Abstract:

Predicting the price of a house becomes very challenging considering that the place, size of the house, number of bedrooms, and number of available amenities have a strong influence on the ultimate price. Our aim, in this project, was to create a Python model that will predict house prices using provided history data. Our data source was Kaggle Housing dataset, and we tested various machine learning techniques, such as Linear Regression, Support vector Regression, to determine which would produce the best outcome. To make the models more precise, we excluded missing data, normalized the data, and converted categorical values into numbers. The models were then validated against metrics such as the Mean Squared Error (MSE) to determine how well they were trained. The inference made was that Support vector regression , being an ensemble method, gave the best results because it can grasp complex relationships in the data better. The purpose of this project was to demonstrate how powerful Python and machine learning are in enhancing decision making in real estate.

Introduction:

The property market is a dynamic and complicated system with numerous variables such as size of property, the number of bedrooms and bathrooms, surrounding facilities, location, infrastructure, and local economic conditions. We can study the dynamics between these variables and how they influence property prices by employing the help of Python. It is useful information that can help developers, purchasers, sellers, and policymakers.
In this project, we apply data science techniques using Python to analyze and forecast house prices. The data contain quantities such as house size, number of bathrooms, and selling price, and yes/no attributes that indicate whether or not particular amenities exist.

We start by preprocessing the data, cleaning it to get rid of missing data and to get it in the correct format. We then get to know the data and plot charts to identify patterns and understand relationships between various features. We also examine basic statistics to determine the strongest factors affecting prices. The primary objective is to train a machine learning model, primarily Linear Regression, to predict house price from the features.

This project demonstrates how Python can be utilized step-by-step, from data cleaning to

model building, to address actual problems in the housing market.

Literature Review:

Title of Paper	Published Year	Dataset Used	Methodology Used in Existing Paper	Key Findings/Results
Housing Price Prediction via Improved Machine Learning Techniques [14]	2020	Beijing Housing Dataset (300k+ records)	Applied Random Forest and XGBoost models and suggested Stacked Generalization (Stacking), a meta-model that learns to make predictions by learning from the predictions of the base learners. Being an ensemble technique, this used multiple predictions for robustness and error reduction.	Stacked Generalization produced the smallest RMSLE and outperformed individual models. The paper validates that ensemble-based methods are more accurate and versatile with very complex and non-linear housing data.
House Price Prediction using Random Forest Machine Learning Technique [15]	2022	Boston Housing Dataset	Targeted towards the Random Forest regression model due to its capability to reduce overfitting and handle a large number of features. Feature importance was also taken into account to select major contributors towards house price.	Random Forest was very accurate and had strong generalization capability. Number of rooms and tax rate were the most significant predictors. Tree-based ensemble models were useful for data with high-order interactions.
A Literature Survey on Housing Price Prediction [18]	2022	Multiple datasets taken from 99acres, nobroker.com and magicbricks.com	Compared different ML models like Linear Regression, XGBoost, and CNN, keeping in mind strengths and weaknesses of models. A focus was on relative accuracy and applicability of models based on availability and shape of data.	The survey concluded by reinforcing that hybrid and ensemble models consistently outperform the one-model methods. It highlighted the integration of geo-tagged and real-time data to enhance prediction accuracy and relevance.

Machine Learning Approach for House Price Prediction [16]	2023	Real estate dataset (not named)	Utilized and compared Linear Regression, Support Vector Machines (SVM), Lasso Regression, Random Forest, and XGBoost to forecast house price. Focused on preprocessing, feature selection, and model performance using accuracy and performance metrics.	XGBoost yielded the best accuracy, further proving the stability of the model in handling non-linear and high-dimensional data. Model tuning and ensemble methods were highlighted in the study to be critical in price prediction.
House Price Prediction Using EDA and ML with Feature Selection [17]	2022	Kaggle House Price Dataset	Carried out EDA and subsequently fit regression models: Linear, Ridge, Lasso, and Elastic Net. Recursive Feature Elimination (RFE) was applied for the majority of relevant feature selection. Models were assessed on R^2 on train and test datasets.	RFE greatly improved model performance.Achieved R^2 values of 0.94 (train) and 0.86 (test), which indicate that including EDA alongside robust feature selection techniques enhances prediction capability.
House Price Prediction Using Machine Learning and Artificial Intelligence [19]	2024	Kaggle Dataset	Performed Exploratory Data Analysis (EDA), followed by Linear Regression and Random Forest Regression. Performed scikit-learn libraries to train model, test, and model evaluation.	Random Forest was superior to Linear Regression, especially in dealing with feature interactions. Determined that using ML regression with proper preprocessing is a reasonable way to estimate prices.
House Price Prediction Modeling Using Machine Learning [12]	2020	Custom dataset (Tadepalligudem, Andhra Pradesh)	Used both the classification and regression algorithms using Decision Tree Classifier and Regressor. Also tried using Multiple Linear Regression. Specifically, experimented with location variables such as transport, schools, and amenities.	Decision Trees could solve both the classification problem and the regression problem. Found locality-based features to have a significant role in price determinations and suggested such models for regional house prediction.

Housing Price Prediction Incorporating Spatio-Temporal Dependency into Machine Learning Algorithms [11]	2022	32-year dataset from SAILIS (Adelaide, Australia)	Trained four ML models—Decision Tree, Linear Regression, Random Forest, and Gradient Boosted Tree—upon which a spatio-temporal lag (ST-lag) feature was implemented combining location and time-based attributes. PySpark and ArcGIS used for preprocessing. Models tested via R^2, MAE, RMSE.	Gradient Boosting Tree reported maximum accuracy ($R^2 = 0.896$). Considering ST-lag highly influenced the performances of all the models. The article revealed the crucial position of space and time-based variables in well modeling real-life housing markets.
Housing Price Prediction with Machine Learning [20]	2022	Boston Housing Dataset	Applied Used Linear Regression, Decision Tree, and Random Forest models and compared them. Applied preprocessing techniques like data cleaning, feature scaling, and correlation-based EDA. Applied the train-test split of the dataset, tested using MSE and cross-validation.	Random Forest was better than Linear Regression and Decision Tree models with the lowest MSE (~2.90) since it was suitable for small data.The research identifies ML as a possible substitute for hand valuation and suggests use of larger and more diverse datasets in subsequent studies.
Machine Learning Based Predicting House Prices Using Regression Techniques [13]	2020	Bengaluru housing dataset from Machine Hackathon platform	Compared five regression algorithms: Linear Regression (OLS), Ridge, Lasso, SVR, and XGBoost. Conducted aggressive preprocessing such as encoding, log scaling, and unit conversion. Tuned using RMSE, R^2, and RMSLE with GridSearchCV hyperparameter tuning.	XGBoost and SVR were best with maximum R^2 and minimum RMSLE. Regularized models outperformed basic regression. The paper points toward SVR and XGBoost as being suitable for urban residential markets and suggesting future work with more detailed information and broader application.

Dataset Description:

The project data is located in a file named traindata.csv. This file has information on houses used to develop models to predict house prices. Each row of the file is one of the houses, and it has information regarding the home.

The cleaning of the data was simply dropping columns that were not relevant (we did not need grade, date, month, quartile_zone, etc.), and we kept the relevant features to a few which consisted of:

bedrooms: How many bedrooms

real_bathrooms: How many bathrooms

living_in_m2: the living area in square feet

single_floors: whether only single floor or more

nice_view: whether a nice view or not

perfect_condition: the general condition of the house

price: the price that the house was sold for (which is the price we wish to predict)

The data includes numerical values and category types of data. This is ideal for a regression model to make price predictions. We pre-cleaned the data before creating the model by dropping columns that were unnecessary, filling in missing values, and rescaling features when needed.

This data set has a lot of information that will be helpful in analyzing the effect and influence of house characteristics on prices.

Methodology:

The project made use of a standard data science workflow under which we undertook data preparation, data exploration and modelling, and model evaluation stages. Later sections detail each stage.

4.1 Data Preparation

In the first step, we preprocessed, cleaned and prepared the dataset for analysis by taking the following actions:

1. We dropped all columns relating to the prediction of house prices which we considered irrelevant, such as grade, date, month, quartile_zone.

2. Found the outliers and dropped them for a better model.

3. We assessed the data for missing values, and handled them by either filling their data where applicable, or dropping those rows.

4. We changed the data types as needed for machine learning algorithms to use the data as efficiently as possible.

5. We used only the most relevant attributes for house price prediction.

4.2 Exploratory Data Analysis (EDA):

The second step was an exploratory data analysis (EDA) that enables us to create graphs and charts with Matplotlib and Seaborn in order to:

- Observe how the numerical features are distributed.

- Look at relationships between features (e.g. bedrooms, bathrooms) and the dependent variable (house price).

- Find observations that could be outliers.

- Look at relationships between features using a correlation heatmap.

4.3 Model Development

In order to forecast house prices, we divide the data into two sets, one for model training (75%) and the other for model testing (25%). The first model that we used was a basic linear regression model that assumes a linear relationship between independent features and the dependent variable price (house price). We also used support vector regression(SVM)to check whether there are more complex nonlinear relationships of independent features with the dependent variable price.

4.4 Model Evaluation
We compared the performance of every model using the following standard metrics.
Mean Squared Error (MSE): Calculates the average of the squared differences between forecasted and actual prices.

Root Mean Squared Error (RMSE): Outputs the error with the same unit of measurement as that of house price, so easier to understand.

Once we applied these steps, we chose the most accurate performing model to predict housing prices.

Results & Discussion / Visualizations:

The performance of the predictive models was assessed after training them using the relevant regression metrics and visualizations that would help in interpretation and insights.

5.1 Model Performance

Linear Regression was the primary model utilized in this project and was utilized as a baseline model for the prediction of housing prices. The following evaluation metrics were computed for the model from the test dataset:

Mean Squared Error (MSE):0.57

Root Mean Squared Error (RMSE): 0.75

Mean Squared Error (MSE):

The MSE gives us a sense of how far off our predictions are from the actual values by averaging the squared differences. A score of 0.57 suggests that the model has a decent predictive ability, with errors staying within a manageable range. This indicates that there are opportunities for systematic improvements while also affirming the model's overall validity.

Root Mean Squared Error (RMSE):

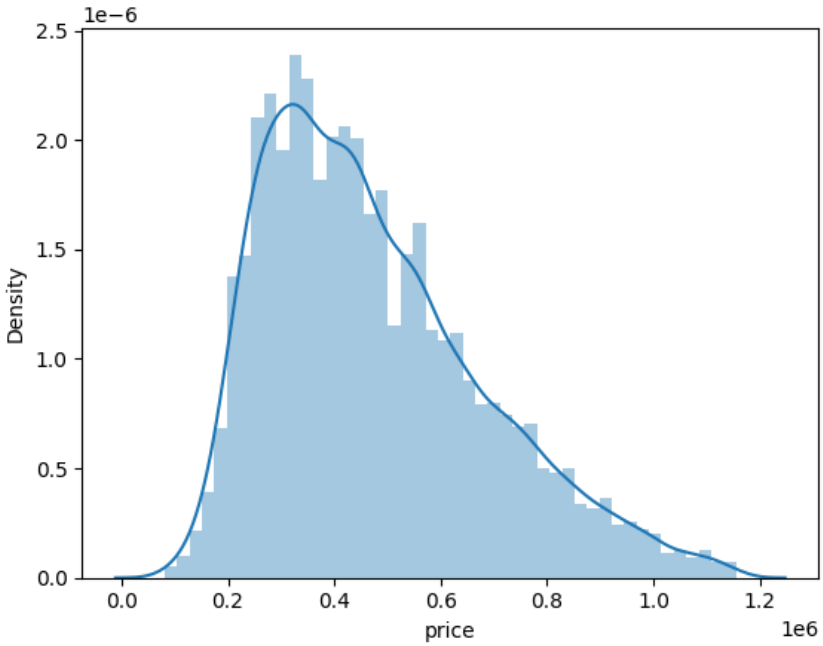
With an RMSE of 0.75, we can translate those prediction errors back into the original scale of the target variable. This means that, on average, our predictions deviate by about 0.75 units, showing that the model can provide useful estimates, though there are still some precision gaps that we can work on optimizing.

Interpretive Significance:

When we look at these measurements together, it’s clear that the model shows a statistically significant correlation with the actual observations. However, making small adjustments could enhance its reliability for more critical applications. The RMSE, being aligned with the units of measurement, allows for easy comparisons against specific tolerance thresholds in the field.

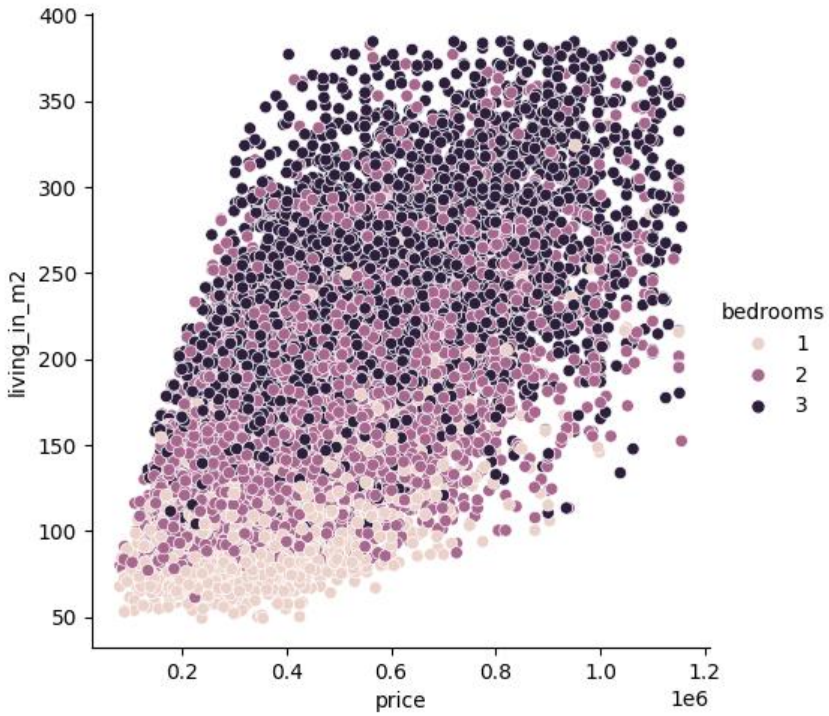
5.2 Visualizations:

The following visualization methods were created to compare the data and model performance:

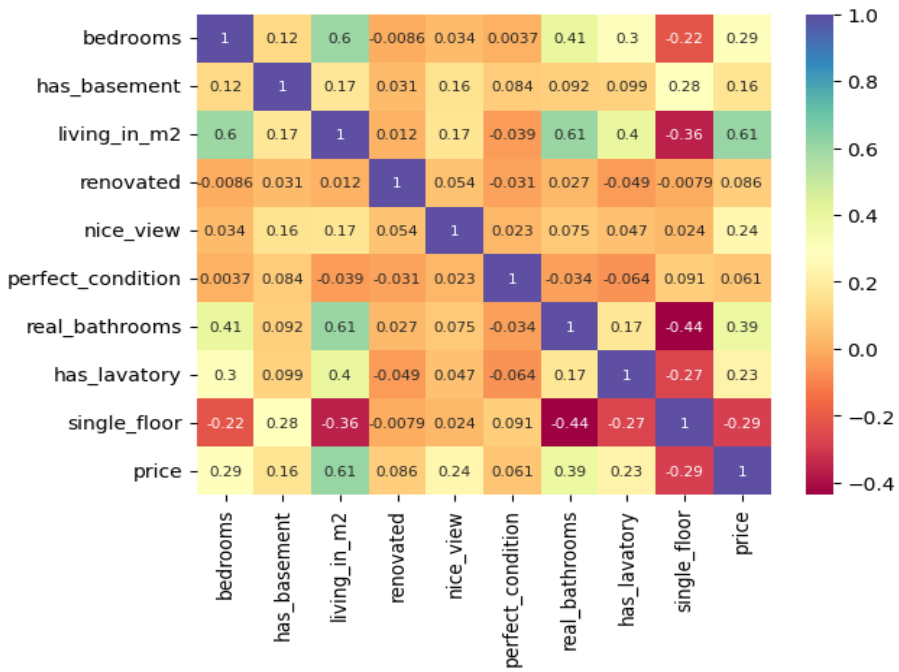


Price Distribution Plot: A histogram of house price distribution was plotted. The data showed a right skewed distribution with bunching of low-price houses and only a few high-outlier values.

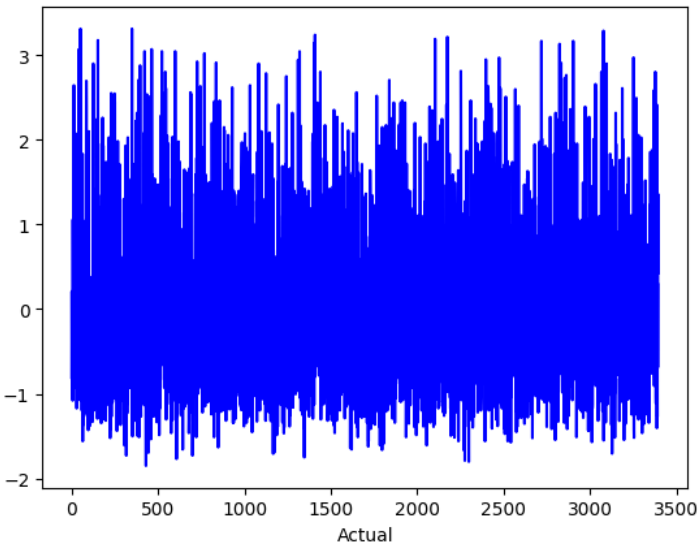
Relational Plot of Size and living in m2 (classified on the basis of no. of rooms): This analysis revealed a fairly positive correlation between size (as measured in sqft_living) and sale price and we also got to know that how price was distributed on the basis of no. of rooms and outliers were also observed as there were some houses of 1 room had price more than those of 3bhk.

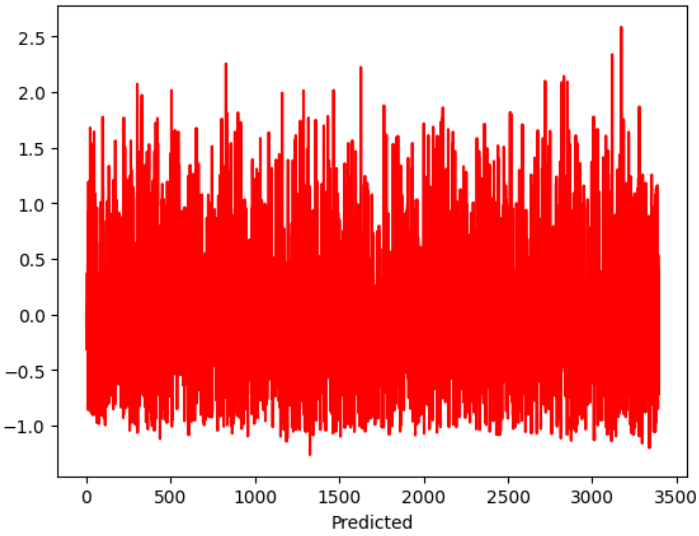


Correlation Heatmap: This plot provided a visual summary of strength between relationships between each feature. bathrooms and sqft_living were more correlated features that strongly related to price whereas yr_renovated and other features were less correlated to price.



Plot Actual vs. Predicted Price: This plot confirmed model performance by plotting actual prices to predicted price values.





5.3 Significant Findings

Larger houses, houses with nice views tended to be in a higher price range.

Bathrooms offered and the condition of the house were some important features.

Outliers and variance with some features influenced the price estimates. More advanced predictive models can better account for non-linear features that are based upon different dimensions of the data.

Conclusion:

In this project we considered using Python machine learning and analytics methods to assist in giving some estimated estimates of house prices based on data from the real estate market. Through this project we began with raw data and then we preprocessed, exploratory data analyzed, created some models, and explored the data to see what factors positively affect price of real estate.The outcomes we encountered verified that square feet, bathrooms, water, condition, and many more alternatives all have a positive effect on housing prices. Our linear regression also provided us with an initial model that was able to determine some linear relations, but was highly affected by outlier data and inter feature non-linear relations.

As part of the project we were able to illustrate that predictive modeling can be applied in decision making regarding the real estate market. By illustrating the property characteristics in conjunction with a predicted price we illustrated a predictive capability. Later, while simple linear models can show predictive associations, we can take a few additional steps to discover nonlinear associations or other interactions (random forests, gradient boosting) and take some time to fine-tune model parameters. Finally, we have illustrated practical applications of machine learning and Python to solve real world issues with a focus on some of the issues related to data pre-processing, visualizations, and evaluation in any analysis/data driven methodology.

References:

[1] Kaggle. House Prices: Advanced Regression Techniques. [Online]. Available: https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques

[2] F. Pedregosa *et al*., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[3] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed. Sebastopol, CA: O'Reilly Media, 2019.

[4] J. Brownlee, *How to Evaluate Machine Learning Algorithms*. Machine Learning Mastery, 2016. [Online]. Available: https://machinelearningmastery.com

[5] M. Kumar and M. Singh, "House Price Prediction Using Machine Learning Algorithms: A Review," *Int. J. Sci. Technol. Res.*, vol. 9, no. 1, pp. 1386–1390, 2020.

[6] W. Iqbal and A. Ali, "Prediction of House Prices Using Machine Learning Techniques," in *Proc. 2021 4th Int. Conf. Artif. Intell. Big Data*, 2021, pp. 31–35.

[7] J. Li, J. Zhang, and Z. Liu, "Real Estate Price Prediction with Regression Models and Ensemble Methods," *Procedia Comput. Sci.*, vol. 127, pp. 377–383, 2018.

[8] S. M. Salaken and M. S. Hossain, "Regression Analysis for Predicting Housing Prices in Real Estate Market," *J. Comput. Commun.*, vol. 8, no. 4, pp. 16–26, 2020.

[9] D. P. Acharjya and A. Mitra, "Data Mining and Machine Learning in Real Estate," *Int. J. Appl. Eng. Res.*, vol. 11, no. 6, pp. 4090–4096, 2016.

[10] Python Software Foundation, *Python 3.12 Documentation*. [Online]. Available: https://docs.python.org/3

[11] A. Soltani, M. Heydari, F. Aghaei, and C. J. Pettit, "Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms," *ISPRS*

International Journal of Geo-Information, vol. 11, no. 4, 2022.

[12] A. Begum, N. J. Kheya, and M. Z. Rahman, "Housing price prediction with machine learning," in *Proc. 5th Int. Conf. on Electrical Engineering and Information & Communication Technology (ICEEICT)*, 2022.

[13] M. J. Manasa, R. Gupta, and N. S. Narahari, "Machine learning based predicting house prices using regression techniques," in *Proc. 2nd Int. Conf. on Innovative Mechanisms for Industry Applications (ICIMIA)*, Bengaluru, India, 2020, pp. 624–630.

[14] Q. Truong, S. Le, V. Nguyen, and T. Nguyen, "Housing price prediction via improved machine learning techniques," in *Proc. 12th Int. Conf. on Knowledge and Systems Engineering (KSE)*, 2020, pp. 146–151.

[15] A. B. Adetunji, S. O. Isewon, and A. O. Olugbara, "House price prediction using random forest machine learning technique," in *Proc. Int. Conf. on Intelligent and Innovative Computing Applications (ICONIC)*, 2022.

[16] M. J. Chowhaan, M. Ramalingam, and M. K. Ragavan, "Machine learning approach for house price prediction," in *Proc. 2nd Int. Conf. on Smart Electronics and Communication (ICOSEC)*, 2023.

[17] F. M. Basysyar and G. Dwilestari, "House price prediction using EDA and ML with feature selection," in *Proc. Int. Conf. on Data Science and Its Applications (ICoDSA)*, 2022, pp. 73–78.

[18] S. S. Yalgudkar, A. G. Patel and V. R. Jangid, "A Literature Survey on Housing Price Prediction," *International Research Journal of Modernization in Engineering Technology and Science (IRJMETS)*, vol. 4, no. 2, pp. 1200–1205, 2022.

[19] F. Maloku, B. Maloku and A. A. D. Kumar, "House Price Prediction Using Machine Learning and Artificial Intelligence," *Journal of Artificial Intelligence & Cloud Computing*, vol. 3, no. 4, pp. 1–10, Aug. 2024, doi: 10.47363/JAICC/2024(3)357.

[20] M. Thamarai and S. P. Malarvizhi, "House Price Prediction Modeling Using Machine Learning," *I.J. Information Engineering and Electronic Business*, vol. 12, no. 2, pp. 15–20, Apr. 2020, doi: 10.5815/ijieeb.2020.02.03.