



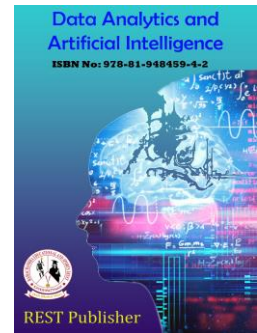
Data Analytics and Artificial Intelligence

Vol: 3(1), 2023

REST Publisher; ISBN: 978-81-948459-4-2

Website: <http://restpublisher.com/book-series/daai/>

DOI: <https://doi.org/10.46632/daai/3/1/2>



Residential Property Price Estimation Using Machine Learning

* S. Kavitha, P. Brinda, J. S. Sowmiya, P.D. Kishore Kumar, S. Kailasanath, T. Jayanthan

*veltech hightech dr.rangarajan dr.sakunthala engineering college,
Chennai, Tamil Nadu, India.*

*Corresponding Author Email: kavithaarunachalam58@gmail.com

Abstract: Property sales is not much straightforward in the country. Property pricing kept on changing day by day and at times that are being hyped rather than being based on valuation. Our system's key component is its ability to estimate property prices using actual variables. This article describes multiple regression algorithms that can be used with the Python library to estimate housing expenses. The suggested approach yields a more precise forecast by accounting for the complex factors employed in calculating home prices. The house price model and the datasets utilized in the suggested model are explained by this system using machine learning. In this system, we had used the linear regression algorithm. In cities like Mumbai, predictive models are still used to estimate residential sales prices due to factors more than just challenging and complex tasks. In cities like Mumbai, the cost of real estate for sale is influenced by a number of variables.

Keywords: Regression algorithm, Basic linear model, XG Boost Regression Model, Random Forest Regression

1. INTRODUCTION

Consider Artificial intelligence as a universal set then machine learning is subset of that .ML is commonly generalized as a capability of a machine to represent intelligence man behaviour. The systems those which use artificial intelligence (AI) are used to finish very hard tasks in a same path which is very common to how people solve problems. Machine learning seeks algorithmic answers to problem rather than just mathematical ones in computer science. It was relayed on finding algorithms which make machine learning possible. Unsupervised learning and supervised learning are the two important partitions of machine learning. So that it can make predictions when fresh data is supplied, the algorithm is trained on a specified set under close supervision. In unsupervised learning, the computer programme searches the data for connections and unobserved patterns. Problem Statement Frequently, the listing pricing and general details are displayed separately . Now the generic and conventional real estate features. They are listed in a separate way and systematically. Those features might be easily compared all over the full potential residences. All home , also having its own distinct qualities. For instance, consider a certain view of shelf or type of marble used for flooring, so home seller would give a summary of every essential components of the home in the detail. Those who could afford can compare with all real estate elements offered. Due to the higher level of variation, it is not possible to provide the instantaneous comparison on all elements. Objective The systems goal is to estimate the appropriate home pricing for real estate clients while taking their priorities and financial situation into account. Looking at previous market trends, price ranges, and impending changes might help predict future pricing. With the help of this new strategy, novice purchasers and clients who lack understanding will find it simpler to comprehend the value of overpriced or underpriced houses. Currently, the cost of a property is determined by the economic and social aspects of the area. We have thought about several fundamental factors (for example, number of rooms, living zone and so forth).The calculations for the linear regression model then make use of these parameter values. We calculated direct linear regression, which is used to predict how quickly an entity will sell its products. In order to effectively train a machine learning model, we must analyse several machine learning techniques in this study. Home price patterns, which also have a direct impact on buyers and sellers, provide insight into the status of the

economy at the moment. Numerous factors affect how much a house will actually cost. The number of bedrooms, bathrooms, and location are among them. In comparison to cities, costs are lower in rural areas. The housing price increased with factors such as proximity to a highway, a mall, a supermarket, employment prospects, top-notch educational institutions, etc. A few years ago, real estate businesses tried manually predicting property prices. For the purpose of estimating the cost of any real estate property, a dedicated management team is present in the organization. They made a choice.

2. RELATED WORK

Machine learning can be used to enhance and make very accurate price predictions in the real estate industry. Housing pricing is a fundamental customer choice model that allows for the estimation of home prices based on a variety of factors. The numerous analytical methodologies available to assess the price of real estate while taking into account various variables connected to the environment, the environment, other structures, etc. are new to the participants in the process. Users will be able to invest in real estate without using an agent thanks to the design. Additionally, it lessens the transaction's risks. We employed a number of machine learning techniques, such as Linear Regression, Random Forest Regression method, Cat Boost Regression algorithm, XGB Regression algorithm, and Ada Boost Regression algorithm, to estimate home values. In the known datasets, the remaining 75% of the data are taken for training purpose, and the remaining 25% are taken for testing purpose. We created several machine learning regression models from scratch throughout the project and learned everything there is to know about power transformers, regression models, and their creation. Support vector regression, linear regression, k-nearest neighbours, Linear Regression, Random Forest Regression method, Cat Boost Regression algorithm, XGB Regression algorithm, and Ada Boost Regression algorithm etc. are some of the techniques we have investigated to increase precision. Here, they made an assessment using an algorithm. This study gone through and reviewed past studies on the key elements that influences property prices as well as the strategies of data mining incorporated to forecast home prices. Technically saying, homes in urban regions with a wide range of amenities are probably less expensive than homes in non-developed locations having less amenities, even if they are located in an advantageous location like being close to a mall or other facility. A realistic estimate of a home's price might be made by investors, home purchasers, or home builders with the help of an accurate prediction model. The features that earlier researchers have utilized to predict the price of a home using multiple prediction models were covered in this work. Overall, the survey's findings have demonstrated the ability of, standard algorithms to forecast property values. These kind of models created and developed based on several input elements and have a large friendly impact on property prices. In summary, the impact of our study was in real to aid and support another researchers in real-world models that could predict house prices easily and accurately. Further work should be done on the actual model with the results This system or research paper carries three techniques to estimate real estate prices: support vector machine learning algorithm, random forest tech., and gradient boost machine technique. The outcomes of various algorithms are then analyzed while employing these techniques on a sample consisting of more than fifty thousand property exchanges that made in Hong Kong over the course of more than 18 years. In terms of prediction performance, forest and gradient boost have been found to outperform vector algorithm. Additionally, these two approaches outperform SVM according to the 3 acting criteria of mean square error find (MSEF), root and mean square error detect (RAMSED), and mean of absolute percentage error method. However, because SVM can generate reasonably accurate predictions internally, as our study demonstrated, it can still be a useful tool for data fitting. specified that the price fluctuation rate should have a method for its traceability. All over the world, forms such as hedonic price method, multiple linear regression analysis and decision tree regression, travel cost method, fuzzy logic system, etc. P. Durganjali proposed a house resales price prediction using classification algorithms. In this paper, the resale price prediction of house is done using different classifications algorithms like Leaner regression, Decision Tree, K-Means and Random Forest is used. There are so many factors are affected on house price include physical attributes, location and also economic factor as well. Here we consider RMSE as the performance matrix for different dataset and these algorithms are applied and find out most accuracy model which predict better results.

3. EXISTING SYSTEM

Predictive analytics is one of the various methods currently used to determine property prices. The original method searches for time series data. The time serial method involves looking for a co-relation between the current exchange price and the current price. The existing system will calculate the price of the house without knowing the necessary future and forecast information. This House Price Predictor project helps homebuyer get an idea of future prices. Real estate price forecasting also helps real estate traders know the value of that property in the future.

4. PROPOSED SYSTEM

The main objective for our project is to precisely predict the price of existing properties in India in the coming years using multiple algorithms. First of all, Linear Regression. This is a supervised learning technique and is responsible for assessing the value of a depend variable (Y) for a given in-depend variable (X). This is a relationship between input (X) and output.

5. MATERIAL AND METHODS

Machine learning: The study of "learning" processes, or processes that use data to improve performance on a set of tasks, is the emphasis of the field of machine learning. It's thought of as a part of artificial intelligence. Machine learning algorithms build a model from sample data, also known as training data, to make predictions or judgments without being explicitly trained to do so. In a variety of fields where it is difficult or impractical to develop conventional algorithms that can complete the necessary tasks, such as computer vision, speech recognition, email filtering, medicine, and agriculture, machine learning algorithms are used. A portion of machine learning is strongly related to computational statistics, which focuses on utilizing computers to make predictions, but not all machine learning is statistical.

Multiple Linear Regression: Regression analyzer is a model that is used to determine the connection between variables. To ascertain how tightly the variable are associated, use the co-relation co-efficacy. or the regressor equation. Multiple regression models could be used to pinpoint the characteristics that are most responsible for the depended variable's explanation. Regression technique also enables some price forecasts by gathering data from independent and dependent factors. The effectiveness of the regression model may be recognized when strength of the connection in between the depend and independent variables has been evaluated..The standard Python interface to the Tcl/Tk and GUI's toolkit is the T-kinter package ("Tk interface"). Most of the Unix platforms, that also includes mac's OS, and also Windows computers, have the feature of Tk and tkinter. Python -m T-kinter should launch a display displaying a simple kinter interface, just let us know if t-kinter is installed on our PC/LAP and also mentions which model of Tcl/Tk is installed, allowing you to read the documentation specific to that model. Numerous Tcl versions, made up either without / with thread's support, are supported by Tkinter. While every other detection variable in the model are "kept fixed," only one linear regression method is to be used to determine the relation between a single detector variable (x j) and the responding variable (y i). The interpretation of j, or the expecting value of the partial differentiation of y w.r. to x j, is the change in y i for a single unit change in x j while the some other variables are maintained constant. This commonly referred to as the special impact of x j on y. The end whole differentiation of y with respect to x j is the minor effect of x j on y, which can be measured use of a co-relation co-efficient or a linear regression model connecting solely x j to y i.

Multiple linear regression makes the following assumptions: the target and predictor variables should be related linearly. The residuals from the regression must be evenly distributed. MLR makes very few, if any, assumptions for multidisciplinary (correlation between the independent variables).

Packages: The tinder package The Tcl/Tk GUI tool kit's standard Python interface is called "Tk interface." Most of the Unix platforms, which also includes mac's OS, and also Windows computers, have te featue of Tk and tkinter You may check if tkinter is correctly installed on your system by typing python -m T-kinter from the command line. version of Tcl/Tk that is downloaded, allowing us to access the folders for that particular version of tk/Tcl. A variety of Tcl/kinder, both without and with thread's support, are supported by T-kinter. The reliable Python's binary launch bundles Tcl/kinter 8.7 thread. See that the supply code of the t-kinter module for greater facts approximately newer versions. T-kinter isn't always a skinny wrapper, however provides a truthful quantity of its very own common sense to make the enjoy greater pythonic. This will give attention to those extra changes, and check with the reliable Tcl/Tk files for information which is not changed.

Regression with ARIMA: R- Arima() is a function which makes a regression model with an R-A error, whether a xreg argument is used the error occurs. The order parameter specifies the order of the Rarima error design. If bias is mentioned, it is applied to all variables in the regression model before the model is determined. For example, the CHEAP command

fit <-Rarima(y, xpar=x, ord=c(0,1,0)) would fit into the model

$y_t = \beta_1 x_t + \eta_t$, where $\eta_t = \phi_1 \eta_{t-1} + \epsilon_t$, $\eta_t = \phi_1 \eta_{t-1} + \epsilon_t$ is an AR(1) error.

This will equivalence to the modal $y_t = \beta_0 + \beta_1 x_t + \eta_t$, $y_t = \beta_0 + \beta_1 x_t + \eta_t$, where η_t is an RARIMA (0,1,0) error.

To be note is the term K disappear because to the difference. To insert a K-constant in the discriminate modal, specify include. drift = TRUE. The auto. Rarima() function will able to take care of regression terms through the

xpar parameter. The subscriber should mention which forecast variable to include, but automatically. Arima () will choose the best R-ARIMA model for the mistakes . If discrimination is needed, all variables were differentiated during the process, even though the final modal is represented as the origin variables. Accuracy is calculated for the last process modal and this value can be used to determine the good estimators. That is, the methodology should be repeated for all subsets of the considered predictors and the model with the least AICc numbering should be chosen. The arguments are as follows: stands for the lagging order, or the number of lag occurrences The degree of subtraction, often known as the how many times do the raw readings are differentiable. moving average window dimension, commonly referred to as the moving average order.

6. METHODOLOGY

Data cleaning refers to taking away or switching data that is not precise very slow unnecessary, duplicated, or formatted wrongly in order to ready it for analysis. It might be getting rid of useless facts like cope with handling missing(null) values with inside the dataset. Data cleaning is the process of maintaining correctness or removing invalid, in consistant values, incorrectly typed, fake values , or unfinished data within a dataset. Plotting various graphs to represent the data and analyzing them. Deal with skewed distribution if found and apply Log transforming on them

Feature engineering refers to the treating of soft transitioning raw info into features that better replicate the unsolved problem to the estimate models, resulting in improved modal precision on unseen info. It will do encoding for categorical variables if found features calling for continuous variables It allows the machine learning algorithm to fast training the models. [Fig 1] It must decrease the complex in of a model and make an easy interpretation. It must improve the precision of a model if the right subset is chosen.It must decrease over fit.

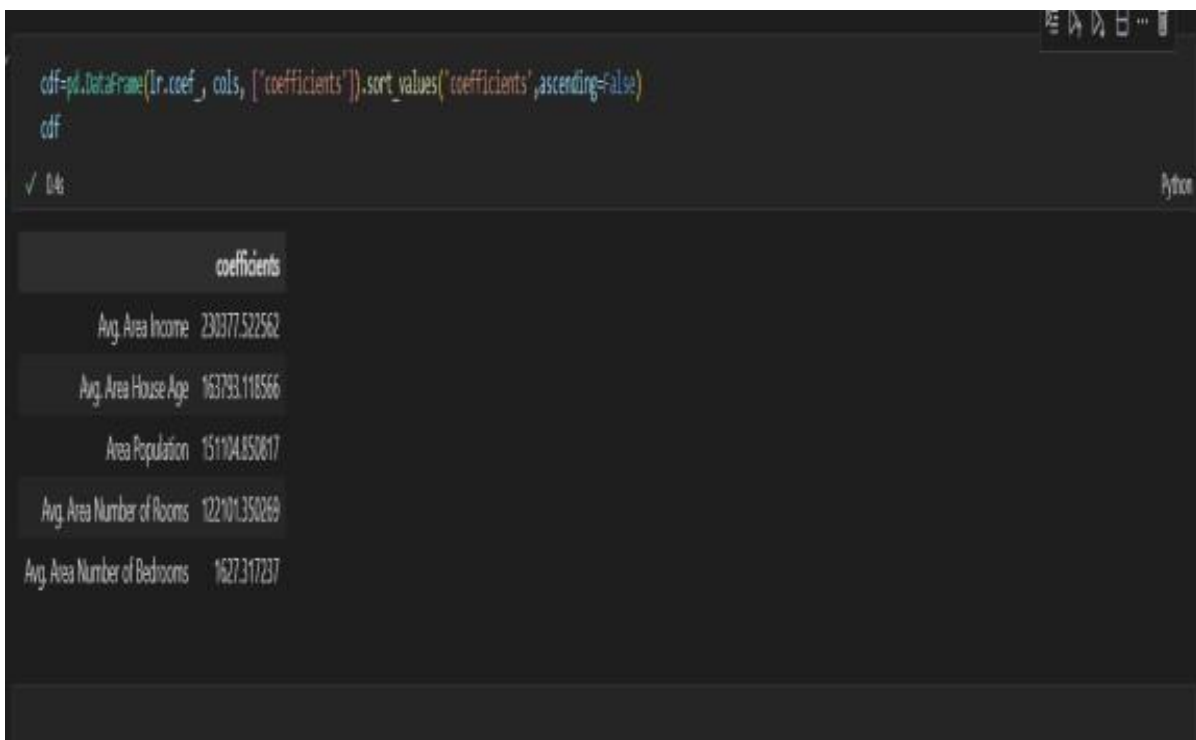


FIGURE 1. Feature engineering

variables	coefficients
Avg. Area Income	230377.5226
Avg. Area House Age	163793.1186
Area Population	151104.8508
Avg. Area Number of Rooms	122101.3503
Avg. Area Number of Bedrooms	1627.317237

FIGURE 2. Coefficient And Variable

Train-Test split and Generates it's prediction score Make our data in to parts for train and test purpose (75/25 a good starting point) Part the training info into training and evaluation (again, 80/20 is a fair split). Sub **sampler** and on selections of our train data, train the splinter with this, and make a note on the acting on the evaluation set. The following graph shows the range of price in the training dataset the negative values are generated by the Linear regression algorithm for it's own reference .

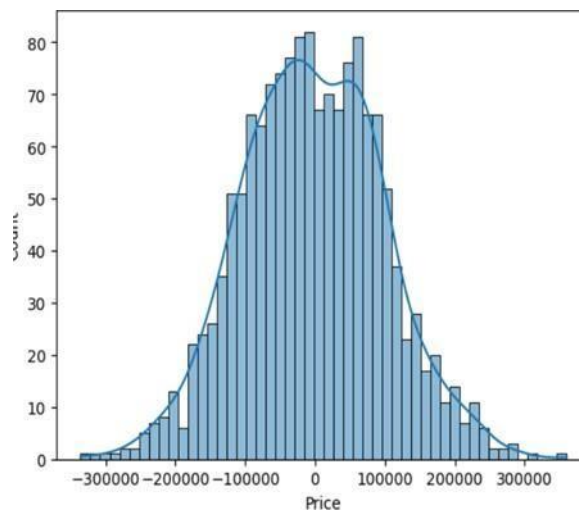


FIGURE 3. price range graph X axis – price Y axis - count(no. Of property with that price)

Linear regression and it Create UI using tkinter to make it user friendly. Regression model assess a number of the X variable given well known number of the Y variables. estimation within the range of numbering in the dataset used for model-fitt is termed better as interpolation. Estimation outside this range of the info is called as extrapolation There are two types of missing numbering the type one is known as missing at random completely. MCAR can beexpressed as $(R/X, Z, \mu) = P(R/\mu)$.

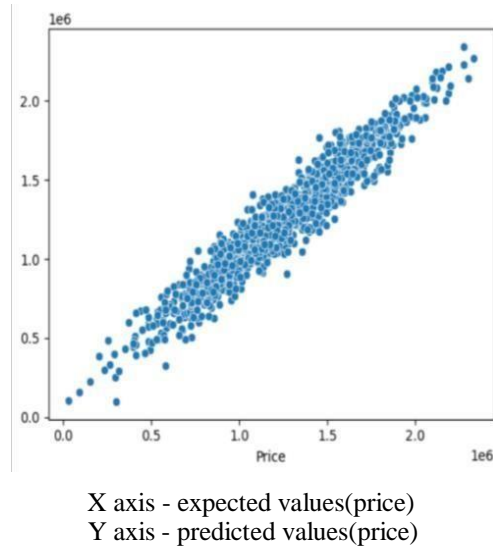


FIGURE 4. Visualization of attribute combination

A method for lowering the carnality of continuous and discrete data is binning, also known as discretion. Binning decreases the number of distinct values by grouping same values together in bins. Binning, a method proposed to decrease the effect of statistic noise, to overcome over fitting, decrease overall complex and make the model most robust. Entropy function is defined as: $En(S) = \sum_{i=1} P(C_i, S) \log(P(C_i, S))$

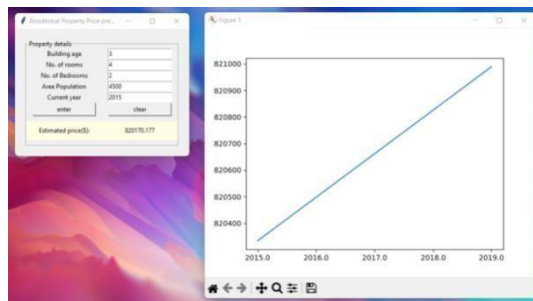
Here C are estimator classes in input dataset S , where maximizing info of the partition is deduced by

$$(T, S) = (|S1| / |S|) Ent(S1) + (|S2| / |S|) Ent(S2)$$

The both data is used for quantify the dependency in random variables, and it could be taken as the amount of data exchanged by two variables. MI is given as: $MI(X; Y) = \sum \sum p(xy) \log p(xy) / yp(x)p(y)$.

7. RESULT AND DISCUSSION

The following graph plots testing values on the x-axis and the predicted values on the y-axis the graph forms an almost diagonal line which shows the accuracy of the model. The accuracy of the model is well around ninety percent which is good compared to other existing systems.



X axis - year Y axis - price

FIGURE 5. output

Advantages of Proposed System: This system observes scoring units obtained with advancement of regression models. We could say that each of them behave in the same way. we could choose one for our price prediction system compared to the base model. Boxplots can be used to look for outliers. If there are outliers, we could remove the outliers and check and improve the models ability to perform. Advanced techniques such as forests method, convolution networks and element heat optimization can

be used for building models and increase estimation accuracy

Future Scope: It can predict the real estate property values at its high accuracy and predict the future property rates at high precision. Even after 5 years it can predict the values.

8. CONCLUSION

This project primarily predicts price based on the age of the property and the size of the property, the system still is able to predict the values of a given property with 90 percent accuracy. The only flaw of this system would be its inability to respect the randomness in the real world, even though it is an AI it still predicts based on a certain pattern. So if we can provide the system with live data it can quickly make predictions relatively faster because of the simplicity and minimal nature of the system. The main objective of this project was to get a rough estimation of a property in the future. This data can also be used to find the best location to build a property. The selling point of this project is its efficiency and rapidness. In the future, the system will observe units obtained with greater regression methods, and we can say that each of them will act much more similarly. You can choose one for house price prediction compared to the base model. Advanced techniques such as forests method, convolution networks and element heat optimization can be used for building models and increase estimation accuracy.

REFERENCES

- [1]. Prediction and Analysis of Housing Price Based on the Generalized Linear Regression Model- Xinshu Li
- [2]. <https://doi.org/10.1155/2022/3590224>. 2022
- [3]. House Price Prediction Analysis using Machine Learning Aniket Singh , Adarsh Kumar Singh , Aditya Raj , Harshit Jain,Mrs. Asha M S,IJCRT – 2022
- [4]. Machine Learning based House Price Prediction Model- Chen Chee Kin; Zailan Arabee Bin Abdul Salam; Kadhar Batcha Nowshath-IEEE 2022.
- [5]. House Price Prediction using Machine Learning-IEEE 2022.
- [6]. Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times-
<https://doi.org/10.3390/land11112100>. 2022
- [7]. A Comparative Study on House Price Prediction using Machine Learning Hardi Joshi , Saket Swarndeep-IRJET 2022.
- [8]. House price prediction with gradient boosted trees under different loss functions – Journal of property research 2022.
- [9]. House price prediction using machine learning-IJME 2022
- [10].House Price Prediction-Journal of Emerging Technologies and Innovative Research- 2022.
- [11].Machine learning based prediction of house pricing- A. Swamy Goud, Dr. Sameera Divanu, B. Nikhil Goud- International Research Journal of Modernization in Engineering Technology and Science-2022.
- [12].G. Gao, Z. Bao, J. Cao, A. K. Qin, and T. Sellis, “Location-centered house price prediction: a multi-task learning approach,” ACM Transactions on Intelligent Systems and Technology (TIST), vol. 13, no. 2, pp. 1–25, 2022.
- [13].D. Cao and X. Tian, “Raw anode volume density prediction algorithm based on the genetic algorithm,” SN Computer Science,2022.
- [14].J. H. Chen, T. Ji, M. C. Su, H. H. Wei, V. T. Azzizi, and S. C. Hsu, “Swarm-inspired data-driven approach for housing market segmentation: a case study of Taipei city,” Journal of Housing and the Built Environment, 2021.
- [15].K. S. Yoon, J. M. Lee, S. J. Ko, H. J. Kim, and J. H. Kim, “Analysing impact of price ceiling system on housing market using machine learning,” Journal of the Architectural Institute of Korea,- 2021.
- [16].House Price Prediction Using Machine Learning June 2021 The Journal of Philosophy Psychology and Scientific Methods 9(6):2455-6211
- [17].HOUSE PRICE PREDICTION USING MACHINE LEARNING Khan Sohil Liyaqatullah , Usman Malik , Qureshi Mohammed Basheeruddin- 2021.
- [18].C. Li, H. Zhu, X. Ye et al., “Study on average housing prices in the inland capital cities of China by night-time light remote sensing and official statistics data,” 2020.
- [19].House Price Prediction Bindu Sivasankar1 , Arun P. Ashok2* , Gouri Madhu3 , Fousiya S.-IJCSE-2020
- [20].Winky K.O ho, Bo Sintang, Siuwa Wong — Predicting property prices with Machine Learning ALgorithms - 2020.
- [21]. W. T. Lim, L. Wang, and Y. Wang, — House Price Prediction Usinf a Machine Learning Model: A Survey Of Litratue(IJMEC) -2020..