**Name: Ishaan Barde**

**Roll No: 116**

**M.SC.IT FINAL PROJECT**

**EXPERIMENTAL SET-UP:**

Setting up experiments for house price prediction involves defining the procedures for

* Data Pre-Processing
* Model Training
* Evaluation
* Comparison

**DATASET: RAW HOUSING DATASET**

The dataset has historical information of houses which are sold. It has information of several features such as Sale Price, Waterfront View, No. of Bedrooms/ Bathrooms, etc.

**DATA EXPLORATION & PRE-PROCESSING**

I have used

* EDA
* Data Cleaning
* Visualization
* Data Exploration & Transformation

**EXPLORATORY DATA ANALYSIS (EDA):**

EDA or **EXPLORATORY DATA ANALYSIS** is an approach that is used to analyse the data and discover trends, patterns, or check assumptions in data with the help of statistical summaries and graphical representations. In the project I have used all the types including Univariate, Bi-variate and Multi-Variate.

**Univariate Analysis** – In univariate analysis, we analyse or deal with only one variable at a time.

**Bi-Variate Analysis** – This type of data involves analysis two different variables.

**Multivariate Analysis** – The data involves three or more variables, it is categorized under multivariate.

**DATA CLEANING:**

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabelled.

**VISUALIZATION:**

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Additionally, it provides an excellent way for employees or business owners to present data to non-technical audiences without confusion. In the world of Big Data, data visualization tools and technologies are essential to analyse massive amounts of information and make data-driven decisions.

**DATA EXPLORATION AND TRANSFORMATION:**

**Data exploration** is an approach similar to initial data analysis, whereby a data analyst uses visual exploration to understand what is in a dataset and the characteristics of the data, rather than through traditional data management systems.

**Data transformation** is the process of converting data from one format to another, typically from the format of a source system into the required format of a destination system. Data transformation is a component of most data integration and data management tasks, such as data wrangling and data warehousing.

**MODEL TRAINING**

Model training in machine language is the process of feeding an ML algorithm with data to help identify and learn good values for all attributes involved. I have used

* Supervised Learning
* Deep Learning

**SUPERVISED LEARNING** also known as supervised machine learning, is a subcategory of machine learning and artificial intelligence. It is defined by its use of labelled datasets to train algorithms that to classify data or predict outcomes accurately.

**DEEP LEARNING** is a method in artificial intelligence (AI) that teaches computers to process data in a way that is inspired by the human brain. Deep learning models can recognize complex patterns in pictures, text, sounds, and other data to produce accurate insights and predictions.

**MODELS USED:**

* Linear Regression
* LSTM

**LINEAR REGRESSION**

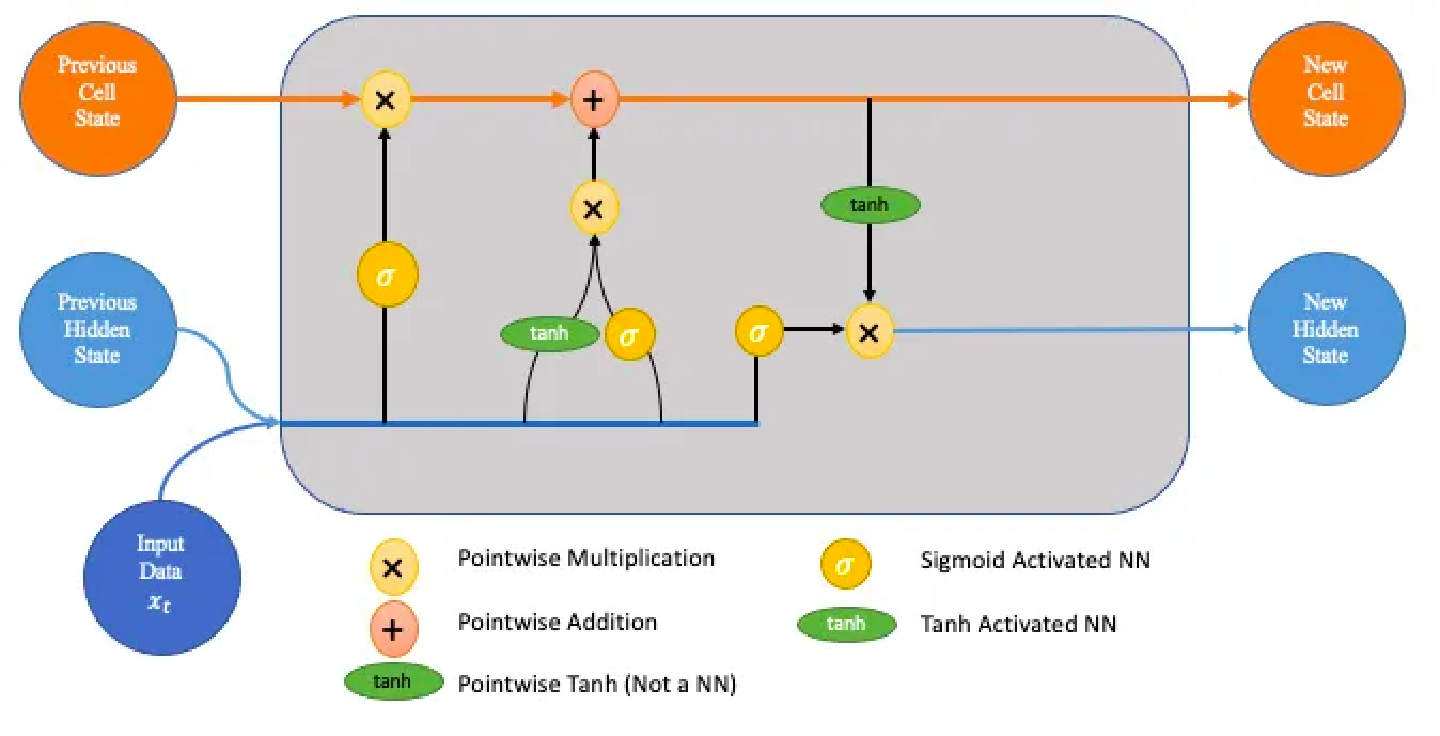
Linear regression analysis is used to predict the value of a variable based on the value of another variable.



The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

**LSTM (Long Short Term Memory Network)**

In concept, an LSTM recurrent unit tries to “remember” all the past knowledge that the network is seen so far and to “forget” irrelevant data. This is done by introducing different activation function layers called “gates” for different purposes.



Each LSTM recurrent unit also maintains a vector called the Internal Cell State which conceptually describes the information that was chosen to be retained by the previous LSTM recurrent unit. LSTM networks are the most commonly used variation of Recurrent Neural Networks (RNNs).

**MODEL EVALUATION**

Model evaluation is the process of using different evaluation metrics to understand a machine learning model's performance, as well as its strengths and weaknesses. Model evaluation is important to assess the efficacy of a model during initial research phases, and it also plays a role in model monitoring.

**COMPARISON STUDY**

Comparative studies are investigations to analyse and evaluate, with quantitative and qualitative methods, a phenomenon and/or facts among different areas, subjects, and/or objects to detect similarities and/or differences. In this I have compared the supervised machine learning algorithms with deep learning algorithms