PREDICTING HOUSE PRICES USING MACHINE LEARNING

INTRODUCTION:

As Artificial Intelligence is involving everywhere in the world there is stupendous amount of goodness in our day-to-day life and implementation of various advanced machines has been increased. As growth of Innovations to Business is going upward computer sciences tend to increase technological transformations. This can put out the vulnerability of security and increase protection of the data. By considering various machine learning models and using the data of real estate forms in Boston we predict the house prices in entire Boston. This project is all about predicting the house prices by considering the datasets of Boston real estate by using different class labels. As we need the data to predict house price, the supervised data is produced which plays key role in predicting the house price and help indealing with the real estate entities. As we are using machine learning it is easier to achieve the target like higher intelligent predictions which are a benefit fact or for futuristic projects and intelligent systems which are linked to robotics as well. Now a days, smartphones are super-advanced and handy devices which could be used for almost every daily tasks instead of laptops. Smartphones applications are widely available, popular and are easily adopted. And so, we developed an Android app which displays the real-time COVID19 data across the globe, through which every user will know about the situation going around the world regarding the COVID19 and thereby they will be able to stay updated and safe.

Main methodology of machine learning is constructing the models using past data as a source to predict the new data. As population is increasing rapidly the market demand is also increasing at the same pace. Most of the public are vacating the rural areas because of scarcity of jobs and increment of unemployment. This ultimately results in increment of houses in cities. If they don’t have enough idea about prices then it results in loss of money.

INNOVATION :

Certainly Predicting house prices using machine learning can be a valuable application. Here are some innovative ideas and approaches:

- **Time Series Analysis**: Incorporate time series data, such as historicalhouse price trends, economic indicators, and seasonality, to improvepredictions.

- **Natural Language Processing (NLP)**: Analyse real estate listings,descriptions, and reviews to extract sentiment and neighbourhoodfeatures that can impact prices.

- **Image Analysis:** Use computer vision to analyse property images to extract features like curb appeal, interior design or outdoor amenities that influence prices.

- **Geospatial Data**: Utilize geographic data like proximity to schools, parks,public transportation, and crime rates to enhance predictions.

- **DeepLearning**: Implement deep learning models like Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) for more accurate predictions based on complex data.

- **Ensemble Methods**: Combine multiple machine learning models, such as Random Forests, Gradient Boosting, and Neural Networks to improve prediction accuracy.

- **Feature Engineering**: Create new features like walkability scores, commute times, or neighbourhood safety indices to capture important aspects of location.

- **Transfer Learning**: Apply pre-trained models (e.g., BERT for NLP tasks)to extract features from textual data related to properties.

- **Anomaly Detection**: Identify unusual patterns or outliers in house pricedata to detect potential investment opportunities or areas for further investigation.

- **Interactive Web Applications**: Develop user-friendly web apps or mobile apps that allow users to input property characteristics and receive instant price predictions.

- **Blockchain and Smart Contracts**: Explore the use of blockchain technology for transparent and secure property transactions, which could impact pricing models.

- **Predicting Future Value**: Rather than just predicting current prices, createmodels that forecast how property values might change over time, helping investors make long-term decisions.

- **Environmental Factors**: Consider environmental factors like climate change risks, flooding, or air quality, which could affect property values in the future.

- **Data Fusion**: Combine various data sources, including social media sentiment, local news articles, and economic forecasts, to build a comprehensive prediction model.

- **Explainable AI**: Ensure your model provides interpretable explanationsfor its predictions, which can be crucial for real estate professionals.

- **Market Segmentation**: Develop models that segment the housing marketinto different buyers categories (e.g., luxury, affordable, starter homes) for more targeted predictions.

- **Collaborative Filtering**: Apply collaborative filtering techniques,similar to recommendation systems, to predict house prices based on user preferences and Behaviour.

- **Energy Efficiency**: Incorporate energy efficiency ratings and potential cost savings into price predictions, as green features become more important to buyers.

- **Legal and Regulatory Analysis**: Consider how changes in local zoninglaws, tax policies, or regulations might impact house prices.

- **Crowdsourced Data**: Use data from platforms like Zillow, Redfin, or Airbnb to supplement your dataset and improve prediction accuracy.

- Remember that innovation in this field often requires a deep understanding of both machine learning techniques and the real estate market, as well as access to diverse and high-quality data sources.

IMPLEMENTATION :

- **Data Pre-Processing :**

  - **Checking for null Values**: Missing item or null value is defined as the data that is not stored or is absent for a variables within the dataset provided. There are multiple reasons why certain values are missing from the info. Some of the explanations are past data might get corrupted because of improper maintenance, observations aren't recorded in sure enough fields because of some reasons, there can be a failure in recording the values because of human error and the user has not provided the values intentionally.

  - **Checking if the data is distributed normally or not**: Normal Distribution is one of the crucial concepts in statistics and therefore also considered the backbone of Machine Learning data distribution. A knowledgeable Mathematician has to fathom distribute after they work with Linear Models and has to perform well if the information is normally distributed and if the central limit theorem and exploratory data analysis are considered. The Distribution if it is normal does carry its assumptions and maybe completely could be specified by 2 parameters those are the mean and therefore the variance. If the mean value and variance value are known, we will be able to access every datum on the data curve.

  - **Checking for Outliers**: In the machine learning perspective, an outlier is defined as a point that is farther from all other points. The above statement tells to us that if an outlier is there like some odd one out or the one that is farther from the gang. Few statistics define the outliers distant from all the other points. There is no need to confuse this statement there upon an imbalanced dataset, though there may be few similarities within the definitions.

- **Checking the correlation between attributes:**

- **Data Correlation**: Thiscould be thanks to understanding the link or dependency between many variables or attributes of the dataset. With the help of correlation, we can obtain some insights.1 or more attributes depend upon another attribute or a reason behind an additional attribute. One or more attributes how they are related to the remaining attributes.

- **Positive Correlation**: That means if a feature X decreases, then the feature Y also decreases or if feature Y increases, then feature X al so increases. Both features move in correspondence and there is a linear relationship between them.

- **Negative Correlation**: implies that if feature X decreases, then feature Y must increase and the vice versa.

- **No Correlation**: There is no link between those 2 attributes.

- **Exploring various ML models:**

- **Linear Regression**: Supervised type of machine learning is supported by Linear Regression algorithm. It accomplishes the regression tasks. The regression technique models the target/output prediction to the independent variables. It is the mostly used algorithm for locating the connection between various attributes and the forecasting. Many regression models are different in supporting the sort of relationship between dependent variables and the independent variable they want, and therefore a good number of independent variables are being used. This regression algorithm accomplishes the task of predicting a variable value (v2) supported the given variables (v*). So, this model finds out the linear relationship between v* (input) and v(output). So, it is also known as statistical regression.

- **K-Nearest Neighbours (KNN)**: The K-Nearest Neighbours algorithm is one in every of the many popular Machine Learning algorithms that support Supervised Learning category. This algorithm presumes the homogeneity between the newest data and available data and put the newest data into the category that is most alike one of all the available categories. This algorithm stores all the pre-fetched data and segregates the fresh information supported by the homogeneity. This helps when the fresh data appears then it is often easily label into the compatible class.

- **Support Vector Regressor (SVR)** : The Support vector regression is one of the supervised learning algorithms that has accustomed prediction of discrete values. This model use the identical principle because the SVM principle. The essential idea behind this SVR is to seek out the most effective fitted line. In SVR, the most effective fit line is that the hyper plane that has the utmost number of points. The regression models of machine learning tries to decrease the error between predicted and important value, this model tries to suit the simplest line near to the threshold value. Edge value is that the distance from the hyper plane to the boundary line. The fitting time complexity of this model is quite quadratic to the quantity of samples taken which makes it hard to scale to available datasets with over pair of 10 thousand samples.

- **Random Forest Regressor (RFR)** *:* Decision Trees can be utilized for boththe regression and classification tasks. They visually flow like trees, hence the name, and within the regression case, they begin with the foundation of the tree and follow splits supported variable outcomes until a leaf node is reached and also the results are given. The forest can be an estimator thatmatches a variety of Labelling decision trees on many sub-samples of given data and makes the use by averaging to boost the prediction accuracyand controls the overfitting.

Maximum samples parameter controls the size of every sub-sample if the bootstrap is True (by default), otherwise the entire data is employed create every tree.

- **AdaBoost Regressor (ABR)** : The AdaBoost regression algorithm may be an estimator which begins by proper fitting of a regressor based on the initial data and so fits additional copies of the regressor on the identical dataset but the weight of every instance is then adjusted in step with the error of this prediction. As such consequence, subsequent regression models focus more with the tougher cases. The decision tree is then boosted using the ADA boost algorithm. This algorithm on the 1D sinusoidal dataset with some quantity of the gaussian noise. 299 boosts on 300 decision trees are analysed.

- **XG-Boost Regressor (XGBR)**: The XG of the XG-Boost stands for Extreme Gradient which is a free and open-source library that produces an effective implementation of the gradient boosting algorithm. Soon after the development and its first release. This algorithm became the go to technique and infrequently the important aspect to win the solutions for a variety of tasks in machine learning contests. Prediction based regression modelling tasks involve the prediction a numeric value like an amount or a distance. This algorithm is often used very directly for prediction-based regression modelling. The gradient boosting points to a class of Machine Learning algorithms related to ensemble learning which can be used for either of classification or regression tasks. Regression predictive modeling problems involve predicting a numerical value such as a dollar amount or a height. XGBoost can be used directly for regression predictive modeling**.**

- **Cat Boost Regressor (CBR)** : The Cat Boost is built upon the speculation of decision trees algorithm and the gradient boosting algorithm. The idea of boosting is to add many weak built models and so by the greedy search technique, we can build a robust competitive model for prediction. As the gradient boosting fits the choice trees one after the other, the learned trees will learn from its mistakes and so, that's how it Reduce the errors. So, this way of adding a new functionality to the existing methodology is sustainable until the chosen loss method isn't any longer reduced.

FEATURES:

So to deal with this kind of issues Today we will be preparing. A Machine Learning Based model, trained on the House Price PredictionDataset.

The dataset contains 13 features:

- Id to count the records.
- MS Sub Class Identifies the type of dwelling involved in the sale.
- MS Zoning Identifies the general zoning classification of the sale.
- Lot Area Lot size in square feet.
- Lot Configuration of the lot
- Bldg Type of dwelling
- Overall Cond Rates the overall condition of the house
- Year Built Original construction year
- Year Remod Add Remodel date (same as construction date if no remodelling or additions).
- Exterior 1st Exterior covering on house
- Bsmt F in SF2 Type 2 finished square feet.
- Total Bsmt SF Total square feet of basement area.
- Sale Price To be predicted.