IMPLEMENTATION OF COMPARATIVE ANALYSIS ON AGRICULTURE DOMAIN BY USING CLASSIFICATION TECHNIQUES

Paramjit[1], NishaCharaya[2]
M.Tech Scholar, Assistant Professor
Computer Science and Engineering
OM Institute of Technology and Management, Juglan (Hisar) ,Haryana,India

Abstract: This paper reviews various articles and documents on agricultural production. Agriculture the important industrial sectors in India and the country's economy is highly dependent on it for rural sustainability. Agriculture contributes round 34% to the gross domestic product (GDP) and offers employments to approximately 36% of households inside the monetary yr 2016. Agriculture influences weather via emissions of greenhouse gases (GHGs) including carbon dioxide, methane, and nitrous oxide. Agriculture can be a solution for weather trade by means of the large adoption of mitigation and edition actions.

***Index Terms*- Climate change, Greenhouse gas emissions**.

I. INTRODUCTION

The main idea of this paper is based on Data Analytics. In this paper different datasets are collected from various sources as input. Classification technique applied on dataset and then analysis of crops, yield production i.e. descriptive analytics is performed on the sugarcane datasets (Soil, Rainfall and Yield). Results are shown by supervise learning algorithmswhere how much crop is produced in last year's and categorized their respective class. The results of different classification technique (KNN, SVN and LS-SVN) are compared.

II.RELEATED WORK

Rajagopalan B, Lall 1999, in this paper, the author developed a random sequence of day to day climate parameters and also metrological parameters which are based on a multivariate, non-parametric time sequence analysis. The historical data of weather and atmosphere is available at the site. A list of metrological variables and weather parameters tends to derive the solar radiation, humidity, maximal and minimal temperature, average wind speed, On the basis of random day data sample sampling is performed which denials of baseness. The author used KNN approach which is equivalent to a non-parametric approximation of a multivariate sample set. It doesn't demand prior assumptions as to the procedure of the combined likelihood compactness function of the variables. The Results are associated with those from the presentation of a multivariate autoregressive model.[1]

Zhenmin, Sudarshan, Srinivasan, Zhifeng Chen 2003, in order to beself-tuning, self-managing, and self-healing and self-protecting, a storage system needs in order to mechanically represent access patterns. This paper proposes a method that uses data mining techniques to systematically mine get right of entry to sequences in a storage device to characterize storage behaviours. More specifically, They use frequent sequence mining algorithms to locate block get right of entry to correlations which may be used to enhance the effectiveness of subsystems such as storage caching and disk scheduling, and for disk power management. This paper reports their Initial consequences of coming across block correlations from storage access sequences using a lately proposed data mining algorithm says CloSpan.[2]

Tripathi S, Srinivas 2006, in this paper, the author studies of climate impact inhydrology which is usually based on climate changes. In last two decades the climate is change drastically. Yet, GCM (General Circulation Models), these are still most advanced and latest tolls to forecast the upcoming weather conditions. These tools operate on a rough scale. Consequently the output from a GCM has to be economized to get the info which is relevant to hydrologic revisions. The author used SVM for numerical rationalizing of rainfall, snowfall, and other factors changes at monthly time scale. The deftness of this method is demonstrated through its application to climatologically sub-divisions in India. Primary, weather variables upsetting spatial-temporal disparity of precipitation at each climatologically sub-divisions in India are identified, and classified using clustering they made two sets, one is represent wet and the other one is represent dry season. For each season, SVM- based economizing model is developed for season so that they can make better prediction about rainfall, snowfall or any other natural factor. The proposed model is shown to be greater too predictable economizing with multi-layer back-propagation artificial neural networks. Afterward, the SVM-based model is realistic to future climate predictions from the additional group United Worldwide Weather Archetypal to obtain forthcoming forecasts of rainfall for the model. The results are then examined to measure the influence of weather alteration on rainfall over India. It is exposed that SVMs offer a likely another conventional artificial neural networks for numerical rationalizing, and are appropriate for leading environment influence studies.[3]

Wu X, Kumar, et al. 2008,in this paper the author reviewed most used widelymethod of data mining, like as KNN, SVM, Apriori, C4.5 algorithm etc. These algorithms are amongst the greatest influential mining algorithms in this paper and as well as commercial field. On this paper they endure describe of each algorithm and application where these algorithms can be

implemented. All these algorithms are used in organize and un organize learning or you can say that machine learning is somehow based on these algorithms.[4]

Yun-lei , Duo ,Dong-fengCai 2010, On this paper we discuss on KNN classification method neighbour.KNN is an algorithms that is based on machine learning. In this training parameters and computational complexity are not high, that the reason we choose KNN as our system framework.In this we use BM25 calculation methods to explain the problem. He says The NTCIR-8 patent mining assignment is to classify research papers written in Japanese or English into the IPC at subclass, main group and subgroup levels. On the results, we can see the KNN+SNN perform best on both corps. The overall performance of both approach has progressed, which suggests that KNN based on shared nearest neighbours (SNN) is closely associated with the density of classes. [5]

III. Classification Algorithms:

KNN (K-nearest neighbour):KNN Has been used in statistical estimation and sample reputation already within the beginning of 1970's.KNN is a simple algorithm that shops all to be had case and classifies new instances based totally on a similarity degree. It is the based on the machine learning where problem is solved by the nearest neighbour . If two classes are existing and new member is coming than select that member to class that near to him. Near means distance and if multiple neighbour than choose close of the class for example

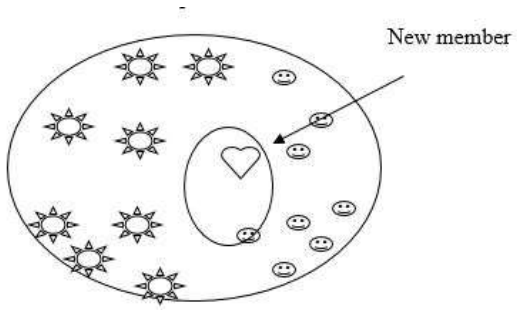

Figure: KNN

In this, two different type classes and one new member add and find the close class and then add that new member on that class that is KNN algorithm.

The main disadvantage of KNN is the involution in search the nearest neighbours for each sample.

SVM (Support Vector Machine):

SVM introduced in COLT-92 by Boser, Guyon and Vapnik in 1963. It is related learning algorithms that investigate information utilized for characterization and relapse examination. Support Agent Apparatus is an adorable adjustment due to its top generalization adequacy and its adeptness to handle high-dimensional ascribe data. Compared to neural networks or accommodation trees, SVM do not ache from the bounded minima problem, it has beneath acquirements ambit to select, and it produces abiding and reproducible results. If two SVMs are accomplished on the aforementioned abstracts with the aforementioned acquirements parameters, they aftermath the aforementioned after-effects absolute of the enhancement algorithm they use. However, SVMs ache from apathetic training abnormally with non-linear kernels and with ample ascribe abstracts size. Support agent machines are primarily bi-fold classifiers. Extensions to multi-class problems are a lot of generally done by accumulation several bi-fold machines in adjustment to aftermath the final multi-classification results. The added difficult botheration of training one SVM to allocate all classes uses abundant added circuitous enhancement algorithms and are abundant slower to alternation than bi-fold classifiers. The next sections, present SVM algebraic foundation for the bi-fold allocation case, again altercate the altered approaches activated for multi-classification.

LS-SVM Vs SVM:

LS-SVM capital advantage is that it is computationally added able than the accepted SVM method. In this case training requires the Band-Aid of a beeline blueprint set instead of the continued and computationally harder boxlike programming botheration complex by the accepted SVM. The adjustment finer reduces the algebraic complexity, about for absolutely ample problems, absolute an actual ample amount of training samples, even this least-squares Band-Aid can become awful anamnesis and time consuming.

Whereas the atomic squares adaptation incorporates all training abstracts in the arrangement to aftermath the result, the acceptable SVM selects some of them (the abutment vectors) that are important in the regression. The absence of acceptable SVM can as well be accomplished with LS–SVM by applying a pruning method. Unfortunately if the acceptable LS–SVM pruning adjustment is applied, the achievement declines proportionally to the alone training samples, back the advice (input output relation) they declared is lost. Another botheration is that this accepted adjustment multiplies the algebraic complexity.

IV. RESULTS AND DISCUSSION

This work will apply the data analytics on agriculture domain and categorize the data into separate classes by performing supervised training on the dataset that are collected from agriculture domain. This system has the capability to perform both the classification as well as regression. In the classification step the data is classified into three classes (low, mid, and high), whereas in regression step the actual cost of yield production is estimated.

4.1 User Interface for the analytics

In this phase, we build graphical user interface so that the person can interaction with our system easily. Our device has the ability to train datasets using various algorithms. For this reason we create user interface to present the power to the user to locate end result using distinctive algorithms and additionally offers facility to evaluate the result.
Figure 3.4 shows user interface.

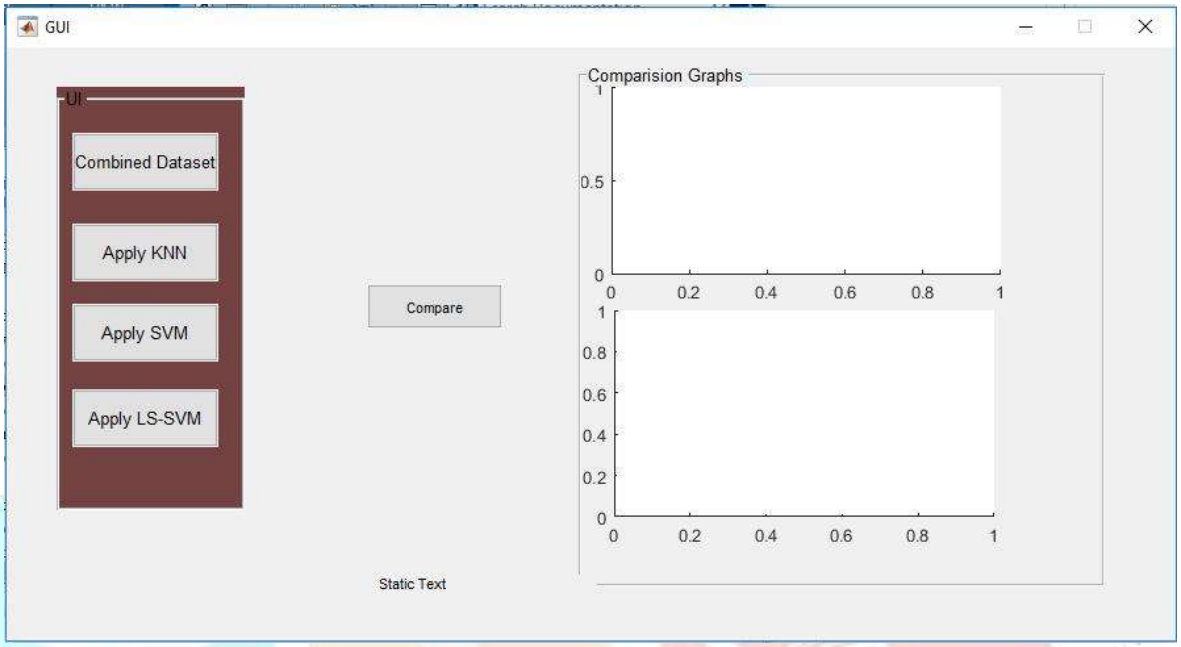

Figure 4.1: User Interface of the Proposed System

UI panel (on the left hand side) has functionality such as Combined Dataset, Apply KNN algorithm, SVM algorithm, and Apply LS-SVM algorithm. When we click on any button the call-back method is called and action is performed. Whereas, on the right hand facet the graphs which suggests the evaluation outcomes.

4.1.1 Combined Dataset

In this function, the parameters of all the dataset are combined into a single table which is called 'Final'. If we do not combine all the datasets together then we need to perform separately each table and somehow it could be extravagant process this approach we have a unique key "Year". On the basis of this key all the parameters are mapped and train by the several supervised mastering algorithms. The results of combined dataset are shown in Figure 4.2, Figure 4.3, Figure 4.4, and Figure 4.5

1 YEAR	2 pH	3 Organiccarbon	4 Nitrogen	5 Phosphorus	6 Potassium	7 Zinc	8 Iron
1901	6.1000	8.5000	282	29	208	412	7.1000
1902	7.7000	6.1000	357	17	232	582	8.1000
1903	6.7000	6.8000	424	9	399	139	8.6000
1904	7	2.7000	223	24	426	426	6.7000
1905	6.8000	7.1000	348	5	215	193	8.1000
1906	5	4.2000	448	38	342	539	5.9000
1907	5.7000	9	232	7	397	482	8.5000
1908	7	6.1000	275	28	253	368	6.4000
1909	6.1000	8.8000	169	9	451	474	5.4000
1910	6.9000	2.7000	578	34	375	311	6.4000
1911	5.7000	2.5000	252	21	377	510	5.2000
1912	8	9.8000	387	8	438	583	7.1000
1913	7.2000	3	314	21	550	547	8.9000
1914	5	5.2000	408	12	402	177	8.2000
1915	6.6000	6.9000	219	28	226	532	7.8000
1916	6.8000	4.8000	305	18	380	216	6.4000
1917	7.4000	8.1000	218	37	423	282	7.6000
1918	5.2000	8	362	12	518	326	6.6000
1919	6.4000	5.7000	439	24	223	226	7.7000
1920	5.3000	8.1000	555	35	228	376	8.6000
1921	5.9000	8.2000	167	22	239	473	9

Figure 4.2: Combined dataset I

The Figure 4.2 shows the partial tabular representation of combined dataset. It includes Year as the key parameter and other parameters of soil dataset.

9 Copper	10 Manganese	11 Sulphur	12 Fibre	13 Sugar	14 JAN	15 FEB	16 MAR
2.6000	1	32	17.2000	10.8667	34.7000	38.6000	17.8000
2	1.3000	26	15.7059	12.1888	7.4000	4.2000	19
2.7000	2.8000	39	15.2000	13.7500	16.7000	8	31.1000
3.5000	2.9000	18	15.4000	13.4133	14.9000	9.7000	31.4000
3.6000	1.9000	8	14.8000	12.7033	24.7000	20.3000	41.8000
3.4000	3.3000	34	15.2200	11.9830	21.4000	49.9000	31.4000
3.2000	2.3000	15	14.9000	12.6400	16	45.5000	37.4000
2.1000	2.7000	38	14.9000	13.2200	19.9000	17.1000	8.3000
3.8000	1	37	15.0333	15.5700	22.7000	15.2000	6.6000
2.3000	1.8000	25	16.7000	13.2300	13.5000	10.3000	13.7000
2.2000	3.3000	23	16.6000	13.2967	40.4000	5.5000	43
2.5000	3.5000	38	15	13.2360	20.3000	21.6000	19.9000
2.4000	1.8000	25	15.4818	11.8591	6.3000	38.1000	23.7000
2.8000	2.1000	12	15.3000	12.3643	5	26.9000	25.4000
2.4000	2.9000	17	16.1000	11	19.8000	37.5000	44.1000
3.7000	3.3000	10	16.8667	12.2200	4.6000	20.1000	11
2.5000	3.2000	16	19.1000	11.5800	7.6000	37.9000	20.5000
2	1	32	15.3000	11.8800	11.8000	4	36.6000
3.2000	1.8000	19	16.8000	12.6400	48.8000	20.2000	19.1000
3.9000	3.6000	17	16.0650	13.3900	23.9000	21.3000	55.1000
2.5000	2.3000	31	15.1143	14.0020	37.6000	7.4000	17.8000

Figure 4.3: Combined dataset II

The figure 4.3 has the parameters of nutrient dataset and some parameters of rainfall dataset (Jan, Feb, and Mar).

17 APR	18 MAY	19 JUN	20 JUL	21 AUG	22 SEP	23 OCT	24 NOV
38.9000	50.6000	113.2000	241.4000	271.6000	124.7000	52.4000	38.7000
44.1000	48.8000	111.7000	284.9000	201	200.2000	62.5000	29.4000
17.1000	59.5000	120.3000	293.2000	274	198.1000	119.5000	40.3000
33.7000	73.8000	165.5000	260.3000	207.7000	130.8000	69.8000	11.2000
33.8000	55.8000	93.7000	253	201.7000	178.1000	54.9000	9.6000
15.8000	37.2000	177	286.5000	251.4000	183.9000	50.6000	17.7000
62	32.7000	153.1000	225.4000	308.3000	95.4000	23	23.1000
31	45.4000	125.6000	320.5000	306	150.8000	38.4000	6.8000
61.6000	51.2000	207.2000	302.3000	228.7000	157.7000	37.5000	10
29	40.8000	211.9000	247.2000	283.4000	185.9000	108.2000	34.6000
23.1000	48.2000	191.3000	163.1000	209.9000	178.5000	71.5000	42.4000
37.9000	43.8000	107.1000	326.3000	259.2000	119.2000	58.2000	51.7000
25.7000	72.9000	214.8000	269.8000	192.6000	109.6000	68.6000	16.8000
42.8000	67.9000	157	342	239.7000	191.3000	45.5000	20.7000
33.6000	63.9000	155.1000	227.9000	226.9000	171.7000	90.5000	45.2000
35.2000	59.4000	232	265	309.7000	199.6000	139.2000	46.3000
40.1000	74	230.7000	282.7000	292.8000	278.1000	161.3000	29.1000
35.8000	103.6000	212.3000	183.8000	240.9000	111.8000	19.5000	44.7000
32.7000	59.5000	194.7000	304.6000	285.3000	163.1000	91.5000	50.1000
38.2000	52.5000	163.7000	295.7000	191.6000	123	45.9000	25.2000
43.9000	51.2000	193.9000	293.7000	274.4000	203.3000	70.5000	16.1000

Figure 4.4: Combined dataset part III

The figure 4.4 has the tabular information of rainfall dataset.

25 DEC	26 ANN	27 JanFeb	28 MarMay	29 JunSep	30 HarvestMonth	31 HarvestDuration	32 TonnHect
8.2000	1.0308e+03	73.2000	107.3000	751	6	3	47.0900
25.2000	1.0384e+03	11.6000	111.9000	797.8000	8	55	1.2434e+03
18	1.1959e+03	24.7000	107.7000	885.6000	9	0	141.5600
16.4000	1.0251e+03	24.5000	138.8000	764.3000	9	0	310.6900
10.1000	977.5000	45	131.4000	726.4000	9	1	219.0200
26.3000	1.1492e+03	71.3000	84.4000	898.9000	7	5	902.4600
12.9000	1.0348e+03	61.5000	132.1000	782.2000	7	0	154.5500
7.4000	1.0774e+03	37	84.7000	903	7	0	128.3800
27.9000	1.1285e+03	37.9000	119.4000	895.7000	8	1	553.9900
5.4000	1.1839e+03	23.8000	83.5000	928.5000	9	0	86.4700
12.1000	1.0289e+03	45.8000	114.3000	742.8000	10	5	136.3000
5.3000	1.0704e+03	41.9000	101.6000	811.8000	10	1	348.9600
23.2000	1.0618e+03	44.4000	122.2000	786.7000	9	19	714.5200
21.6000	1.1859e+03	31.9000	136.1000	930	10	4	428.3600
8.2000	1.1244e+03	57.3000	141.6000	781.5000	11	25	292.0800
2.9000	1.3248e+03	24.7000	105.6000	1.0062e+03	6	1	667.5500
9.3000	1.4639e+03	45.5000	134.5000	1.0843e+03	6	0	24.0200
15.5000	1.0202e+03	15.8000	176	748.8000	9	0	46.8300
18.2000	1.2879e+03	69	111.4000	947.7000	9	0	23.2900
3	1.0391e+03	45.2000	145.7000	774.1000	10	48	40.5500
15.3000	1225	45	112.9000	965.2000	10	5	527.9700

Figure 4.5: Combined dataset part IV

4.1.2 Results of KNN algorithm

In this phase, when we click on "Apply KNN" button a callback method is called which works to apply KNN algorithm on the combined dataset and produce the result in term of classified classes (low, mid and high) and also crop yield estimated cost in term of regression.

Figure 4.6 shows the results, which have accuracy of KNN classifier, data points belong from which class label after training the data and the actual crop yield production estimated cost.

4.1.3 Results of SVM algorithm

In this phase, when we click on "Apply SVM" button a call-back method is called which works to apply SVM algorithm on the combined dataset and produce the result in term of classified classes (low, mid and high) and also crop yield estimated cost in term of regression.

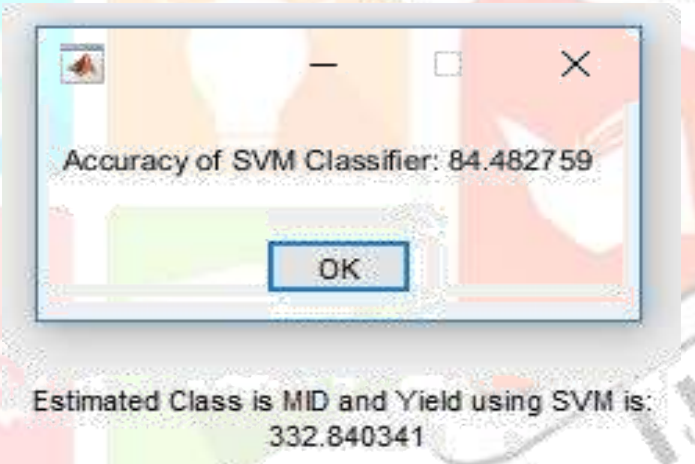

Figure 4.7: Result of SVM classifier

Figure 4.7 shows the results, which have accuracy of SVM classifier and the data points belongs from which class label (low, mid, and high) after training the data and the actual crop yield production estimated cost.

V. Conclusions

The concept of descriptive analytics in agriculture domain is a major research area. This research work provides the information about how it could apply data analytics on sugarcane crop datasets. There are three dataset named as Soil, Rainfall, Yield. These datasets includes several parameters which are helpful to know the condition of crops and classify the data into separate classes by performing supervised training on the dataset that are collected from agriculture domain. This system has the capability to perform both the classification as well as regression. In the classification step the data is classified into three classes (low, mid, and high), whereas in regression step the actual cost of yield production is estimated. This work is basically provides the comparative study of various algorithm and it shows the accuracy of each algorithms to train the datasets and also mean squared error at the cross-validation phase of the sample data. This work is domain independent. This system can be build system for other domain like as medical, product comparison, retails etc. In this, just need to pass the datasets through this system but dataset should be in consistent form.

VI. FUTURE SCOPE

This paper work can be enhancing to the next level. We can build a recommender system of agriculture production and distribution for farmer. By which farmers can make decision in which season which crop should sow so that they can get more benefit. This system is work for structured dataset. In **future, data** independent system can be implement. It means format of data

whatever, our system should work with same efficiency.

References

[1] Rajagopalan B, Lall U, A K-Nearest Neighbor simulator for daily precipitation and other weather variables.Wat Res Res35(10) : 3089–3101, 1999.
[2] Using Data Mining to Discover Patterns in Autonomic Storage Systems. Zhenmin Li, Sudarshan M. Srinivasan, Zhifeng Chen, Yuanyuan Zhou, Peter Tzvetkov, Xifeng Yan, and Jiawei Han. 1st Workshop on Algorithms and Architectures for SelfManaging Systems in conjunction with ISCA and SIGMETRICS, June 2003.
[3]Tripathi S, Srinivas VV, Nanjundiah RS Downscaling of precipitation for climate change scenarios: a Support Vector Machine approach. J Hydrol ss330:621–640, 2006.
[4] Wu X, Kumar V, Quilan JR, Ghosh J, Yang Q, Motoda H, McLanchlan GJ, Ng A, Liu B, Yu PS, Zhou Z-H, Steinbach M, Hand DJ, Steinberg D, Top 10 algorithms in data mining. KnowlInfSyst14: 1-37, 2008.
[5]Yun-lei Cai, Duo Ji ,Dong-fengCai
Natural Language Processing Research Laboratory, Shenyang Institute of Aeronautical Engineering Proceedings of NTCIR-8 Workshop Meeting, June 15–18, 2010, Tokyo, Japan
[6] https://www.wikipedia.org