IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Intelligent Sales Prediction using Machine Learning

Nikita Lemos Department of Information Technology Xavier Institute of Engineering
Mumbai, India nikita.l@xavier.ac.in

Ismail Pawaskar Department of Information Technology Xavier Institute of Engineering
Mumbai, India ismailpawaskar55@gma il.com

Deepak Ramchandani Department of Information Technology Xavier Institute of Engineering
Mumbai, India deepak.ramchandani8@ gmail.com

Taman Poojary Department of Information Technology Xavier Institute of Engineering
Mumbai, India tamanpoojary@gmail.com

Abstract*—Intelligent Decision Analytical System requires combination of decision exploration and prediction. The greater part of the business associations intensely rely upon an information base also, request expectation of deals patterns. The precision in deals figure gives a major effect in business. Pattern Analytical System and Data Mining techniques are powerful tools that helps in removing covered up information from a huge dataset to upgrade exactness and productivity of estimating. Traditional prediction frameworks are hard to manage the enormous information and exactness of sales measure. Valuable metals like diamonds and gold are sought after because of their financial prizes. In this manner, different methods are commonly utilized to estimate costs of jewels and valuable metals with the point of quick and precise outcomes. The costs vacillate day by day making it hard to predict the next future worth. Subsequently, by analyzing the example of past costs we can apply relapse models for future expectation. This paper targets forecasting the future costs of valuable metals like diamond, utilizing various machine learning techniques, intending to get the most precise consequence of all. Group models are utilized for expanding the exactness of costs.

Keywords*—Machine Learning, Prediction, Data Mining, Analytics, Estimation, Pattern.

INTRODUCTION

One of the significant destinations of this examination work is to discover the dependable deals pattern forecast instrument which is executed by utilizing data mining analytics procedures to accomplish the most ideal income. The present business handles gigantic storehouse of information. The volume of information is relied upon to become further in an exponential way. The measures are compulsory so as to oblige process speed of exchange and to upgrade the normal development in information volume and client conduct. For a considerable length of time, human advancements all around the globe have desired precious stones for their tasteful magnificence and charming sparkle [1]. Viewed as an image for riches, they are estimated for the well-to-do by the diamond ventures that overwhelm the market. While these precious stone makers and advertisers have been amazingly effective before, with the beginning of the data age, purchasers are presently capable to cross- reference costs of comparable precious stones from other organizations before settling on buying choices. Laboratory assessed quality can be promptly decided for diamonds also, encouraging the order of the characteristics of a precious stone [2]. With these assets comes the assignment of assessing the cost of cut precious stones given their properties. There are many factors that might affect the price of a diamond, but the most common ones are referred to as the 4 Cs: carat, cut, color, and clarity.

***Carat*:** Carat is the mass of the diamond. 1 carat (ct) is equivalent to 200mg. This is the main quantitative proportion of the 4 Cs. Carat is non-straight identified with the cost.

***Cut*:** Cut alludes to both the state of the stone and the nature of its glitter. The cut flawlessness is grouped from "Fair" to "Ideal".

***Color*:** Diamond hues differ from dreary to a light yellow. The more vapid a precious stone is, the more costly it is probably going to be. The standard is a characterization created by the Gemmological Foundation of America and is the most utilized out of the entirety of the shading reviewing plans; it utilizes an in sequential order score, "D" being the most vapid and "Z" being a prominent yellow.

***Clarity*:** Diamonds may have inside imperfections and breaks which decline their straightforwardness, which thus diminishes their worth. Lucidity is evaluated on a scale from FL (Flawless) to I3 (Obvious Inclusions) in light of the size, nature, position, and amount of inward imperfections.

There are a few different properties that may influence the cost of a precious stone. While these probably won't be as well-known as the ordinarily utilized 4 Cs, they despite everything may quantify a few variables of the precious stone that could influence the quality and thusly the cost of the jewels.

***Depth*:** Depth is estimated as the proportion among z and the normal width of the head of the diamond.

***Table*:** Table is estimated as the width of the head of the diamond at its most stretched out point.

The dataset contains information on prices of diamonds, as well as various attributes of diamonds, some of which are known to influence their price: the 4 Cs (carat, cut, color, and clarity) , as well as some physical measurements (depth, table, price, x, y, and z) [6]. The figure below shows what these measurements represent.

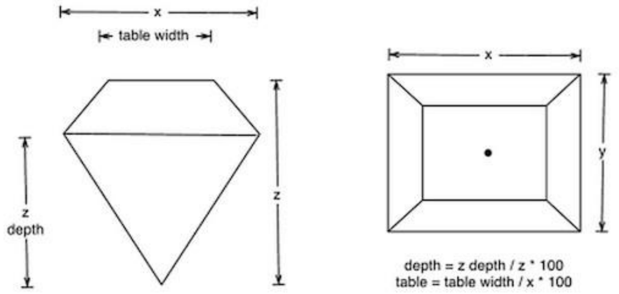


Fig.1: Exploring dimensions of Diamond

The carat is estimated in carats, the cut is estimated subjectively as an evaluation from "Fair", "Good", "Very Good", "Premium", and "Ideal", the shading is estimated with grades from "J" being the most exceedingly awful to "D" being the best, the clarity is estimated utilizing the standard from "I1", "SI2", "SI1", "VS2", "VS1", "VVS2", "VVS1", and "IF", the depth is measured as the proportion of the diamond’s z axis to its normal distance across, the table is estimated as the width of the diamond at its broadest point, the directions are estimated in millimetres, and the price is estimated in US dollars [7].

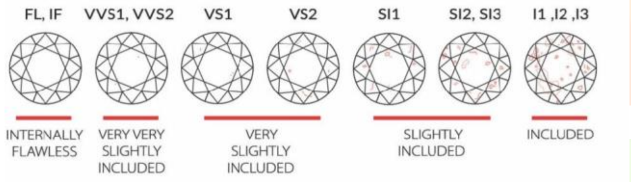


Fig.2: Diamond Clarity Chart

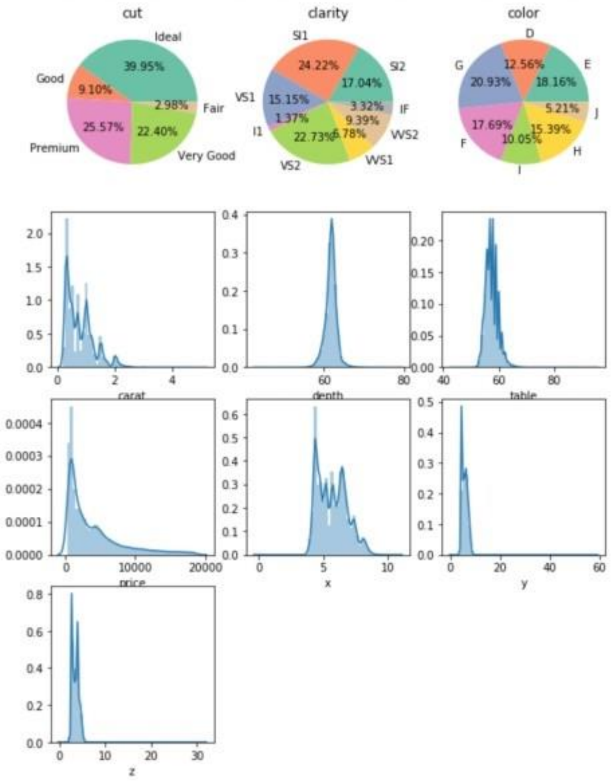


Fig.3: Distributions of data on different properties

LITERATURE SURVEY

Exact forecasts permit the association to improve advertise development with more elevated level of income age. Information mining strategies are exceptionally powerful in tuning enormous volume of information into helpful data for cost expectation and deals estimate, it is the fundamental of sound planning. Exception location this procedure plays out every fundamental datum pre-handling and model streamlining. Exception recognition procedure can be utilized to convey the model or as a beginning point for additional enhancements and supportive in demonstrating conventional data which is free of the models. The oversupply of rough stones and the increasingly strained finances of middlemen have hit miners' balance sheets in recent months as they try to manage the surplus and increase the value of existing stones[1].

Gradient Boosted Trees Gradient boosting is a machine learning technique for regression and classification problem. This approach could ensemble learning method that combines large number of decision trees to produce final prediction model. This model is built on a principle that a collection of weak learners combined together can produce a strong learner by using boosting process. Traditional estimating strategies face difficulties in delivering precise deals information for new items and buyer situated merchandise. [2].

The information input squares are demonstrated by the "Recover" administrator. The "Recover" administrator loads a Rapid Miner object into the information stream process. In the particular case it permit to choose information put away in the nearby store separated by the csv trial dataset. So as to structure a practical preparing dataset, a rationale association must applied on each of the three datasets, acquiring a solitary dataset to process [3].

This exploration paper is with respect to the business forecast dependent on current apparel style. In current society being chic is a pattern. There are different elements impact attire and style deals, including value, kind of material, and size, along with the conventional variables like season and material. This exploration paper utilized choice tree and ID3 calculation to lead the expectation examination with respect to the dress deals [4].

Strong expectations profit by having high caliber and effectively open information. Contingent upon the result of the organization, various types of outside information could be utilized. Along these lines, it is significant not exclusively to make forecasts dependent on the numbers close by yet in addition to match these numbers with subjective data so as to get an increasingly practical perspective on the business . This can be accomplished with fitting correspondence and cooperation between the business and the group associated with the development of the estimating model [5].

In this paper, it is actualized a framework to assess the nature of system assets, which embraces four AI calculations to figure the bundle scores. The entire model contains numerous segments: information preprocessing, model preparing and expectation utilizing the calculations like Random Forest KNN and Logistic Regression. We looked at the presentation of four expectation strategies on a dataset with 100 normal programming bundles. It shows that these strategies have a high precision, which is from 82.84% to 90.52%.[6].

In this paper, we propose a period discovery calculation coordinated pattern forecast. The expectation ability is accomplished by learning automata. Another pattern expectation and period discovery calculation is proposed in this paper. As far as we could possibly know, this is the main calculation that can naturally discover the time of information, which is additionally used to anticipate its future pattern [7].

So as to adapt to the intricacy of equipment plan streamlining, we address the issue by beginning from the estimation of one of the objective highlights that is the territory of a equipment segment. To accomplish this objective, we resort to various Machine Learning calculations to play out the expectation by utilizing calculations like Random Forest and Gradient Lift.

Three ML calculations have been considered to foresee the zone of a RI part [8].

Particularly in clinical field, where those strategies are broadly utilized in determination and examination to make choices by utilizing Support Vector Machines, Naive Bayes and KNN. In this investigation, we utilized four fundamental calculations: SVM, NB, k-NN and C4.5 on the Wisconsin Breast Cancer (unique) datasets [9].

With the advancement of large information examination innovation, more consideration has been paid to sickness forecast from the point of view of large information investigation, different explores have been led by choosing the qualities naturally from an enormous number of information to improve the precision of hazard grouping, instead of the recently chose qualities [10].

Hence, we can reach the inference that the precision of CNN-UDRP (T-information) and CNNMDRP (S&T-information) calculations have little distinction however the review of CNN-MDRP (S&Tdata) calculation is higher and its union speed is quicker. In synopsis, the presentation of CNN-MDRP (S&T-information) is better than CNNUDRP (T-information) [10].

Machine Learning procedures can be applied to all disciplines. AI utilizes insights to explain numerous characterization and bunching issues. The ML calculations are characterized in three classifications . They are managed, unaided and semi directed. In this paper we examined around three AI calculations which can be applied to forecast, as Generalized Linear Model (GLM), Decision Tree (DT) and Gradient Boost Tree (GBT) [11].

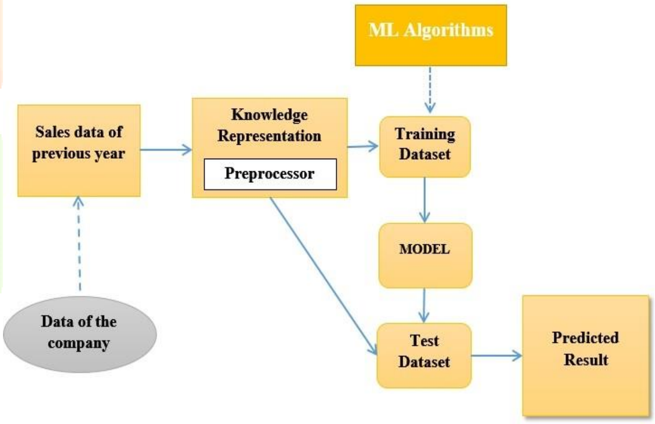


Fig.4: System Architecture

To forecast the electric utilization, chronicled information on the past utilization, temperature, moistness, total national output (GDP), populace, kinds of families and their connection coefficients, and client practices are regularly used. The conduct projection is known to utilize the information of individuals' propensities in utilizing electrical gadgets (which isn't connected with the ecological conditions and the climate) for expectation. [12].

GOAL AND SCOPE DEFINITION

The initial step requires characterizing the objective and extent of the appraisal, including the kind **of functional result** considered as a source of perspective premise and **cost segments.**

The **functional result** (normalized unit of capacity conveyance) viable is the 'change of one harsh into at least one cleaned precious stones with maximal cost'. The cost of a precious stone relies upon a complex connection of various boundaries, known in the business as the 'four C's': shading (when in doubt a white precious stone is more significant than a jewel that is increasingly yellow), lucidity (subject to the quantity of material deformities, assessed by a clearness reviewing scale), cut (which mirrors the evenness, extents and clean of a jewel) and carat (the stone's weight communicated in carats, for example units of 200 mg). Since jewels are expended not for their inherent utility however for the impression they make on others, precious stone estimating shows irregularities, for example, value premiums of 25% that clients are happy to pay for a 0.50ct precious stone over a 0.49ct jewel [13].

The **cost segments** thought about mirror the money related assets expended so as to understand a utilitarian outcome in the current situation and in the GIP situation. These expenses are, for the current situation, from one perspective the cost paid to subcontractors for cleaning in India or China, that are communicated in US$ per carat of harsh jewel, and then again the expenses of shipping the jewels to and fro to the subcontractor, that are communicated in US$ per 1000$ of worth that is moved.

PROPOSED MODEL

This model is used to pre-process the data like removing the redundant attributes or records, data cleansing, filtering the dataset, etc. Numerical variables can be classified as continuous or discrete based on whether or not the variable can take on an infinite number of values or only non-negative whole numbers, respectively. If the variable is categorical, we can determine if it is ordinal based on whether or not the levels have a natural ordering [11].

Data Cleansing or information cleaning is the way toward recognizing and adjusting (or expelling) degenerate or erroneous records from a record set, table, or database and alludes to distinguishing fragmented, mistaken, off base or superfluous pieces of the information and afterward supplanting, changing, or erasing the messy or coarse information. Information purging might be performed intelligently with information fighting devices, or as group preparing through scripting [12].

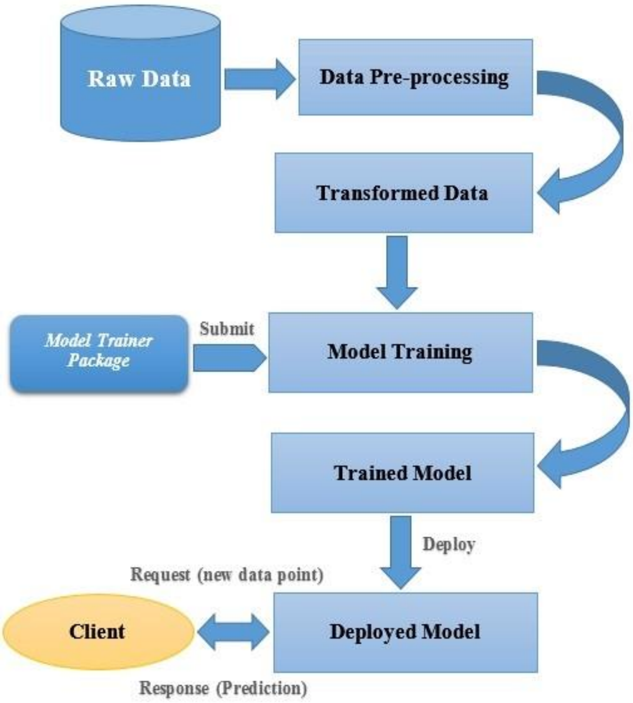


Fig.5: Data Pre-processing Flow

Statistical Data Analysis for the most part includes some type of measurable apparatuses, which a layman can't perform without having any measurable information. There are different programming bundles to perform factual information examination. This product incorporates Statistical Analysis System (SAS), Statistical Package for the Social Sciences (SPSS), Stat delicate, and so forth. After information preprocessing, so as to obviously comprehend the idea of our information, an exploratory investigation was conducted [13]. Outlier Detection procedure plays out every essential datum preprocessing also, model advancement.

Our objective was to foresee the cost of diamonds utilizing highlights, for example, carat, clarity, color, cut, depth, table length, x, y, and z axis lengths in millimeters. The depth, table length, x, y, and z pivot lengths were given as numerical information. Along these lines, we standardized the information by taking away the mean from every information point and isolating by the standard deviation.

Machine Learning procedures can be applied to all disciplines. AI utilizes insights to settle numerous characterization and bunching issues. The ML calculations are characterized in three classifications. They are regulated, unaided and semi regulated. In this paper we examined around three AI calculations which can be applied to forecast.

Fig.7: Histogram of clarity vs price

Fig.8: Histogram of clarity vs carat

The carat, clarity, color, what's more, cut were given as absolute information so we changed over it to an element vector by planning them to a consecutive number run with zero mean. For instance, the cut element of a diamond. is spoken to as Fair, Good, Very Good, Premium, or on the other ful while be spok to as po

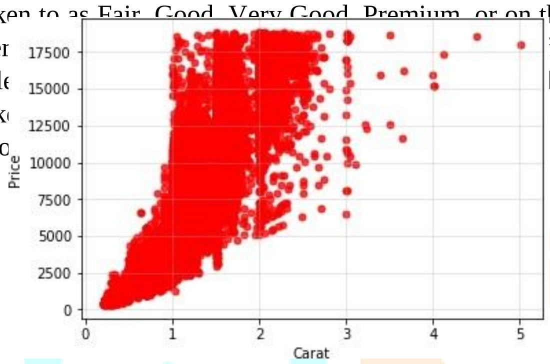
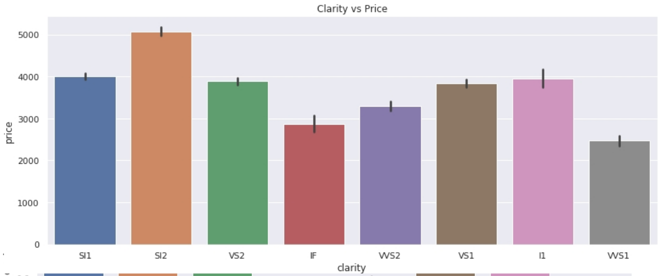


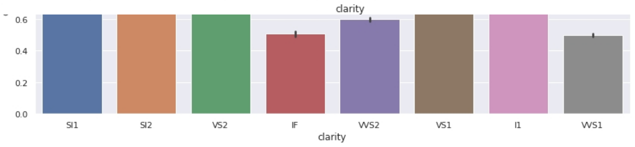
Fig.6: Scatterplot of Carat vs Price

RESERCH METHODOLOGIES

1. Gradient Boost

Gradient boosting is a Machine Learning strategy for relapse and order issues, which delivers an expectation model as a gathering of powerless forecast models, ordinarily choice trees. It assembles the model in a phase savvy design like other boosting strategies do, and it sums them up by permitting improvement of a discretionary differentiable misfortune work.





$$MSE=1/2(y-F(x))^2$$

$$F0(X) = \arg\min \sum_{i=1}^{n} L(yi, \Upsilon)$$

$$F0(X) = \arg\min \sum_{i=1}^{n} L\,(yi,F(m-1)(xi) + \Upsilon hm(xi))$$

$$Fm(x) = Fm\,-1(x) + \Upsilon m\ hm(x)$$

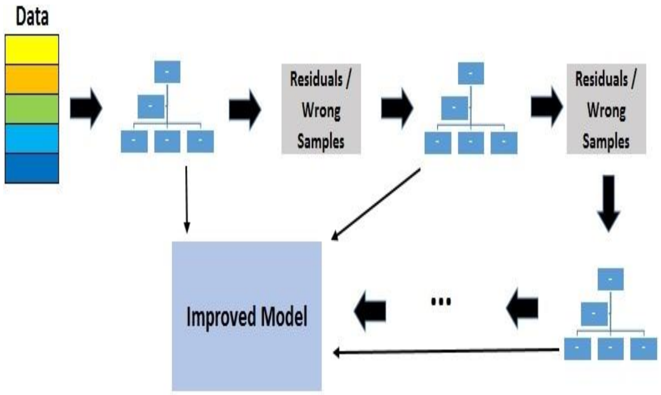


Fig.9: Gradient Boost Model

2. Naive Bayes Classifier

In Statistics and Machine Learning, Naïve Bayes classifiers are a group of basic "probabilistic classifiers" in light of applying Bayes' hypothesis with solid (credulous) autonomy suppositions between the highlights. They are among the least complex Bayesian system models. However, they could be combined with Kernel thickness estimation and accomplish higher precision levels.
Naive Bayes classifiers are exceptionally adaptable, requiring various boundaries straight in the quantity of factors (highlights/indicators) in a learning issue.

$$X = (x1,\ldots..,xn)$$

$$P\,(Ck|x1,\ldots.xn)$$

Where p is instant probablities for K outcomes

$$P\,(Ck|x) = p\,(Ck)\ p\,(x|Ck)/px$$

$$\text{Posterior} = (\text{prior} * \text{Likelihood})\ /\text{evidence}$$

$$Z = p(x) = \Sigma p(Ck)p(x|Ck)$$

3. Polynomial Regression

In Statistics and Machine Learning, polynomial regression is a type of regression wherein the connection between the a independent variable x and the dependent variable y is displayed as a furthest limit polynomial in x. Polynomial relapse fits a nonlinear connection between the estimation of x and the comparing contingent mean of y, indicated E(y |x). Albeit polynomial relapse fits a nonlinear model to the information, as a factual estimation issue it is straight, as in the relapse work E(y | x) is direct in the obscure boundaries that are assessed from the information.

$$Y = \beta 0 + \beta 1x1 + \beta 2x^2 + \varepsilon$$

$$Y = \beta 0 + \beta 1x + \beta 2x^2 + \beta 3x^3 + \beta nx^n + \varepsilon$$

4. KNN (K Nearest Neighbors)

In design acknowledgment, the k-closest neighbor's calculation (k-NN) is a non-parametric strategy proposed by Thomas Cover utilized for grouping and regression. In the two cases, the info comprises of the k nearest preparing models in the element space. The yield relies upon whether k-NN is utilized for arrangement or relapse:
In k-NN arrangement, the yield is a class participation. An item is characterized by a majority vote of its neighbors, with the article being doled out to the class generally normal among its k closest neighbors (k is a positive whole number, regularly little). In the event that k = 1, at that point the item is just doled out to the class of that solitary closest neighbor.
In k-NN relapse, the yield is the property estimation for the article. This worth is the normal of the estimations of k closest neighbors.

The optimal Weighting Scheme is given by

$$W*ni = 1/k*[1+ d/2 - d/2k*^2/d\{i^1+2/d - (i-1)^1+2/d\}]$$
$$\text{for } i = 1,2,\ldots..k*$$

$$W*ni = 0 \text{ for } i=k* + 1,\ldots., n.$$

$$\text{Set } k* = [Bn^4/d+4]$$

$$RR\,(Cn^knn) - RR(C^Bayes) = \{B1\ 1/k + B2\ (k/n)\ ^4/d\}\ \{1+o(n)\}$$

Where B1 and B2 are constants.

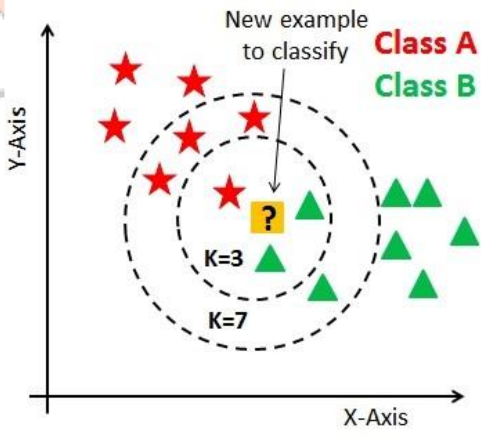


Fig.10: K-Nearest Neighbors

FUTURE SCOPE AND CONCLUSION

The historical backdrop of man-made reasoning gives us that there has been a slow and trans-formative improvement inside the extraordinary parts of computational sciences that underlie the advances common in AI. A large portion of the innovations comprise of techniques characterized by so-called computational insight, including neural systems, developmental calculations and fluffy frameworks. In spite of the fact that it might appear to be unavoidable that such a ground-breaking business instrument will be embraced as once huge mob, actually more nuanced than that.

AI model expectations permit organizations to make profoundly exact conjectures regarding the presumable results of an inquiry dependent on authentic information, which can be pretty much a wide range of things – client agitate probability. As a matter of first importance, it is a benchmark. We can utilize it as the same old thing level we will accomplish if nothing changes in our methodology. Besides, we can compute the steady estimation of our new activities on head of this benchmark.it tends to be used for arranging. We can design our interest and gracefully activities by taking a gander at the gauges. It assists with seeing where to contribute more.

REFERENCES

[1] Sales Prediction using Machine Learning, Sunitha Cheriyan, Shaniba Ibrahim, Saju Mohanan, Published 2018 Computer Science 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE)

[2] A survey on retail sales forecasting and prediction in fashion markets, Systems Science & Control Engineering: An Open Access Journal, 3:1, 154-161, DOI: 10.1080/21642583.2014.999389

[3] Data Mining Model Performance of Sales Predictive Algorithms Based on Rapidminer Workflows, Alessandro Massaro, Vincenzo Maritati, Angelo Galiano Published 2018 Computer Science International Journal of Computer Science and Information Technology

[4] The Implementation of Data Mining Techniques for Sales Analysis using Daily Sales Data in International Journal of Advanced Trends in Computer Science and Engineering 8(1.5):74-80 · November 2019 DOI: 10.30534/ijatcse/2019/1681.52019

[5] Deloitte Sales Forecasting Deloitte Analytics Approach The growing world of data https://www2.deloitte.com/content/dam/Deloitte/it/Documents/technology/Sales%20forecasting_Deloitte%20Analytics%20Approach_Deloitte%20Italy.pdf

[6] Resource Quality Prediction Based on Machine Learning Algorithms 4th International Conference on Systems and Informatics (ICSAI 2017) DOI: 110.1109/Cybermatics_2018.00161 2017 4th International Conference on Systems and Informatics (ICSAI)

[7] Period Detection and Future Trend Prediction Using Machine Learning Techniques 21st Euromicro Conference on Digital Systems and Electronics Conference: 2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom)

[8] A Machine Learning Approach for Area Prediction of Hardware Designs from Abstract Specifications Volume 71, November 2019, 102853 published at IEEE Machine Learning Design productivity Area estimation

[9] Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis Volume 83, 2016, Pages 1064-1069 published at Elsevier Procedia Computer Science DOI: 10.1109/CIMCA.2018.8739696 published at IEEE.

[10] Disease Prediction by Machine Learning over Big Data from Healthcare Communities International Journal of Innovative Research in Computer and Communication Engineering Vol. 5, Issue 12 December 2017 DOI: 10.1109/ACCESS.2017.2694446

[11] Intelligent Sales Prediction using Machine Learning Techniques 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, August 2018.

[12] A Machine Learning-based Approach for The Prediction of Electricity Consumption Volume 134, December 2019 published at Elsevier State-of-the-art approaches to estimate energy consumption in machine learning

[13] A PSS model for diamond gemstone processing: economic feasibility analysis /Forty Sixth CIRP Conference on Manufacturing Systems 2017 published by Elsevier B.V 300A, box 2422, 3001 Leuven, Belgium doi: 10.1016/j.procir.2013.06.005.