APPLICATION OF MACHINE LEARNING TECHNIQUES IN VARIOUS EMERGING FIELDS

[1]Mr. Mohit Sharma [2]Ms. Harshita Sharma [3]Ms. Archana Sharma [4]Dr. B. K. Sharma

[1]Computer Science & Engg. ABESIT, Ghaziabad –UP (India)

[2]Electronics & Communication Engg. Jaypee Institute of Information Tech. Sector-62, Noida –UP

[3]Assistant Professor, Computer Science & Engg. ABESIT, Ghaziabad –UP (India)

[4]Principal Scientific Officer & Head Software Development Centre and Computer Science and Engg. Division Northern India Textile Research Association, Ghaziabad (India)

ABSTRACT:

Machine learning is the part of artificial intelligence that emphasis on the learning of the system through training on data sets and developing experience without being commutated by a mathematical expression. The machine learning algorithm can be applied if the pattern exist among the training data sets , we cannot pin down the target function mathematically, there is sufficient data for the system.

The application of machine learning techniques used in various fields like Speech Recognition, Weather Forecasting, Image Processing, Fraud Detection, Stock Market Analysis, Energy Conservation, AI, Medical Diagnosis , various records for better automation in industry related data and many more. In this paper, we have elaborated some applications of machine learning including Energy Conservation and Stock Market Analysis.

Keywords: Machine Learning, K-Mean Clustering, Naïve Bayes Classification, Neural Network, Artificial Intelligence, Boiler Efficiency, Stock Market.

INTRODUCTION:

Machine learning is the part of artificial intelligence that emphasis on the learning of the system through training on data sets and developing experience without being commutated by a mathematical expression. The machine learning algorithm can be applied if the pattern exist among the training data sets , we cannot pin down the target function mathematically, there is sufficient data for the system. The different categories of Machine learning are as listed:

SUPERVISED MACHINE LEARNING:

In this type of learning algorithm the training data sets provided to form the Learning Model for predicting the future events are labeled (i.e. provided with the output).The framework can alter learning by focusing on new contributions. It's the better model used to find errors and better yield.

UNSUPERVISED MACHINE LEARNING:

Unsupervised learning algorithms are used to prepare the learning model with ungrouped or unlabelled data points. The main essence of this learning is to portray a concealed structure from unlabeled data. It doesn't make a correct result approximation, however it helps in extracting major information and deductions about the learning model from the unlabeled data sets.

SEMI-SUPERVISED MACHINE LEARNING:

This learning algorithm falls in the middle of supervised and unsupervised learning since this makes use of both labeled and unlabelled data points. This model utilizes more of the unlabelled data than the named information.

REINFORCEMENT MACHINE LEARNING:

This learning acquires sufficient knowledge or information about the input data point and has some labeled data to project the learning model.

MACHINE LEARNING METHODS:

Decision trees: tree- shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decisions tree methods include Classification and regression trees (CART) and chi Square Automatic interaction Detection (CHAID). CART and CHAID are decisions tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2- way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

CLASSIFICATION:

Classification is a machine learning technique used to predict group membership for data instance. Classification is a learning function that maps a data item into one of several predefined classes. The various Method for Classification like Bayes Classification Theorem, Classification by Back propagation and K Nearest Neighbor Classifiers. The Naive Bayesian classifier is based on Bayes' theorem with the independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

CLUSTERING:

Clustering algorithms divide data into meaningful groups such that patterns in the same group are similar in some sense and patterns in different group are dissimilar in the same sense. Searching for clusters involves unsupervised learning. For example, the search engine clusters billions of web pages into different groups, such as news, reviews, videos, and audios. Hierarchical clustering method combines data objects into subgroups; those subgroups merge into larger and high level groups and so forth and form a hierarchy tree. Hierarchical clustering methods are of two types Agglomerative (bottom-up) and Divisive (top-down) approaches. The agglomerative clustering start with one-point clusters and recursively merges two or more of the clusters. CURE (Clustering Using Representatives) are its further extension. The divisive clustering starts with a single cluster containing all data points and recursively splits that cluster into appropriate sub clusters and SVD (Singular Value Decomposition) are its further research. (ii) Partitioning algorithms discover clusters either by iteratively relocating points between subsets or by identifying areas heavily populated with data. Its further research includes SNOB, MCLUST, *k*-medoids, and *k*-means, DBSCAN (Density Based Spatial Clustering of Applications with Noise).

ARTIFICIAL NEURAL NETWORKS:

Artificial Neural Networks are used in a wide range of applications. Some of the applications are detecting the fraudulent use of credit cards. They can be used for finding of credit risk prediction for increasing rate of targeted mailings. The Non linear predictive models that learn through training and resemble biological neural networks in structure. Optimization techniques that use processes such as genetic combination, mutation, and selection in a design based on the concepts of natural evolution.

GENETIC ALGORITHMS:

Optimization techniques that use processes such as genetic combination, mutation, and selection in a design based on the concepts of natural evolution.

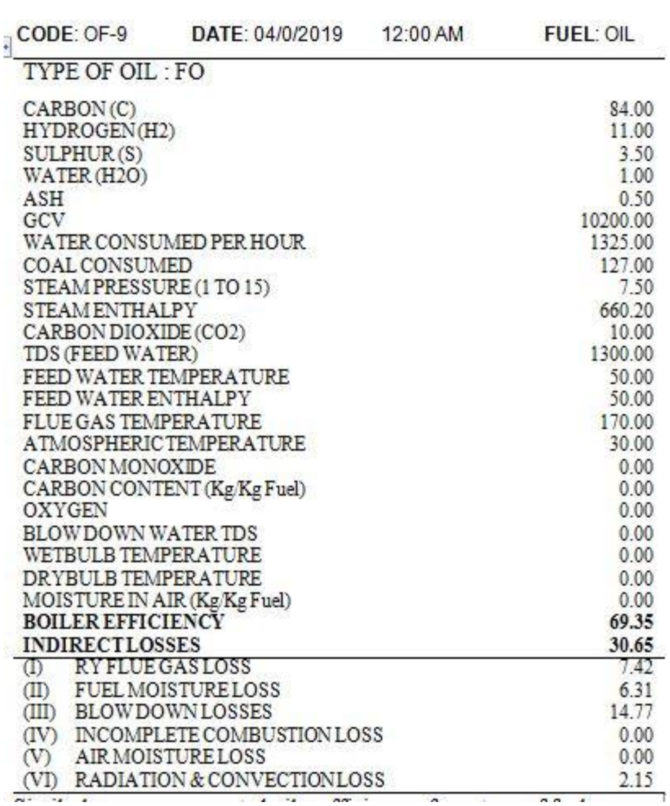
In this paper, we have elaborated some applications of machine learning techniques used in Energy Conservation and Stock Market Analysis.

ENERGY CONSERVATION:

In Industrial processing unit, major emphasis is usually laid upon productivity and quality, whereas energy conservation is considered as a second priority. However, due to the alarming increase in energy cost, every effort should be given to minimize it. As we are aware that Boiler is main source of fuel consumption in any industry. It is a matter of great concern to that entire boiler should run at its maximum efficiency with minimum indirect losses. For attaining maximum boiler efficiency, exact assessment of boiler efficiency and all indirect losses are very important. Boiler consumes measure chunk of fuel in any industry. Efficiency of any Boiler depends upon minimization of various indirect losses of the boiler so that amount of energy input in the boiler by burning the fuel can be maximum utilized for generation of steam and cost of steam can be minimized ultimately. The direct efficiency of the boiler is based on fuel consumption and steam generation for a particular time period as per standards. The following indirect losses can be minimized for efficient boiler efficiency.

- Dry Flue Gas Loss
- Fuel Moisture Loss
- Blow Down Losses
- Incomplete Combustion Loss
- Air Moisture Loss
- Radiation and Convection Loss

After knowing the various heat losses it is possible to take action to improve boiler efficiency. A model report format of the boiler efficiency is shown in Table-1 where all the input details can be fed to computer like fuel analysis, calorific value, and steam pressure, enthalpy, %CO_2, TDS etc. Subsequently the boiler efficiency and the indirect losses are calculated and displayed. Similarly, we can generate boiler efficiency report of any type of fuel like coal, husk, etc. Graphical representation of Boiler efficiency can be displayed as shown in Fig.1.



CODE: OF-9 DATE: 04/0/2019 12:00 AM FUEL: OIL

TYPE OF OIL : FO

CARBON (C)	84.00
HYDROGEN (H2)	11.00
SULPHUR (S)	3.50
WATER (H2O)	1.00
ASH	0.50
GCV	10200.00
WATER CONSUMED PER HOUR	1325.00
COAL CONSUMED	127.00
STEAM PRESSURE (1 TO 15)	7.50
STEAM ENTHALPY	660.20
CARBON DIOXIDE (CO2)	10.00
TDS (FEED WATER)	1300.00
FEED WATER TEMPERATURE	50.00
FEED WATER ENTHALPY	50.00
FLUE GAS TEMPERATURE	170.00
ATMOSPHERIC TEMPERATURE	30.00
CARBON MONOXIDE	0.00
CARBON CONTENT (Kg/Kg Fuel)	0.00
OXYGEN	0.00
BLOW DOWN WATER TDS	0.00
WETBULB TEMPERATURE	0.00
DRYBULB TEMPERATURE	0.00
MOISTURE IN AIR (Kg/Kg Fuel)	0.00
BOILER EFFICIENCY	**69.35**
INDIRECT LOSSES	**30.65**
(I) RY FLUE GAS LOSS	7.42
(II) FUEL MOISTURE LOSS	6.31
(III) BLOW DOWN LOSSES	14.77
(IV) INCOMPLETE COMBUSTION LOSS	0.00
(V) AIR MOISTURE LOSS	0.00
(VI) RADIATION & CONVECTION LOSS	2.15

Table 1 : Model Report for Boiler Efficiency & Indirect Losses System

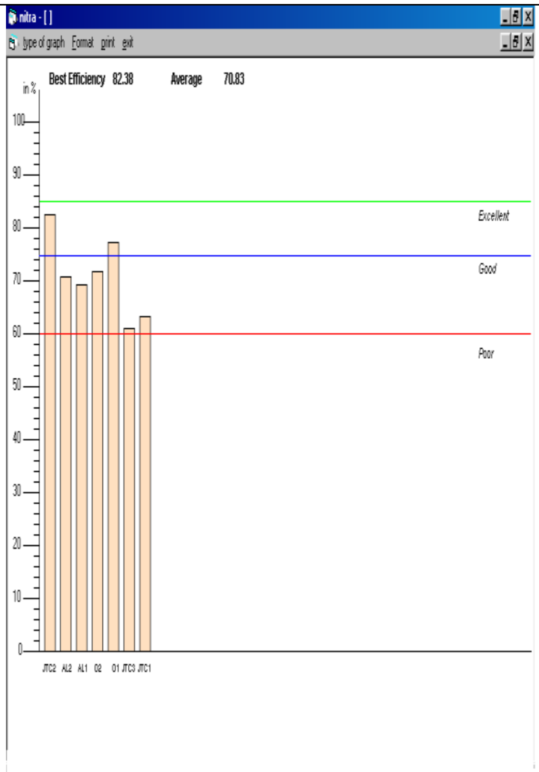


Fig 1 : Graphical Presentation of Boiler Efficiency

NAÏVE BAYES CLASSIFICATION TECHNIQUES IN BOILER EFFICIENCY:

In a industrial process house, the utility department collect the data for fifteen sample. The details of boiler efficiency and its performance are provided in Table-2.

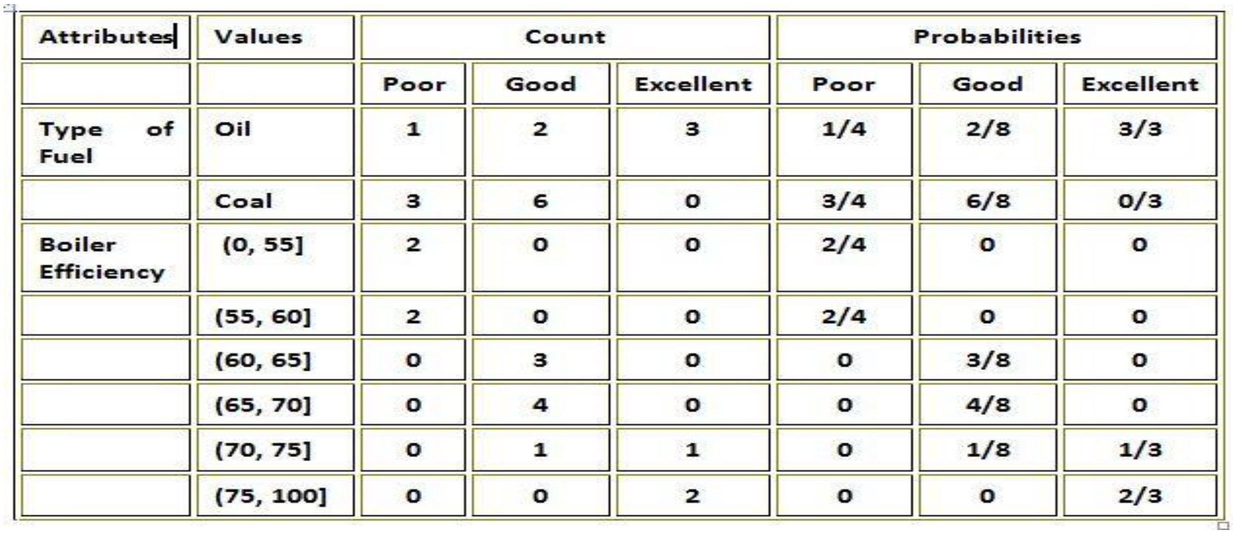
Code ID	Type of Fuel	Boiler Efficiency in %	Performance of Boiler Efficiency
A1	Coal	55	Poor
A2	Oil	75	Excellent
A3	Coal	67	Good
A4	Coal	70	Good
A5	Coal	59	Poor
A6	Oil	68	Good
A7	Coal	55	Poor
A8	Oil	59	Poor
A9	Oil	80	Excellent
A10	Oil	78	Excellent
A11	Coal	65	Good
A12	Oil	74	Good
A13	Coal	67	Good
A14	Coal	65	Good
A15	Coal	62	Good

Using the performance of boiler efficiency classification results for table, there are four records classified as poor, eight as good and three as excellent. We divide the boiler efficiency attribute into six ranges:

(0, 55] , (55, 60] , (60, 65] , (65, 70] , (70, 75] , (75, 100]

We estimate the prior probability P (h) of boiler efficiency:

P (Poor) = 4/15 = 0.267
P (Good) = 8/15 = 0.533
P (Excellent) = 3/15 = 0.200



Attributes	Values	Count			Probabilities		
		Poor	Good	Excellent	Poor	Good	Excellent
Type of Fuel	Oil	1	2	3	1/4	2/8	3/3
	Coal	3	6	0	3/4	6/8	0/3
Boiler Efficiency	(0, 55]	2	0	0	2/4	0	0
	(55, 60]	2	0	0	2/4	0	0
	(60, 65]	0	3	0	0	3/8	0
	(65, 70]	0	4	0	0	4/8	0
	(70, 75]	0	1	1	0	1/8	1/3
	(75, 100]	0	0	2	0	0	2/3

We use these values to classify new boiler efficiency. For example, suppose we wish to classify
t = {A20, Oil , 74% }.

Using these values and the associated probabilities of fuel and boiler efficiency, we obtain the following estimate i.e, conditional probability P (Xi | h):

P {t | poor} = 1/4 * 0 = 0.00
P {t | Good} = 2/8 * 1/8 = 0.031
P {t | Excellent} = 3/3 * 1/3 = 0.333

Combining these P (Xi | h) * P (h)

Poor = 0 * 0.267 = 0.00
Good = 0.031 * 0.533 = 0.0166
Excellent = 0.333* 0.2 = 0.066

We estimate P (xi) = 0 + 0.0166 + 0.066 = 0.0826

Finally, we obtain the posterior probability P (h1 | xi)
P (Poor | t) = 0 / 0.0826 = 0.00

P (Good| t) = 0.0166 / 0.0826 = 0.20
P (Excellent | t) = 0.066 / 0.0826 = 0.799

Therefore, based on these probabilities, we classify the Boiler Efficiency as Excellent because it has the highest probability.

STOCK MARKET ANALYSIS

The stock market is a backbone of fast emerging economies in the world. It is considered too uncertain to be predictable due to various factors such as company's economic growth, investments, company and country's strategically plans etc. Stock market forecasting includes uncovering market trends, planning investment strategies, identifying the best time to purchase the stocks and what stocks to purchase. Financial institutions produce huge data sets that build a foundation for approaching these enormously complex and dynamic problems with various techniques. The stock market data can be effectively analyzed using various statistical methods and machine learning algorithms. Introducing Machine Learning algorithm into stock market prediction processes can achieve a substantial increase in growth of stock market and demand of investors in the market. It is here that Machine Learning algorithm plays a vital role.

CONSISTENCY BETWEEN SHARE PRICES:

The consistency between shares prices can be evaluated using statistical coefficient of variation method. The comparison of dispersion for such dataset can be made by calculating coefficient of variation. Greater the coefficient of variation higher is the value of the standard deviation relative to the mean. The lesser value of coefficient of variation more consistent as compare to others.

ANALYSIS OF J48 AND NAÏVE BAYES CLASSIFIERS ON STOCK DATASET AND DETERMINE THE ACCURACY:

J48 is an algorithm used to generate a decision tree which is C4.5 and can be used for classification. The additional features of J48 are accounting for missing values, decision tree pruning, continuous attribute value range, derivation of rules etc. The share holder in most of the cases invests on profit. The detail of sample dataset of a particular company has been given in Table-3.

Outlook	Purchasing Price	Selling Price	Company Growth	Nation Policy effect on stock market	Investment
Buy	82	82	Down	Good	No
Buy	80	92	Up	Good	Yes
Sell	83	89	Down	Good	Yes
Hold	70	96	Down	Bad	No
Hold	68	80	Down	Good	Yes
Hold	65	70	Up	Good	Yes
Sell	64	65	UP	Good	No
Buy	72	95	Up	Good	Yes
Buy	69	70	Down	Good	Yes
Hold	75	80	Down	Good	Yes
Buy	75	70	Up	Good	Yes
Sell	72	95	Up	Bad	Yes
Sell	81	75	Down	Bad	Yes
Hold	71	91	Up	Good	No

OUTPUT RUN INFORMATION BY J48 ON WEKA SOFTWARE

```
Scheme:        weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:      stock_market
Instances:     14
Attributes:    6
               outlook
               purchasing_price
               selling_price
               company_growth
               nation_policy_stock
               invest
Test mode:   evaluate on training data and Classifier model (full training set)
J48 pruned tree : yes (14.0/5.0)
Number of Leaves  :     1
Size of the tree     :     1

Summary
Correctly Classified Instances        9            64.2857 %
Incorrectly Classified Instances      5            35.7143 %
Kappa statistic                        0
Mean absolute error                    0.4592
Root mean squared error               0.4792
Relative absolute error               98.9011 %
Root relative squared error           99.9306 %
Total Number of Instances             14

Detailed Accuracy By Class
                 TP Rate  FP Rate  Precision      Recall   F-Measure  MCC     ROC Area  PRC Area  Class
                  1.000    1.000    0.643          1.000    0.783      ?       0.500     0.643      yes
                  0.000    0.000     ?             0.000     ?         ?       0.500     0.357      No
Weighted Avg.0.643    0.643     ?             0.643     ?         ?       0.500     0.541
Confusion Matrix
 a b   <-- classified as
 9 0 | a = yes
 5 0 | b = no
```

The Naive Bayesian classifier is based on Bayes' theorem with the independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods. The above mentioned dataset given in the table has been applied for Naïve bayes classifier on the weka software.

Output Run information by Naive Bayes Classifier on weka software

```
Scheme         :   weka.classifiers.bayes.NaiveBayes
Relation       :   stock_market
Instances      :  14
Attributes     :   6
                  outlook
                  purchasing_price
                  selling_price
                  company_growth
                  nation_policy_stock
                  invest
Test mode      :  evaluate on training data and Classifier model (full training set)
                     Class
Attribute          yes     no
```

```
                 (0.63)  (0.38)
===============================
Outlook
  buy              4.0    3.0
  sell             4.0    2.0
  hold             4.0    3.0
  [total]         12.0    8.0

purchasing_price
  mean            73.3131 73.2364
  std. dev.        5.4689  6.3323
  weight sum           9       5
  precision        1.7273  1.7273

selling_price
  mean            80.3704  85.4222
  std. dev.       10.6475  10.9791
  weight sum           9       5
  precision        3.4444   3.4444

company_growth
  up               5.0    4.0
  down             6.0    3.0
  [total]         11.0    7.0

nation_policy_stock
  good             8.0    5.0
  bad              3.0    2.0
  [total]         11.0    7.0
Summary
Correctly Classified Instances     9              64.2857 %
Incorrectly Classified Instances   5              35.7143 %
Kappa statistic                      0.1026
Mean absolute error                  0.4231
Root mean squared error              0.4506
Relative absolute error             91.1245 %
Root relative squared error         93.9810 %
Total Number of Instances              14

Detailed Accuracy By Class

                 TP Rate  FP Rate    Precision  Recall   F-Measure   MCC     ROC Area  PRC Area  Class
                 0.889    0.800        0.667    0.889    0.762       0.122   0.689     0.821     yes
                 0.200    0.111        0.500    0.200    0.286       0.122   0.689     0.551     no
Weighted Avg.    0.643    0.554        0.607    0.643    0.592       0.122   0.689     0.725

Confusion Matrix
 a b : classified as
 8 1 | a = yes
 4 1 | b = no
```

CONCLUSION

In the present scenario, the energy cost is the highest among all other cost component in the production of industrial goods. So far there have not been much efforts to evolve a systematic approach to analyze and monitor energy consumption in each process and translate into financial terms. The proposed case study for analysis of boiler efficiency and indirect losses system are a step forward towards those objectives. By proper use of this machine learning techniques it is possible to analyze k-mean cluster wise boiler efficiency and also consistency of boiler efficiency can also be maintained within a cluster. The Naive Bayes classification approach towards estimation of performance of boiler efficiency yield agreeable results for industrial purposes. In order to compete with international products, there is no other alternative but to go for automation in near future. This approach may act as a precursor to that.

The proposed case study of machine learning techniques for analysis of stock market is a step forward towards those objectives. By proper use of this statistical and machine learning technique it is possible to analyze the consistency of two share price and important decision making regarding investment of shares on Stock Dataset and determine its Accuracy.

REFERENCES

1. R. Agrawal, H. Mannila, R. Srikant, H. Toivonen and A. Verkamo, Fast Discovery of Association Rules, in Advances in Knowledge and Data Mining, U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy Eds, MIT Press, 1996.
2. H. Toivonen. Sampling large databases for association rules. Proc. of the Int'l Conf. on Very Large Data Bases (VLDB), 1996.
3. R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Proceedings of the 20th VLDB Conference, Santiago, Chile, 1994.
4. R. Agrawal, T. Imilienski, and A. Swami. Mining Association Rules between Sets of Items in Large Databases, Proc. Of the ACM SIGMOD int'l Conf. on Management of Data, pages 207-216, May 1993.
5. R. Agrawal, T. Imilienski, and A. Swami. Data base Mining: A Performance Perspective, IEEE Transactions on Knowledge and Data Engineering, 5(6):914-925, December 1993.
6. Patricia L. Carbone, Expanding the Meaning of and Applications for Data Mining 2000 IEEE
7. Dr. B.K. Sharma, D.K. Sharma, Application of Information Technology in Textile Wet Processing for Strategic Decision Making, International Journal of Management and System, Australia
8. Prof. S.M. Ishtiaque, Dr. B.K. Sharma, D.K. Sharma, Management Information System in Textile Wet Processing for batter decision-making, International Conference on IT in Textile Sector organized by Ministry of Textile, Govt. of India
9. Dr. B.K. Sharma & Dr. A. Das, Application of information Technology to manage the textile quality control text data, The Textile Industry and Trade Journal, Vol. No. 44, No. 5-6, May-June 2006
10. Soler, S. and D. Yankelevich (2001). "Quality Mining: A data Mining Based Method for Data Quality Evaluation", Processing of the Sixth international Conference on Data Quality, MIT.
11. Dongsong Zhang and Lina Zhou Discovering Goden Nuggets: Data Mining in Financial Application 2004 IEEE.
12. Dr. B.K. Sharma, Prof. S.K.B & Abhay Bansal, Data Mining Tools and Techniques in Textile Industry for Effective Decision Making and Corrective Action, Asian Textile Journal, Vol No. 15, No. 8, August 2006
13. C. Brunk, J. Kelly, and R. Kohavi, "MineSet: An Integrated System for Data Access, Visual Data Mining, and Analytical Data Mining," Proceedings of the Third Conference on Knowledge Discovery and Data Mining (KDD-97), Newport Beach, CA, August 1997.
14. G.K. Gupta, A book on introduction to Data Mining with Case Studies.
15. Prof. S.K. Tyagi & Dr. B.K. Sharma "Evaluation of energy efficiency, monitoring and improvement through data mining tools and techniques for oil conservation in textile industry, International Journal of computer sciences software engineering and electrical communication.
16. D.K. Bhattacharya, Dr. B.K. Sharma & Sanjeev Saxena "Energy balance & accounting for boiler and process house in textile industry through software approach at 43rd Joint technology conference, IIT Delhi on 2-3rd march 2002.
17. Sanjeev Saxena & Dr. B.K. Sharma " Evaluation of Energy Efficiency, Monitoring and improvement through software for oil conservation at 5th international petroleum conference (PETROTACH-2003) organized by ministry of petroleum and natural gas on 9-12 january,2003at vigan bhawan, New Delhi
18. B.B. Agarwal and S.P. Tayal , A book on Data Mining and Data Warehousing.
19. Dr. Seema Gupta and Shweta Nanda "Machine Learning Techniques with Image Classification, IPEM Journal of Computer Application and Research, Vol.3, December 2018.
20. Harshita Sharma, Mohit Sharma, Dr. Bhisham Kapoor, Dr. B.K. Sharma "An Analysis on Stock Market Predication using Data Mining Techniques for Effective Decision Making , IPEM Journal of Computer Application and Research, Vol.3, December 2018.
21. Harshita Sharma, Mohit Sharma, Dr. Bhisham Kapoor, Dr. B.K. Sharma "Evaluation of Boiler Efficiency through Machine Learning Techniques, 21st Annual Conference on Modeling, Optimization and Computing for Technological and Sustainable Development, SRM University, Modi Nagar, 26th – 28th April 2019.