiber properties section is the link between (vegetative and fruit properties) and industry spinning section so it is a useful tool in cotton fiber properties studies in the sense that it is able to study and describe cotton fiber technological properties as well as make a relatively accurate generalization on the basis of the interpreted data rather than making decisions on inferences as regard to point pattern based on subject and manual methods. The role of this specific machine learning model; nearest neighbor analysis was already pointed out for fiber quality properties. Additionally relationships were predicted for cotton vegetative and fruit properties with fiber quality properties and spinning properties which will need further research experiments.

References

[1] Abbott, A. M., Higgerson, G. J., Long, R. L., Lucas, S. R., Naylor, G. R. S., Tischler, C. R. and Purmalis, M. M. (2010). An instrument for determining the average fiber linear density (fineness) of cotton lint samples. Journal of Research Textile. 80: 822-833.

[2] Arshad, M., Hanif, M., Ilahi, N. and Shah, S. M. (1993). Correlation studies on some commercial cotton varities of G. hirsutum. Journal of Agricultural Sarhad. 9: 49-53.

[3] ASTM (2016). ASTM standard D1776/D1776M-16. Standard practice for conditioning and testing textiles. ASTM International, west Conshohocken. 1-5. with fiber characteristics in Gossypium hirsutum L. International Journal of Agriculture and Biology. 4: 129-131.

[4] Bartlett, M. S. (1978). Nearest neighbor models in the analysis of field experiments. 40(2): 147-174.

[5] Berman, M. (1986). Testing for spatial association between a point processes and another stochastic process. Applied Statistics. 35: 54-62.

[6] Braden, C. A. (2005). Inheritance of cotton fiber length and distribution. Texas A and M University. Ph.D. Diss.

[7] Bradow, J. M., Wartelle, L. I. I., Bauer, P. J. and Sassenrath-cole, G. F. (1997). Small fiber cotton quality quantitation. Journal of Cotton Science. 1: 48-60.

[8] C. A. T. G. O. (2016). Cotton Attribution and Testing General Organization, Publication of the Information and Documentation Center, Data of Cotton Season 2015/2016. November.

[9] Chaudhry, M. R. and Guitchounts, A. (2003). Fiber quality cotton facts. International Cotton Advisory Committee Technical Paper. 25: 85-89.

[10] Clouvel, P., Goze, E., Sequeira, R., Dusserre, J. and Cretener, M. (1998). Variability of cotton fiber quality, Proc. Of the World Cotton Research Conference 2, Athens, Greece, pp. 963-966.

[11] Docey, M. F. (1960). A note on the derivation of nearest neighbor distances. Journal of Regional Science. 81-87.

[12] E. C. G. (2015). The Egyptian Cotton Gazette, No. 144. April.

[13] Fiori, L. A. and Brown, J. J. (1951). Effects of cotton fiber fineness on the physical properties of single yarns. Journal of Textile Research. Industrial Section. October. 750-757.

[14] Foulk, J. A. and McAlister, D. D. (2002). Single cotton fiber properties of low, ideal and high micronaire values. Journal of Textile Research. 72(10): 885-891.

[15] Freeland, Jt. T. B., Pettigrew, B., Thaxton, P. and Andrews, G. L. (2006). Agronometeorology and cotton production. Chapter 13. A guide to agricultural meteorological practices. PP. 1-17.

[16] Ghosh, A. and Chatterjee, P. (2010). Prediction of cotton yarn properties using support vector machine. Journal of Fiber and Polymers. 11(1): 84-88.

[17] Ghosh, A., Das, S. and Majumder, A. (2015). A statistical analysis of cotton fiber properties. Journal of Institution of Engineers (India). 97(1): 1-7.

[18] Kumar, S. and Bindu, S. (2019). Medical image analysis using deep learning: A systematic literature review. In proceedings of emerging technologies in computer engineering: Microservices in big data analytics. Journal of Communications in Computer and Information Science. 985: 81-97.

[19] Karademir, E., Karademir, C., Ekinci, R. and Gencer, O. (2010). Relationship between yield, fiber length and other fiber-related traits in advanced cotton strains. Journal of Notulae Botanicae Horti Agrobotanici Cluj-Napoca. 38(3): 111-116.

[20] Long, R. L., Bange, M. P., Delhom, C. D., Church, J. S. and Constable, G. A. (2013). An assessment of alternative cotton fiber quality attributes and their relationship with yarn strength. Journal of Crop and Pasture Science. 64(8): 750-762.

[21] Lokhande, S. B. and Reddy K. R. (2015). Cotton reproductive and fiber quality responses to nitrogen nutrition. International Journal of Plant Production. 9: 191-209.

[22] Ma, C. M., Yang, W. S. and Cheng, B. W. (2014).How the parameter k-nearest neighbor algorithm impact on the best classification accuracy: in case of Parkinson dataset. Journal of Applied Sciences. 14: 171-176.

[23] Moujalvo, J. G. (2005). Relationship between micronaire, fineness and maturity. Part 1. Fundamentals. Journal of Cotton Science. 81-88.

[24] Moura, S. A. and Cordeiro, N. M. (2020). Got to write a classic : classical and perturbation based QSAR methods, machine learning, and the monitoring of nanoparticle ecotoxicity. Journal of Ecotoxicology. 495-213.

[25] Majumdar, A., Majumdar, P. K. and Sarkar, B. (2005). Determination of the technological value of cotton fiber: A comparative study of the traditional and multiple-criteria decision making approaches. Journal of AUTEX Research. 5(2): 71-80.

[26] Rodriguez, Y., De Baets, B., Garica, M. M., Morell, C. and Grau, R. (2008). A correlation based distance function for nearest neighbor classification. Iberoamerican Congress on Pattern Recognition. 284-291.

[27] Saputra, M. E., Mawengkang, H. and Nababan, E. B. (2019). Gini index with local mean based for determining K-value in K-nearest neighbor classification. The 3rd International Conference on

Computing and Applied Informatics. Journal of Physics: Conference series. doi: 10.1088/1742-6596/1235/1/012006.
[28] Shakhovska, N., Kaminskyy, R., Zasoba, E. and Tsiutsiura, M. (2018). Association roles mining in big data. Journal of Institute Computer. 17(1): 25-32.
[29] Smith, W. S. (1947). Air gauge measures fiber fineness. Journal of Textile Industries. 111: 86-88.
[30] SPSS (2012). IBM SPSS Statistics 21 Core System. User's Guide (edited by IBM ® SPSS ® statistics 21). U. S. government users restricted rights by GSA and ADP.
[31] Steel, R. G. D. and Torrie, J. H. (1980). Principles and procedures of statistics. McGraw- Hill Book Co., New York.
[32] Screenivasa,M. H. V. and Samanta, S. K. (2000). A fresh look at fiber quality index. Journal of The Indian Textile. 111(3): 29-37.
[33] Thompson, H. R. (1956). Distribution of distance to N-th neighbor in a population of randomly distributed neighbors. Journal of Ecology. 37: 391-394.
[34] Wang, J., Neskovic, P. and Cooper, L. N. (2007). Improving nearest neighbor rule with a simple adaptive distance measure. Pattern Recognition Lett., 28: 207-213.
[35] Wang, Y. H., Zheng, M., Gao, X. B. and Zhou, Z. G. (2012). Protein differential expression in the elongation cotton (Gossypium hirsutum L.) fiber under nitrogen stress. Journal of Science China. 55: 984-992.
[36] Wilson, R. and Martinez, T. (1997). Improved heterogeneous distance functions. Journal of Artificial Intelligence Research. 6: 1-34.
[37] Wright, S. (1921). Correlation and causation. Journal of Agricultural Research. 20: 557-585.
[38] Zhou, L., Pan, S., Wang, J. and Vasilakos, A. (2017). Machine learning on big data: opportunities and challenges. Journal of Neurocomputing. 237: 350-361.
[39] Zhao, X., Guo, X., Luo, J. and Tan, X. (2018). Efficient detection method for foreign fibers in cotton. Journal of Information in Agriculture. 5: 320-328.

Author Profile

Mona Shalaby was born on 16th march, 1977 in Sweden. She has master (2007) and Ph.D. (2015) in applied statistics from Faculty of Agriculture in Cairo University. She is currently a researcher in Cotton Grading Section. Cotton Research Institute at Agricultural Research Center (ARC), Egypt. Formerly, she was a researcher at Central Laboratory for Design and Statistical Analyses, ARC.