Original Article

Data Preprocessing Using Enhanced Principal Component Analysis (Epca) For Agricultural Datasets

[1] HARSHINI. N, [2]Dr. M. RATHAMANI

[1 2]Assistant professor,
[1 2]Department of Computer Science,
[1 2]NGM College, Pollachi.
DOI: 10.47750/pnr.2022.13.S10.495

Abstract

Data pre-processing is considered as the core stage in core and data mining. Standardization, discretization, and dimensionality decrease are notable strategies in data pre-processing. In this paper proposed Enhanced Principal component Analysis (EPCA) the effects of pre-processing strategies on the Agricultural for the accuracy of the dataset. Experiments were conducted utilizing the above-listed techniques and their singular outcomes were contrasted with one another. Enhanced PCA were tried for dimensionality decrease; besides, an existing methodology of PCA and KNN was attempted and the presentation showed superior characterization accuracy contrasted with the individual strategies.

Keywords: Agricultural, Data Mining, Data pre-processing, principal component analysis;

1. Introduction

Agricultural sector organizations attempt consistently to search information in immense databases for decision making. The condition of decision making can be changed by the utilization of information technology by which the farmers can yield in much better manner. In the agricultural field which is an exceptionally prevailing and significant field, data mining assumes a critical part. Normally the situation is that the answer for their concerns is far inside their reach. For this situation Data mining ends up being viable for making decisions on different issues related to agriculture field. Data mining, through better data analysis and management, can help related organizations to accomplish more prominent advantages. Data mining additionally gives client situated admittance to track down secret pattern in data. In this paper have examined different issues looked in agriculture sector and how data mining tackles the issue. Agricultural institutions use data mining applications for various regions, like prediction of problem, disease detection, enhancing the pesticide, etc for making optimal decisions. Consequently say that Data mining has turned into a shelter to agriculture sector.

1.1 Data Mining

Data mining is the computing system of finding patterns in huge data sets including strategies at the intersection of artificial intelligence, machine learning, statistics, and database frameworks. Data mining includes five-steps. They are:

1. Identifying the source information.

2. Separating the applicable information from the data.
3. Recognizing the key values from the removed data set.
4. Interpreting and reporting the outcomes.

Data mining is a process utilized by organizations to transform crude data into valuable information. By utilizing software to search for designs in huge groups of data, businesses can get familiar with their clients to foster more effective marketing strategies, increase sales and reduction costs.

1.2 Data Mining Is Used In Agriculture Sector

Excessive utilization of pesticides hampers the in general agricultural efficiency and to handle this issue there is a need to limit the utilization of pesticides in agriculture. Data mining can be utilized to configuration computerized frameworks to distinguish weeds filling in fields.

Agriculture is one of the significant sources in Indian Economy. Step by step, the populace increments, so the interest of the food additionally increments. To get cleanse of these circumstances of ranchers, analyst and agricultural researchers are endeavoring to get better crop yield, analyzing cycle of stowed away patterns according to various perspectives for classification and changed over into important information is called as data mining in which data is organized specifically regions like data repository. The productive examination utilizing data mining methods assist ranchers with taking decisions. This information assists them to decrease of expenses and expanding the creation with rating. This strategy utilized the Food and Horticulture Association of the Unified Countries (FAOSTAT) dataset which can be utilized as the open source. After gathering data, the dataset will be pre-processed to reduce the undesirable data to create data can be decipherable in the next stage.

1.3 Application of Data mining in smart agriculture

After some time, farmers should produce more food utilizing less resource. They should consume less water, and utilize fewer chemicals. Farmers need to maximize development and decrease costs, while consumers request good food. Thus, the agriculture industry is searching for new products, practices and technologies. These various requirements can be met through smart agriculture or accuracy agriculture.

Smart agriculture relies upon a progression of innovations that cooperate to empower data collection and analysis. By and large, these new innovations create tremendous measures of data. Smart agriculture varies from traditional agriculture in focusing on the collection and utilization of data to decide.

Outfitting a lot of existing yield, soil and environment data and examining new non-trial data enhances creation and makes agriculture stronger to climate change. Simultaneously, the data are immediately collected in sets that are excessively huge and complex to be dissected without adequate software. Data alone doesn't create experiences. Analyzes are expected to assist farmers with utilizing the entirety of this data. Applications in light of Data Mining are turning out to be smarter. In any case, the steady expansion in how much data can prompt data quality issues, and as these applications develop into large, real-time monitoring systems. In this way, for Data Mining, intelligent processing and analysis is an additional challenge due to the enormous number of heterogeneous and frequently unstructured data. Coordinating a wide range of data sources is another challenge.

The key characteristics of Data Mining in smart agriculture are as follows:

- Large volumes of data, so want to explore fast and effective mechanisms.
- Heterogeneous data sources and data types to coordinate: in smart agriculture, the data sources are assorted; for instance, really want to coordinate sensors data, cameras data, GPS data, etc and this multitude of data

are different in design, byte, binary, string, number, etc. Want to speak with various sorts of devices and various systems and furthermore need to remove data from web pages.

Data mining in smart agriculture are being utilized mostly for planning soil and water use, monitoring crops health, reducing and improving the utilization of natural resources, limiting the utilization of pollutants (for example pesticides, herbicides), working on the nature of the creation and so on. In this part, the Data Mining strategies used to take care of different agricultural issues are examined.

2. Literature Survey

1. PCA as a Dimensionality Reduction and Data Preprocessing Technique

Raksha Upadhyay et.al proposed Principal Component Analysis as a Dimensionality Reduction and Data Preprocessing Technique. These days, managing high aspect data has become normal in many fields, for example, image processing, medical diagnosis, biometric systems, e-commerce and so on. In data base with higher aspect the immediate utilization of data is costlier as far as its storage, processing, and mathematical computations and so on. In this paper, we examined Principal Component Analysis (PCA) as a dimensionality decrease method. PCA should be possible either by computing the covariance matrix or by Singular Value Decomposition (SVD). In this paper, we examined Principal Component Analysis (PCA) as a dimensionality decrease strategy. PCA should be possible either by working out the covariance matrix or by Singular Value Decomposition (SVD). Recording of RSS tests in numerous applications creates gigantic measure of data and its use is perplexing in limitation hubs as in WSN or IoT (Soni et al., 2017). So this data should be molded with a dimensionality decrease procedures of some kind or another. The significant point of PCA is to diminish data sets in higher aspect data to give calculation ease in different applications. In this paper, we concentrated on the PCA as a dimensionality decrease procedure. Among the two sorts determined we thought about covariance matrix computation strategy, because of its simplicity, for examining the principal components in the simulated RSS design. Principal components produced after PCA are symmetrical and uncorrelated, this reality is additionally justified by our simulation results.

2. Preprocessing Model for Handling Imbalanced Data using kNN

Preeti Nair et.al proposed Optimization of kNN Classifier Using Hybrid Preprocessing Model for Handling Imbalanced Data. The ordinary k Nearest Neighbor (kNN) classifier has many difficulties while managing issues brought about by imbalanced data sets. The classifiers are typically intended to further develop accuracy by decreasing the errors and accordingly, they don't depend on class distribution or proportion or balance of classes. So for handling imbalanced data issues data preprocessing techniques, for example, inspecting strategies are broadly utilized. Classification is a broadly utilized data mining technique. It is a cycle to dole out a case in a dataset to an objective class. The target of a classification is to precisely predict the class mark for every unknown question occasion. There are numerous classification techniques, which are broadly utilized, some of them are Innocent bayes, decision tree, kNN and some more. Here, our principal center is around kNN classification technique. KNN classifiers depend on advancing by tracking down similitudes between the occurrences. The likenesses are estimated by a distance recipe. An endeavor has made to support the presentation of kNN classification algorithm while handling imbalanced datasets. While characterizing imbalanced datasets there is an issue of high misleading negative rates, so to diminish bogus negative rates and to improve the accuracy, the dataset adjusted must be adjusted. Indeed, even with a well-trained classifier like kNN, we can't lessen the misleading negative rates.

3. Semantic Information Extraction from Multi-Corpora

Kumar S et.al proposed Semantic information extraction from multi-corpora using deep learning. Information extraction assumes an imperative part in natural language processing, to remove named substances and occasions from

unstructured data. Because of the dramatic data development in the agricultural area, extricating huge information has turned into a difficult undertaking. However existing deep learning based procedures have been applied in smart agriculture for crop development, crop illness location, weed expulsion, and yield production, still it is challenging to track down the semantics between separated information because of unswerving impacts of weather, soil, pest, and manure data. An underlying stage, which proposes a data preprocessing method for expulsion of equivocalness in input corpora, and the subsequent stage proposes an original deep learning-based long momentary memory with correction in Adam optimizer and multi-facet perceptron to track down agricultural-based named element recognition, occasions, and relations between them. The proposed algorithm has been prepared, tested on four information corpora i.e., agriculture, weather, soil, and pest and fertilizers. The trial results have been contrasted and existing methods and it was seen that the proposed algorithm outperforms Weighted-SOM, LSTM+RAO, PLR-DBN, KNN, and Naïve Bayes on standard boundaries like accuracy, sensitivity, and specificity.

4. Preprocessing Model Long Short-Term Memory (LSTM)

Zhang J et.al proposed developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas. Anticipating water table profundity over the drawn out in agricultural regions presents exceptional hardships in light of the fact that these regions have complicated and heterogeneous hydrogeological characteristics, limit conditions, and human exercises; likewise, nonlinear communications happen among these elements. Therefore, a new time series model in view of Long Transient Memory (LSTM) was created in this concentrate as an option in contrast to computationally costly actual models. The proposed model is made out of a LSTM layer with one more completely associated layer on top of it, with a dropout strategy applied in the primary LSTM layer. The proposed model purposes month to month water diversion, evaporation, precipitation, temperature, and time as information data to predict water table depth. A straightforward yet compelling normalization strategy was utilized to pre-process data to guarantee data on a similar scale. 14 years of data are isolated into two sets: training set (2000-2011) and approval set (2012-2013) in the analysis. True to form, the proposed model accomplishes higher R2 scores (0.789-0.952) in water table profundity prediction, when differentiated and the eventual outcomes of conventional feed-forward neural network (FFNN), which simply shows up at to some degree low R2 scores (0.004-0.495), demonstrating that the proposed model can preserve and learn previous information well. In this way, one can reason that the proposed model can act as an elective methodology predicting water table depth, particularly in regions where hydrogeological data are challenging to get.

5. Preprocessing Model Advanced Decision Tree (ADT)

Rajeswari S et.al proposed C5. 0: Advanced Decision Tree (ADT) classification model for agricultural data analysis on cloud. Agriculture assumes an essential part in India's economy and around 70% of individuals procure their pay through it and furthermore gives enormous work opportunities. The technological headway has prompted remarkable accomplishments in creating Agricultural based software applications to get quicker information. Yet, numerous ranchers are as yet applying the customary techniques for cultivating and thus the consequence of efficiency turns out to be exceptionally low. Agribusiness prediction process for natural and inorganic residences is an open issue and it depends on weather, soil fertility, water, seasons and ware costs, and so on. Soil fertility factor is principal critical to keep up with the yield development and increment the creation. The soil fertility levels assist the ranchers with distinguishing the lacks in the soil, in particular supplement content, soil type, pH esteem, EC (Electrical Conductivity) worth and soil texture and to pick the right harvests to expand the creation. In this work, as a curiosity, the soil fertility level is predicted by breaking down the Virudhunagar Locale Soil information and recommendations are offered for crop selection and planting by utilizing C5.0: Advanced Decision Tree (ADT) classifier algorithm. Utilizing this method, an Android based wireless applications named as Design of Smart Information System (DSIS) application has been made. The proposed application enacts the Global Positioning System (GPS) to distinguish the client area. The performance of proposed model is dissected and it is contrasted and the current arrangement model for agricultural data.

3. Proposed Methodology

3.1 Data Acquisition

The objective of this step is to distinguish and get all data-related issues. In this step, really want to distinguish the various data sources, as data can be gathered from different sources like records and databases. The quantity and quality of the gathered data will decide the productivity of the output. The more data, the more exact the prediction will be.

	State Name	District Name	Season	Crop	Area	Production
0	Tamil Nadu	ARIYALUR	Kharif	Rice	24574	NaN
1	Tamil Nadu	ARIYALUR	Whole Year	Arhar/Tur	909	NaN
2	Tamil Nadu	ARIYALUR	Whole Year	Bajra	565	NaN
3	Tamil Nadu	ARIYALUR	Whole Year	Banana	190	NaN
4	Tamil Nadu	ARIYALUR	Whole Year	Cashew nut	31113	NaN
…	…	…	…	…	…	…
13522	Tamil Nadu	VIRUDHUNAGAR	Whole Year	Small millets	1187	676.0
13523	Tamil Nadu	VIRUDHUNAGAR	Whole Year	Sugarcane	2917	296017.0
13524	Tamil Nadu	VIRUDHUNAGAR	Whole Year	Sweet potato	4	84.0
13525	Tamil Nadu	VIRUDHUNAGAR	Whole Year	Tapioca	4	120.0
13526	Tamil Nadu	VIRUDHUNAGAR	Whole Year	Turmeric	4	15.0
13527 rows x 7 columns						

Table 1.Data Acquisition

3.2 Data Preparation

Data preparation is a step where put our data into a reasonable spot and prepare it to use in our AI preparing. Understanding the idea of data that should work with is utilized. This research really wants to grasp the characteristics, format, and nature of data. Data wrangling is the method involved with cleaning and changing over crude data into a usable format. It is the most common way of cleaning the data, choosing the variable to utilize, and transforming the data in a legitimate format to make it more reasonable for analysis in the following stage. It is one of the main steps of the complete interaction. Cleaning of data is expected to address the quality issues.

Different crop Dataset in array form

array(['Rice', 'Small millets', 'Arhar/Tur', 'Bajra', 'Banana',

'Cashewnut', 'Castor seed', 'Coconut ', 'Coriander',

'Cotton(lint)', 'Dry chillies', 'Groundnut', 'Jowar', 'Maize',

'Moong(Green Gram)', 'Onion', 'Ragi', 'Sesamum', 'Sugarcane',

'Sunflower', 'Sweet potato', 'Tapioca', 'Turmeric', 'Urad',

'Horse-gram', 'Tobacco', 'Black pepper', 'Cardamom', 'Gram',

'Pulses total', 'Total foodgrain', 'Wheat', 'Sannhamp', 'Korra',

'Samai', 'Guar seed', 'Other Cereals & Millets',

'Other Kharif pulses', 'Rapeseed &Mustard', 'Varagu', 'Ash Gourd',

'Beans & Mutter(Vegetable)', 'Beet Root', 'Bhindi', 'Bitter Gourd',

'Bottle Gourd', 'Brinjal', 'Cauliflower', 'Citrus Fruit',

'Cucumber', 'Drum Stick', 'Garlic', 'Grapes', 'Jack Fruit',

'Lab-Lab', 'Mango', 'Orange', 'Other Citrus Fruit',

'Other Fresh Fruits', 'Other Vegetables', 'Papaya', 'Pome Fruit',

'Pome Granet', 'Redish', 'Ribed Guard', 'Snak Guard', 'Tomato',

'Water Melon', 'Yam', 'Cabbage', 'Pump Kin', 'Dry ginger',

'Arecanut', 'Potato', 'Carrot', 'Pineapple', 'Mesta', 'Plums',

'Turnip', 'Pear', 'Ber']

Different district_ name wise Dataset in array form

array(['ARIYALUR', 'COIMBATORE', 'CUDDALORE', 'DHARMAPURI', 'DINDIGUL',

'ERODE', 'KANCHIPURAM', 'KANNIYAKUMARI', 'KARUR', 'KRISHNAGIRI',

'MADURAI', 'NAGAPATTINAM', 'NAMAKKAL', 'PERAMBALUR', 'PUDUKKOTTAI',

'RAMANATHAPURAM', 'SALEM', 'SIVAGANGA', 'THANJAVUR',

'THE NILGIRIS', 'THENI', 'THIRUVALLUR', 'THIRUVARUR',

'TIRUCHIRAPPALLI', 'TIRUNELVELI', 'TIRUPPUR', 'TIRUVANNAMALAI',

'TUTICORIN', 'VELLORE', 'VILLUPURAM', 'VIRUDHUNAGAR']

3.3 Data Pre-processing

The objective of this step is to study and comprehend the idea of data that was procured in the previous step and to know the nature of data. In this step, will check for any invalid values and eliminate them as they might influence the effectiveness. Recognizing copies in the dataset and eliminating them is likewise finished in this step. In the pre-processing utilized a strategy called data fighting that is utilized to choose the features to utilize, changing over the obtained data in the dataset to a format that would be reasonable for subsequent stages and cleaning of data points.

To optimize prediction, need to clean and prune the dataset as the preprocessing and selection has the more prominent impact on the computational effectiveness and predictive accuracy. Incomplete and conflicting information altogether affect analysis and may lead the most obviously terrible prediction.

Datasets might incorporate multiple boisterous, incomplete, conflicting and superfluous features that ought to be tended to. Critically, the selection of legitimate dataset for grouping extensively affects prediction accuracy. Waikato Environment for Knowledge Analysis (WEKA) is an open source artificial intelligence to instrument, contains different data mining computation including, portraying data calculations. In this paper are involving WEKA for data mining functions and approaches to concentrate and develop the standards. In excess of 1000 data passages are gathered for this paper. These sections are then changed over completely to the ARFF format, a reasonable format for WEKA. The channels accessible for preprocessing in WEKA for example Eliminate R-1, Replace Missing Values and Discretized have assisted with changing over this data into cognizant and commotion Free State. The resulted dataset comprises of 760 data sections of various attributes as portrayed previously.

In the first stage, pre-processing of data was performed by predicting the missing values and normalizing the information. The missing values that were in "this present reality" agricultural datasets were predicted utilizing the "mean imputation" strategy. The agricultural information was exposed to standardization in the scope of 0 to 1 for diminishing the predominance of features with higher values over features with lower values. Prediction of missing values was performed since they lead to inaccurate predictions as the data won't be complete. Enhanced PCA is utilized for choosing applicable features and concentrate features from these chose features.

Missing data values cause the data incomplete which causes the ML models to predict inaccurately. In this paper, the 8 "realworld datasets" contained "missing values". These values were predicted utilizing the "mean imputation" strategy in which the omitted values were replaced by the mean/normal of the known values of that feature. "Mean imputation" strategy was picked since its presentation was viewed as better compared to different methods like "complete case analysis" and "multiple imputations".

The first cultivating data was pre-handled between values 0 and 1. This forbids the higher esteemed features from overwhelming the lower esteemed features during prediction. Eq. (1) shows how esteem (ti) is normalized for the feature G in column for 'i'.

$$\boldsymbol{Normalized}\ (\boldsymbol{t_i}) = \frac{t_i - G_{min}}{G_{max} - G_{min}} \quad \textbf{(1)}$$

Where, G_{min} and G_{max} are the minimum and maximum values of feature G, respectively.

EPCA is a method that changes over a bunch of perceptions of potentially connected factors into a bunch of values directly uncorrelated factors called principal components. The transformed dataset is defined in such a way that the first principal components account for much of the variance. Principal components are destined to be autonomous assuming that the data set is mutually typically dispersed.

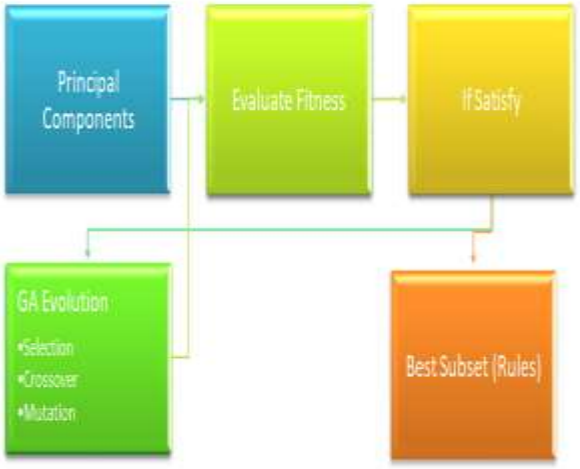

Figure 1.Enhanced Principal Component Analysis

The idea is to execute the application of Principal Component Analysis to diminish the dimensionality of a dataset to a feature set called principal components. The principal components are then utilized as info population in the pursuit space of the GA in looking for the ideal arrangement. This system proficiently works on the data mining process utilizing the representative data of the first dataset, to which decreases computational time and further develops order execution of classifiers Be that as it may, the EPCA procedure tends to lose data interpretability yet has high discriminative power. To defeat the inadequacies of this interaction, a feature subset selection method in view of a changed GA is utilized. In this specific circumstance, utilizing different classifiers is investigated and taken on as the fitness capability. The fitness capability in GA is altered likewise utilizing productive variety of distance measures between features, this gives better partition of the pattern classes, which, thus, diminishes complexity and works on the presentation of classifiers and decrease computational costs.

The Enhanced PCA Algorithm

Step 1: Start the process

Step 2: Fix the Principal components as population

Step 3: Register and assess the fitness of every principal component in the population

Step 4: Test the end condition is fulfilled, pause and return the best arrangement in current population, in any case,

Step 5: Make new population by rehashing the accompanying strides until the new population is finished

Step 6: Select two parent chromosomes from a populace according to their fitness (the better fitness, the more prominent chance to be picked)

Step 7: With a crossover probability moves past the watchmen to approach another successors (child). Assuming no crossover was performed, any kind of family down the line is a precise duplicate of guardians.

Step 8: With mutation likelihood mutate new posterity

Step 9: Accept Spot new posterity in a new population

Step 10: Replace and Utilize new produced population

Step 11: Loop then Go to step 3.

Step 12: Stop the Process

4. Experimental Result

1. Accuracy

Accuracy is the degree of closeness between a measurement and its true value. The formula for accuracy is:

$$\boldsymbol{Accuracy} = \frac{(\boldsymbol{true\ value} - \boldsymbol{measured\ value})}{\boldsymbol{true\ value}} * \mathbf{100}$$

Dataset	KNN	PCA	Proposed Enhanced PCA
100	85.12	84.37	98.67
200	81.69	82.82	96.26
300	78.62	80.54	95.21
400	75.55	78.63	92.58
500	73.94	74.72	89.87

Table 2.Comparison tale of Accuracy

The Comparison table 2 of Accuracy demonstrates the different values of existing KNN, PCA and proposed Enhanced PCA. While comparing the Existing algorithm and proposed Enhanced PCA, provides the better results. The existing algorithm values start from 73.94 to 85.12, 74.72 to 84.37 and proposed Enhanced PCA values starts from 89.87 to 98.67. The proposed method provides the great results.

Figure 2.Comparison chart of Accuracy

The Figure 2 Shows the comparison chart of Accuracy demonstrates the existing KNN, PCA and proposed Enhanced PCA. X axis denote the Dataset and y axis denotes the Accuracy ratio. The proposed Enhanced PCA values are better than the existing algorithm. The existing algorithm values start from 73.94 to 85.12, 74.72 to 84.37 and proposed Enhanced PCA values starts from 89.87 to 98.67. The proposed method provides the great results.

2. Precision

Precision is a measure of how well a model can predict a value based on a given input. The precision of a model is the ratio of true positive predictions to all positive predictions.

$$\boldsymbol{Precision} = \frac{\boldsymbol{true\ positive}}{(\boldsymbol{true\ positive}\ +\ \boldsymbol{false\ positive})}$$

Dataset	KNN	PCA	Proposed Enhanced PCA
100	68	73	89
200	70	70	90
300	75	66	91
400	80	69	94
500	87	64	98

Table 3.Comparison table of Precision

The Comparison table 3 of Precision demonstrates the different values of existing KNN, PCA and proposed Enhanced PCA. While comparing the Existing algorithm and proposed Enhanced PCA, provides the better results. The existing algorithm values start from 68 to 87, 64 to 73 and proposed Enhanced PCA values starts from 89 to 98. The proposed method provides the great results.

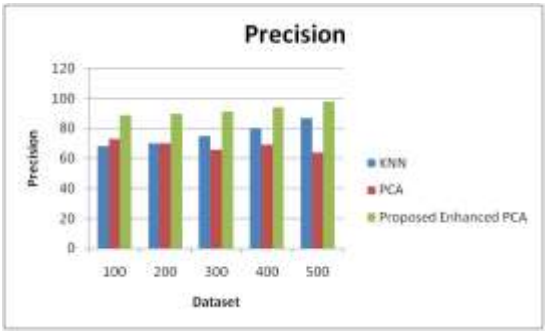

Figure 3.Comparison chart of Precision

The Figure 3 Shows the comparison chart of Precision demonstrates the existing KNN, PCA and proposed Enhanced PCA. X axis denote the Dataset and y axis denotes the Precision ratio. The proposed Enhanced PCA values are better than the existing algorithm. The existing algorithm values start from 68 to 87, 64 to 73 and proposed Enhanced PCA values starts from 89 to 98. The proposed method provides the great results.

3. Recall

Recall is a measure of a model's ability to correctly identify positive examples from the test set:

$$Recall = \frac{True\ Positives}{(True\ Positives + False\ Negatives)}$$

Dataset	KNN	PCA	Proposed Enhanced PCA
100	0.72	0.82	0.85
200	0.76	0.75	0.88
300	0.80	0.69	0.92
400	0.82	0.72	0.95
500	0.85	0.69	0.98

Table 3.Comparison tale of Recall

The Comparison table 3 of Recall demonstrates the different values of existing KNN, PCA and proposed Enhanced PCA. While comparing the Existing algorithm and proposed Enhanced PCA, provides the better results. The existing algorithm values start from 0.72 to 0.85, 0.69 to 0.82 and proposed Enhanced PCA values starts from 0.85 to 0.98. The proposed method provides the great results.

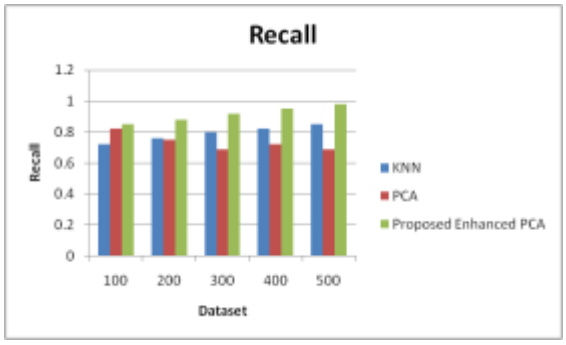

Figure 3.Comparison chart of Recall

The Figure 3 Shows the comparison chart of Recall demonstrates the existing KNN, PCA and proposed Enhanced PCA. X axis denote the Dataset and y axis denotes the Recall ratio. The proposed Enhanced PCA values are better than the existing algorithm. The existing algorithm values start from 0.72 to 0.85, 0.69 to 0.82 and proposed Enhanced PCA values starts from 0.85 to 0.98. The proposed method provides the great results.

F -Measure

F1-measure is a test's accuracy that combines precision and recall. It is calculated by taking the harmonic mean of precision and recall.

$$\mathbf{F1 - Measure} = \frac{(\mathbf{2 * Precision * Recall})}{(\mathbf{Precision + Recall})}$$

Dataset	KNN	PCA	Proposed Enhanced PCA
100	0.88	0.73	0.98
200	0.86	0.71	0.96

300	0.84	0.68	0.94
400	0.82	0.65	0.93
500	0.80	0.63	0.92

Table 4.Comparison tale of F -Measure

The Comparison table 4 of F -Measure Values explains the different values of existing KNN, PCA and proposed Enhanced PCA. While comparing the Existing algorithm and proposed Enhanced PCA, provides the better results. The existing algorithm values start from 0.80 to 0.88, 0.63 to 0.73 and proposed Enhanced PCA values starts from 0.92 to 0.98. The proposed method provides the great results.

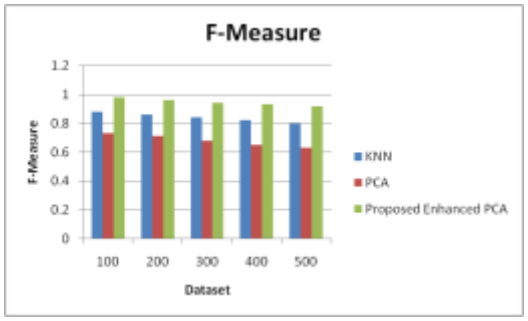

Figure 4.Comparison chart of F -Measure

The Figure 4 Shows the comparison chart of F -Measure demonstrates the existing KNN, PCA and proposed Enhanced PCA. X axis denote the Dataset and y axis denotes the F -Measure ratio. The proposed Enhanced PCA values are better than the existing algorithm. The existing algorithm values start from 0.80 to 0.88, 0.63 to 0.73 and proposed Enhanced PCA values starts from 0.92to 0.98. The proposed method provides the great results.

5. Conclusion

Agricultural institutions use data mining applications for various regions, like prediction of problem, disease detection, enhancing the pesticide, etc for making optimal decisions. Data pre-processing is a basic stage in data mining. In this paper we proposed Enhanced Principal component Analysis (PCA) the impacts of pre-processing strategies on the Agricultural dataset was utilized in the trial while on the accuracy of the data pre-processing. In light of the aftereffects of the examination, the execution of the calculation in view of Enhanced PCA is proficient in optimizing the data mining process, creating grouping models and rules for agricultural. From the outcomes, shows the Enhanced PCA to acquire high accuracy with least time.

References

1. A. Mukherjee, M. Velez-Reyes and B. Roysam, "Interest Points for Hyperspectral Image Data," in IEEE Transactions on Geoscience and Remote Sensing, vol. 47, no. 3, pp. 748-760, March 2009, doi: 10.1109/TGRS.2008.2011280.
2. A. Schlamm and D. Messinger, "A euclidean distance transformation for improved anomaly detection in spectral imagery," 2010 Western New York Image Processing Workshop, Rochester, NY, USA, 2010, pp. 26-29, doi: 10.1109/WNYIPW.2010.5649762.
3. H. Wang, G. Li, Z. Ma and X. Li, "Image recognition of plant diseases based on backpropagation networks," 2012 5th International Congress on Image and Signal Processing, Chongqing, China, 2012, pp. 894-900, doi: 10.1109/CISP.2012.6469998.
4. J. Karami, A. Alimohammadi and S. Modabberi, "Analysis of the Spatio-Temporal Patterns of Water Pollution and Source Contribution Using the MODIS Sensor Products and Multivariate Statistical Techniques," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 5, no. 4, pp. 1243-1255, Aug. 2012, doi: 10.1109/JSTARS.2012.2187273.
5. Kumar S, Sastry GH, Marriboyina V, Alshazly H, Idris SA, Verma M, Kaur M. Semantic information extraction from multi-corpora using deep learning. Computers, Materials and Continua. 2021:1-7.

6. L. Feng et al., "Detection of Subtle Bruises on Winter Jujube Using Hyperspectral Imaging With Pixel-Wise Deep Learning Method," in IEEE Access, vol. 7, pp. 64494-64505, 2019, doi: 10.1109/ACCESS.2019.2917267.
7. L. Franceschelli, C. Cevoli, A. Benelli, E. Iaccheri, M. Tartagni and A. Berardinelli, "Vis/NIR hyperspectral imaging to assess freshness of sardines (Sardina pilchardus)," 2020 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor), Trento, Italy, 2020, pp. 124-128, doi: 10.1109/MetroAgriFor50201.2020.9277603.
8. Nair P, Kashyap I. Optimization of kNN classifier using hybrid preprocessing model for handling imbalanced data. Int. J. Eng. Res. Technol. 2019;12(5):697-704.
9. Rajeswari S, Suthendran K. C5. 0: Advanced Decision Tree (ADT) classification model for agricultural data analysis on cloud. Computers and Electronics in Agriculture. 2019 Jan 1;156:530-9.
10. S. -L. Wu, H. -Y. Tung and Y. -L. Hsu, "Deep Learning for Automatic Quality Grading of Mangoes: Methods and Insights," 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), Miami, FL, USA, 2020, pp. 446-453, doi: 10.1109/ICMLA51294.2020.00076.
11. S. Mishra, D. Mishra, S. Das and A. K. Rath, "Feature reduction using principal component analysis for agricultural data set," 2011 3rd International Conference on Electronics Computer Technology, Kanyakumari, India, 2011, pp. 209-213, doi: 10.1109/ICECTECH.2011.5941686.
12. Upadhyay R, Panse P, Soni A, Rathore Bhatt U. Principal component analysis as a dimensionality reduction and data preprocessing technique. Proceedings of Recent Advances in Interdisciplinary Trends in Engineering & Applications (RAITEA). 2019 Apr 2.
13. Y. Tang et al., "Apple Bruise Grading Using Piecewise Nonlinear Curve Fitting for Hyperspectral Imaging Data," in IEEE Access, vol. 8, pp. 147494-147506, 2020, doi: 10.1109/ACCESS.2020.3015808.
14. Y. Yao, H. Si and D. Wang, "Object oriented extraction of reserve resources area for cultivated land using RapidEye image data," 2014 The Third International Conference on Agro-Geoinformatics, Beijing, China, 2014, pp. 1-4, doi: 10.1109/Agro-Geoinformatics.2014.6910671.
15. Zhang J, Zhu Y, Zhang X, Ye M, Yang J. Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas. Journal of hydrology. 2018 Jun 1;561:918-29.