**A**

**PROJECT REPORT**

**ON**

“**House Price Prediction By Advanced Regression Method On Data Science**”

# Submitted by

Student name Roll no

Omkar Dhapte 15

##### **Under the Guidance of**

### Prof. Abidali Shaikh

### (ME,CS)

##### ***In partial fulfillment of***

# Savitribai Phule, Pune University

# Department of Information Technology

At



PARVATIBAI GENBA MOZE COLLEGE OF ENGINEERING,

**Parvatibai Genba Moze College**

Department of Information Technology

## Wagholi-412207



***CERTIFICATE***

This is to certify that the project entitled “**House Price Prediction By Advanced Regression Method On Data Science”** submitted by **Omkar Sunil Dhapte** is a record of bonafide work carried out by him/her, in the partial fulfillment of the requirement for the award of Engineering in Information Technology at Parvatibai Genba Moze College, Wagholi under the Savitribai Phule Pune University This work is done during year 2021-2022, under our guidance.

------------------------ -------------------------------- ----------------------------

**Project Coordinator Project Guide HOD IT**

[Prof. Supriya More] [Prof. Abidali Shaikh] [Prof. Abidali Shaikh]

**Examiner Principal**

(External) (PGMCOE)

Date:

Place: Wagholi

**Acknowledgement**

I am profoundly grateful to Twinkle Shuklafor her expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion.

I would like to express my deepest appreciation towards **Prof. M.G.Jadhav** Principal, Parvatibai Genba Moze College ,**Prof. Abidali Shaikh**, HOD Information Technology Department whose invaluable guidance supported me in completing this project.

Student name Roll no

Mr. Omkar Dhapte 15

Date:

Place: Wagholi

**Engineering in Information Technology**

**List of Tables**

**Contents Page No.**

Risk Table 14

Risk probability Definition 15

Risk Impact Definition 15

Overview of Risk Mitigation, Monitoring, Management 17

Dissertation Schedule Table 32

Hardware and Software requirements 34

DFD Notations 39

Test Cases 47

**List of Diagrams**

**Contents Page No.**

Waterfall model 21

System Architecture 26

Task Network 30

DFD Diagram 40

Use case Diagram 42

Activity Diagram 43

**Abstract**

The increasing and decreasing occurrence of house prices changes from time to time. There are more reasons that effect the fluctuation of house prices. Some are build year, location, physical amenities, size of house etc. Predicting the value of house (Sale price) helps the customers to take right choice of buying the house. Machine learning is being adapted for various fields that could build prediction model and estimate the outcomes.

In this paper, we are contemplating the issue of rise and fall of house rates as a regression problem. Regression is a process that aims to predict the correlation between target dependent feature and a sequence of other changing independent features. In our experimental analysis, we are using Decision Tree Regression, Linear Regression, Ada Boost Regression, Gradient Boost Regression, Random Forest Regression techniques.

In addition, we are also used AutoML to predict the House sale price. AutoML is a system that takes labelled trained data as input and automatically build a suitable optimized model that the dataset fits.

**Keywords:**

House price prediction, Machine learning Decision Tree Regression, Random Forest Regression, Gradient Boost Regression, Ada Boost Regression, AutoML

**INDEX**

|  |  |  |
| --- | --- | --- |
|  | CONTENTS | PAGE NO. |

ACKNOWLEDGEMENT

List of Tables

List of Diagram

ABSTRACT

1. INTRODUCTION ……………………………………………………... 8-13

1.1 Motivation ……………………………………………………... 10

1.2 Need ……………………………………………………... 11

1.3 Literature Survey ……………………………………………………... 12

2. PROPOSED WORK …………………………………………………… 14-25

2.1 Problem Statement 15

2.2 Features 16

2.3 Risk Management 17

2.4 Life Cycle of the Project 23

2.5 Feasibility Study 24

3. SYSTEM DESCRIPTION ……………………………………………… 26-30

3.1 Data Retrieval 27

3.2 The Correlation between Variables Analysis 27

3.3 Data Preparation 28

3.4 Prediction with Advanced Regression Algorithm 28

3.5 Data Validation 28

3.6 Data Analysis 29

4. SCHEDULE OF WORK ………………………………………………. 31-33

4.1 Project Schedule 32

4.2 Task Network 32

4.3 Dissertation schedule table 34

5. PROJECT DESIGN …………………………………………………… 35-46

5.1 System Requirement 36

5.2 Number of People 39

5.3 Risk Management w.r.t NP- Hard Analysis 39

5.4 DFD Diagram 41

|  |  |
| --- | --- |
| CONTENTS | PAGE NO. |

5.5 Use Case 45

5.6 Activity Diagram 46

6. IMPLEMENTATION …………………………………………………. 47-60

6.1 Testing 48

6.2 Code Sample 51

6.3 Screenshots 56

7. CONCLUSION ………………………...………………………………. 61

8. REFERENCES ………………………………………………………… 62

1. **INTRODUCTION**
   1. **Introduction**

Real estate is not only the key sector of the national economy, but also one of the citizen’s major concerns. Due to the housing demands, people's attention to the housing price continues increasing. It is critical to provide accurate predictions of housing prices. Housing price is impacted by multiple factors including time and space, house ages, surrounding conditions, communities, transportation, etc.

Existing prediction models are usually single predictor ones, i.e., a single forecasting model are applied for the prediction. The prediction accuracy of this type model is not satisfactory when datasets are noisy. In this paper we have considered various intrinsic parameters (such as number of bedrooms, living area and construction material) and also external parameters (such as location, proximity, upcoming projects, etc.). Then we have applied these parameter values to four different machine learning algorithms i.e.

Random forest, Ridge Regression, Lasso Regression and Elastic net Regression and comparing them based on Root Mean Squared Error (RMSE). In addition to this we will also discuss the significance our approach and the methodology used.

All the experiments were run using the open-source software Python on a system with configurations 4 GB RAM Mac iOS 1.1 GHz Intel Core i5

* 1. **Motivation**

Create an effective price prediction model. Validate the model’s prediction accuracy .Identify the important home price attributes which feed the model’s predictive power. So, to maintain the transparency among customers and also the comparison can be made easy through this model. If customer finds the price of house at some given website higher than the price predicted by the model, so he can reject that house.The highest accuracy of 91% is from gradient boost regression. The customers and sellers will be benefitted with house price prediction.

* 1. **Need**

The increasing and decreasing occurrence of house prices changes from time to time. There are more reasons that effect the fluctuation of house prices. Some are build year,location, physical amenities, size of house etc. Predicting the value of house (Sale price) helps the customers to take right choice of buying the house. Machine learning is being adapted for various fields that could build prediction model and estimate the outcomes. In this paper, we are contemplating the issue of rise and fall of house rates as a regression problem. Regression is a process that aims to predict the correlation between target dependent feature and a sequence of other changing independent features.

* 1. **Literature Survey**

Actual cost of house is depending on so many factors like number of bathroom,bedroom,floors.

The house price grate with like near to highway, mall, super market, job opportunities, good educational facilities etc

the real estate companies trying to predict price of property by manually. Its giving price is not accurate than real price.

The different Machine Learning models like Linear Regression, Advanced Regression, Decision Tree and Random forest are used to build a predictive model

it was found that advanced regression had the best accuracy of 90% approx.

1. **PROPOSED WORKS**
   1. **Problem Statement**

House deals can be managed effectively with a combination of lifestyle changes and profit able in both partyes. With the right ways, the time of hosue losses can be reduced and the functioning of the claraity infomation. The predicted results can be used to prevent and thus reduce cost for government values and other expensive. The overall objective of my work will be to predict accurately with more information and attributes the presence of losses. Attributes considered form the primary basis for factors and give accurate results more or less. Many more input attributes can be taken but our goal is to predict with few attributes and faster efficiency the risk of having fanicacal losses. Decisions are often made based on borkers and experience rather than on the knowledge rich data hidden in the data set and databases.

This practice leads to unwanted biases, errors and excessive paper costs which affects the quality of service provided to provider. Data mining holds great potential for the real esate industry to enable real easte systems to systematically use data and analytics to identify inefficiencies and best dealers that improve care and reduce costs. The successful application of data mining in highly visible fields like e-business, marketing and retail has led to its application in other industries and sectors. Among these sectors just discovering is healthcare.

The healthcare environment is still „information rich‟ but „knowledge poor‟. There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in the data for African genres.

* 1. **Features**

The features of House Price Prediction Using Machine Learning are as follows.

* This Project will predict the Price of the Customer based on the Area and other general information using the datasets.
* This proposed system can try to use this data to create a model that tries to predict (reading data and data Exploration) if a Property has thisPriceor not. In this proposed system, using a Advanced regression (classification) algorithm we use the sklearn library to calculate the score.
* Finally, analysing the results with the help of Comparing Models and Confusion Matrix. From the data we are having, it is classified into different structured data based on the features of the land uses.
* From the availability of the data, we have to create a model that predicts the Property’s price using a Advanced regression algorithm.
* This is done based on the previous datasets of the brokes so after comparing it can provide up to 95% of accurate results, and the project is still developing further to get the 100% accurate results.
* With the help of price prediction, it can predict the Price of the Property and can solve various problems and prevents from various aspects.
* It provides security for the system so that no one can break into that and no one can make any changes in the system.
* The Price is predicted using the algorithms and the user has to enter the factors from the given drop-down menu, in order to get correct accuracy, the user has to enter all the factors.
* Here we can easily prepare the data and transform that data into algorithm, which will reduce the overall work of the project.
  1. **Risk Management**

The risks for the Project can be analysed within the constraints of time and quality

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ID** | **Risk Description** | **Probability** | **Impact** | | |  |
|  |  |  |  |
| **Schedule** | **Quality** | **Overall** |  |
|  |  |  |  |
|  |  |  |  |  |  |  |
| 1 | Domain knowledge | Low | Low | High | High |  |
| 2 | Technology will not Meet Expectations | Low | High | High | High |  |
| 3 | Lack of Development Experience | Medium | High | High | High |  |
| 4 | Poor Quality Documentation | Low | Low | Low | Low |  |
| 5 | Deviation from Software Engineering Standards | High | Low | High | High |  |
| 6 | Poor Comments in Code | Low | Low | Medium | Medium |  |
| 7 | Changes in Requirements | Medium | High | High | High |  |
|  |  |  |  |  |  |  |

**Table: Risk Table**

|  |  |  |
| --- | --- | --- |
| **Probability** | **Value** | **Description** |
|  |  |  |
| High | Probability of occurrence is | > 75% |
|  |  |  |
| Medium | Probability of occurrence is | 26 -75% |
| Low | Probability of occurrence is | < 25% |
|  |  |  |

**Table: Risk Probability definitions**

|  |  |  |  |
| --- | --- | --- | --- |
| **Impact** | **Value** | | **Description** |
|  |  |  |  |
| Very high | > 10% |  | Schedule impact or Unacceptable quality |
|  |  |  |  |
| High | 5 -10% |  | Schedule impact or Some parts of the project have low |
|  |  |  | Quality |
|  |  |  |  |
| Medium | < 5% |  | Schedule impact or barely noticeable degradation in quality Low Impact on schedule or Quality can be incorporated |
|  |  |  |  |
|  |  | **Table: Risk Impact definitions** | |

* + 1. **Risk Mitigation, Monitoring, Management**

A risk management strategy can be defined as a software project plan or the risk management steps. It can be organized into a separate Risk Mitigation, Monitoring and Management Plan. The RMMM plan documents all work performed as part of risk analysis and is used by the project manager as part of the overall project plan.

Teams do not develop a formal RMMM document. Rather, each risk is documented individually using a risk information sheet. In most cases, the RIS is maintained using a database system, so that creation and information entry, priority ordering, searches, and other analysis may be accomplished easily.

Once RMMM has been documented and the project has begun, risk mitigation and monitoring steps commence. As we have already discussed, risk mitigation is a problem avoidance activity. Risk monitoring is a project tracking activity with three primary objectives:

(1) to assess whether predicted risks occur.

(2) to ensure that risk aversion steps defined for the risk are being properly applied; and

(3) to collect information that can be used for future risk analysis.

Effective strategy must consider three issues:

* risk avoidance
* risk monitoring
* risk management and contingency planning. Proactive approach to risk – avoidance strategy. Develop risk mitigation plan. Develop a strategy to mitigate this risk for reducing turnover. Meet with current staff to determine causes for turnover. Mitigate those causes that are under our control before the project starts.
* Organize project teams so that information about each development activity is widely dispersed.
* Define documentation standards and establish mechanisms to be sure that documents are developed in a timely manner. Project manager monitors for likelihood of risk, Project manager should monitor the effectiveness of risk mitigation steps. Risk management and contingency planning assumes that mitigation efforts have failed and that the risk has become a reality. RMMM steps incur additional project cost.

**2.3.2 THE RMMM PLAN**

Risk Mitigation, Monitoring and Management Plan (RMMM) – documents all work performed as part of risk analysis and is used by the project manager as part of the overall project plan.RIS is maintained using a database system, so that creation and information entry, priority ordering, searches, and other analysis may be accomplished easily. Risk monitoring is a project tracking activity

Three primary objectives:

* assess whether predicted risks do, in fact, occur
* ensure that risk aversion steps defined for the risk are being properly applied
* collect information that can be used for future risk analysis.

Following are the details for each risk.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Risk** | **Category** | **Probability** | **Impact** | **RMMM plan (Solution)** | |
| Computer Crash | Technical Issue | 59% | High | S2 |
| Late Delivery | Business issue | 38 % | Medium | S6 |
| Technology will not Meet Expectations | Technology risk | 25 % | Low | S1, S5 |
| End Users Resist System | Business issue | 30 % | Low | S2, S4 |
| Changes in Requirements | Product size risk | 40 % | Medium | S6 |
| Lack of Development Experience | Technical Issue | 20 % | Low | S1 |
| Lack of Database Stability | Technical Issue | 60 % | High | S5, S1 |
| Poor Quality Documentation | Business issue | 35 % | Medium | S5 |
| Deviation from Software Engineering Standards | Process risk | 12 % | Low | S3 |

**Table: Overview of Risk Mitigation, Monitoring, Management**

**Solution 1:**

When working on the product or documentation, the staff member should always be aware of the stability of the computing environment they’re working in. Any changes in the stability of the environment should be recognized and taken seriously.

**Solution 2:**

The schedule will be followed closely during all development stages. Steps have been taken to ensure a timely delivery by gauging the scope of project based on the delivery deadline.

**Solution 3:**

The meetings with the customer should ensure that the customer and our organization understand each other and the requirements for the product. Should the development team come to the realization that their idea of the product specifications differs from those of the customer, the customer should be immediately notified and whatever steps necessary to rectify this problem should be done

**Solution 4:**

In order to prevent this from happening, the software will be developed with the end user in mind. The user-interface will be designed in a way to make use of the program convenient and pleasurable.

**Solution 5:**

In order to prevent this from happening, meetings (formal and informal) will be held with the customer on a routine business. This ensures that the product we are producing and the requirements of the customer are equivalent.

**Solution 6:**

Poor code commenting can be minimized if commenting standards are better expressed. While standards have been discussed informally, no formal standard yet exists. A formal written standard must be established to ensure quality of comments in all code.

**2.3.3 Risk Identiﬁcation**

Risk identiﬁcation is the process of determining risks that could potentially preven theprogram, enterprise, or investment from achieving its objectives. It includes documenting and communicating the concern. The risk identiﬁcation function should not be left to chance but should be explicitly covered in a number of project documents:

• Statement of work (SOW)

• Work breakdown structure (WBS)

• Budget

• Schedule

• Acquisition plan, and

• Execution plan

**2.3.4 Risk Analysis**

Risk analysis is the process of deﬁning and analysing the dangers to individuals, business and government agencies posed by potential natural and human-caused averments events. In IT, a risk analysis report can be used to align technology-related objectives with a company’s business objection. Risk Analysis is a process that helps you identify and manage potential problems that could undermine key business initiatives or projects.

To carry out a Risk Analysis, you must ﬁrst identify the possible threats that you face, and then estimate the likelihood that these threats will materialize.

Risk Analysis can be complex, as you’ll need to draw on detailed information such as project plans, ﬁnancial data, security protocols, marketing forecasts, and other relevant information. However, it’s an essential planning tool, and one that could save time, money, and reputations. To carry out a risk analysis, follow these steps:

1. Identify Threats

2. Estimate Risk

**2.3.5 Implementation Constraint**

* Login of user
* Taking user inputs and displaying message, heartPriceor not
* Show dashboard of user

**2.4 Life cycle of project**

Waterfall model is sequential approach, where each fundamental activity of a process represented as a separate phase, arranged in linear order.

In a waterfall model, you must plan and schedule of all the activities before starting, working on them. The waterfall model is a sequential design process, used in a software development process on which progress is seen as flowing steadily downwards through the phases of Conception, Initiation, Analysis, Design, Construction, Testing, Production/Implementation and Maintenance.

The waterfall development model originates in the manufacturing and construction industries: highly structures physical environments in which after-the-fact changes are prohibitively costly, if not impossible. Since no formal software development methodologies existed at the time, this hardware-oriented model was simply adapted for software development.

**Diagram:**

**2.5 Feasibility Study:**

In Feasibility study the assign project is analysed, information about project participants is collected, and the requirements for the system are gathered and analysed. During the Feasibility Study stage, the project’s goals, parameters and restraints are agreed and a conceptual problem solution is prepared. Three key considerations are involved in the feasibility analysis.

The feasibility study comprise of an initial investigation into personnel will be required. Feasibility study will help you make informed and transparent decisions at crucial points during the developmental process. All projects are feasible given unlimited times and resources. Unfortunately, the development of computer-based system is more likely to be plagued to scarcity of resources.

**2.5.1 Economic Feasibility:**

The procedure is to determine the benefits and saving that are expected from a candidate and compare them with costs. If benefits outweigh costs, then the decision is made to design and implement the system. Otherwise, future justification in proposed system will have to be made if it is to have a chance of being approved.

Economic feasibility looks at the financial aspects of the project. Economic feasibility concerns with the returns from the investments in a project. It determines whether it is worthwhile to invest the money in the proposed system. In our analysis during the economic feasibility, we have found that once this project is implemented successfully it will bring the below benefits to the organization:

1. The proposed system automatically generates the arguments.

2. The system provides accurate results.

3. System extracts active as well as passive sentences from knowledge base.

**2.5.2 Technical Feasibility:**

Technical feasibility centre of the existing computer system (hardware, software, etc.) and to what extent it can support the proposed addition. If the budget is a serious constraint, then the project is judged is not feasible.We have done technical feasibility for this project by identifying the inputs required, output generated and the procedures required to generate the required output which are mentioned below:

* Input Required; input voice
* Output Generated: Device.

**2.5.3 Operational feasibility:**

People are inherently resistant to change, and computers have been known to facilitated change. It is understandable that the introduction of a candidate. System requires special efforts to educate, sell and train the staff on new ways of conducting business.

**2.5.4 Performance Feasibility:**

The system provides a user-friendly graphical interface where the user does not need to know the technical requirements hence it’s very easy to operate the system. Without much training proved the any user could be able to use it very effectively.

**2.5.5 Conclusion drawn from feasibility study:**

As a whole, it has been concluded that this system will help the organization to have a higher level of performance. Hence, the system is feasible.

**Major Constraints**

* System should recognize to any voices in his lists without any fault.
* With ideal conditions, system response should be fast and error free.
* System performance shall not decrease with time or by usage.
* Performance and speed should not depend on newer or older mobile, but that depends on mobile phone capable of running android programs or not.

1. **SYSTEM DESCRIPTION**
   1. **Data Retrieval**

The First process is Data Retrieval. In this process, HeartPriceUCI Dataset -published by Ronit in Kaggle website (https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data)- will be used. It will be imported into the PyCharm software. The data obtained are categorical data and numerical data. The data in this study contain 14 variables with 76 attributes and 304 responses as the basis for analysis. First variable is MSSubClass with units in Rupess (MSSubClass:). Second, the feet with value one means x driesction and value 0 means y-direction (FeetSqu).

Third, the variable type of floor(st). Fourth, the variable age of floot in 0 at buyer land to x-land user (FlAge). Fifth, area it located Agriculture(a),Commercial(c) Floating Village Residential(FVL), Industrial Residential(IR), High Density Residential(HDR), Low Density Residential(LDR), Low Density Park Residential(LDPR).Sixth Rang of rong in Medium Densit Linear feet of street connected to property Lot size in square feet: Type of road access to property Grvl Pave so more factors in there more than 54 so in future there more factors include make accurucy

* 1. **The Correlation between Variables Analysis**

Besides, to facilitate data analysis, all variables inthe imported dataset will be visualized in the form ofa histogram to facilitate the reading of the data ingeneral. In the process, Analyze the Correlationbetween Variables; the correlation between variablesis examined to prove that the method to be used is theadvanced regression model is the right model.Relationships between variables in the availabledataset will be plotted in the form of a matrix. This isalso done to check whether there is multicollinearitybetween variables in the dataset.

Database

* 1. **Data Preparation**

The dataset imported in PyCharm will be dividedinto two parts, namely training data and testing data.Training data is used as a basis for building models.Meanwhile, testing data is used as a basis for testingor validating the model. In this data preparationprocess, 293 data will be sampled. Then the data willbe partitioned into train data and test data.

* 1. **Prediction with Advanced Regression Algorithm**

In this process, the data that has been partitioned inthe previous process will be used. Prediction using linear regression method will produce several datathat can be used as a basis for concluding tomake predictions.

* 1. **Data Validation**

The technique used to validate the results is themethod of the confusion matrix and K-fold cross validation with 10-fold. By using a confusion matrix,the accuracy of the use of the advanced regressionmodel can be known. Besides, the use of the K-foldcross-validation method¸ produces values of errorsthat may occur when using a advancedregression model.

* 1. **Data Analysis**

The dataset obtained by the researcher as a basis for analysis is imported into PyCharm. The data retrieval process is also performed in the data visualization to see the value of each variable involved in the overall research analysis.

In this process, the correlation between variables will be examined, which will be used as a basis for analysis to predict land area also. Based on the matrix in Fig. 4, it was found that the variables induced First variable is MSSubClass with units in Rupess (MSSubClass:). Second, the feet with value one means x driesction and value 0 means y-direction (FeetSqu). Third, the variable type of floor(st). Fourth, the variable age of floot in 0 at buyer land to x-land user (FlAge). Fifth, area it located Agriculture(a),Commercial(c) Floating Village Residential(FVL), Industrial Residential(IR), High Density Residential(HDR), Low Density Residential(LDR), Low Density Park Residential(LDPR).Sixth Rang of rong in Medium Densit Linear feet of street connected to property Lot size in square feet: Type of road access to property Grvl Pave so more factors in there more than 54 so in future there more factors include make accurucy .

Data that has been imported will be taken as many as 293 random data as a basis for analysis. The data is divided into train data and test data. The data shown in Fig. 5 on the next page is data from train\_data and test\_data that will be used in this study. The training data is used to build a advanced regression model using the glm () function because advanced regression is included in the generalized linear model with binomial type families. Based on the results of using the advanced regression method, it is predicted that the variables influence the target variable at an alpha value of 5% significantly. The selected variables are the variables that significantly affect the target variable. In advanced regression, the effect of each variable on the target variable can be seen from the odds ratio value.

For example, for the age variable having a coefficient value of -1.547601 with a reference category with a male value, the odds ratio value is 4.2655 which means that for new build Propertys, the odds of getting deals 4.2655 times the x owner odds or it can be said the tendency of men to deals higher than women. For the price variable with a coefficient value of -0.029713, it is found that the odds ratio value is 0.0822 which means that for the properties age variable there will be a significant increase when properties age enters the value 0.0822 mmHg. On the other hand, the properties age with uses variable with a coefficient of 0.032028 will have an odd of 0.08856 which means that at that value there will be a significant change in the performance of the heart rate or cardiovascular rate.

The exang1 variable is exercise-induced angina with an estimated coefficient of -1.05855 so that the exang variable with a reference value of 1 will have an odd of 2.92710 which means that if the value is achieved then cardiovascular performance will decrease. Next is the variable ca with reference ca values 1, 2, and 3. Ca1 with an estimated coefficient of - 1.430110 will have odds of 3.955, while ca2 with an estimated ratio of -3.329874 will have odds of 9.1777 and ca3 with an estimated factor of -0.553711 will have odds in the amount of 1.5261. It proves that when the number of fluoroscopy vessels reaches its value odds, this will have an impact on decreasing cardiac performance which will affect the increased potential for cardiovascular disease. Besides, the composition of value 0 and value 1 on variable target is 97:116, which is still fairly balance, so the result will be reliable and free from any imbalanced data set problems.

1. **SCHEDULE OF WORK**

**4.1 Project Schedule:**

* + 1. Project task set in the system following main tasks are considered:

1. Task 1(T1): Proposal of the system defining system flow and gathering of all information related to it.

2. Task 2(T2): Gathering information about similar approaches

3. Task 3(T3): Mathematical model design for proposed system Mathematical model design for verification of database transaction using set theory

4. Task 4(T4): UML diagrams formation, such as class dig., sequence dig., etc.

5. Task 5(T5): Installing all required software All required software will be installed and classes will be defined

6. Task 6(T6): Preparation of input file Different inputs required are stored

7. Task 7(T7): Designing User Interface Task

8. Task 8(T8): Implementation of system

9. Task 9(T9): Analysis of the user interface and system

10. Task 10(T10): Corrective actions to be made in system.

**4.2System Engineering:**

As this project is implemented by a group of 4-person, sequential designing method is used. Waterfall model for software development is used.

* + 1. **Task network:**

Here are the project tasks and their dependencies in this diagrammatic form.

**T1: Communication**

Software development process starts with the communication between customer and developer. According to need of project, we gathered the requirements related to project.Project requirements are collected in this activity. This framework activity is the main focus of the project managers and stakeholders. This framework activity includes communication and coordination with the clients.

**T2: Planning**

This framework action incorporates data about the technical work to be planned, risks to be faced, resources needed for task completion, the decision of the milestone deadline to release the product in production.

**T3: Risk Management**

It includes identifying the risks during project development and managing the risks which are affecting the project development. Risk management means risk containment and mitigation. First, you’ve got to identify and plan. Then be ready to act when a risk arises, drawing upon the experience and knowledge of the entire team to minimize the impact to the project.

**T4: Designing**

It Includes designing the modules and internal, external interfaces of the entire system. Software design is the process of envisioning and defining software solutions to one or more sets of problems. One of the main components of software design is the software requirements analysis (SRA).

**T5: Modelling**

It includes detailed requirement analysis and project design. Modelling is the process of developing abstract models of a system, with each model presenting a different view or perspective of that system.

**T6: Coding**

It includes actual coding of the entire project. In this framework activity, the actual development of the product begins with code construction and then testing to fix errors and issues.

**T7: Testing**

Here all the modules are tested and bugs are discovered. The bugs are then corrected and entire system is tested thoroughly.  Testing is an investigation conducted to provide stakeholders with information about the quality of the software product or service under test.

* 1. **Dissertation Schedule table:**

|  |  |
| --- | --- |
| **Phases** | **Months** |
| Requirement Gathering | December 2021 |
| Specification | December 2021 |
| Literature survey | December 2021, January 2022 |
| Design Analysis | December 2021, January 2022 |
| Implementation/Coding | January 2022 |
| Testing | January 2022 |
| Deployment/Result Analysis | January 2022 |
| Documentation | February 2022 |

1. **PROJECT DESIGN**
2. **System requirement:**
3. **Hardware Requirement**

|  |  |
| --- | --- |
| **Name** | **Details** |
| Processor | Intel i3 Processor |
| RAM | 4Gb and above |
| Hard Drive | 500GB |

1. **Software Requirement**

|  |  |
| --- | --- |
| **Name** | **Details** |
| Operating System | Windows 10 |
| Software Used | PyCharm, Visual Studio |
| Technology Used | Python, Flutter, Firebase |

**Python:**

* Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales. Van Rossum led the language community until stepping down as leader in July 2018.
* Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented,imperative, functional and procedural, and has a large and comprehensive standard library.
* Python interpreters are available for many operating systems. CPython, the reference implementation of Python, is open-source software and has a community-based development model, as do nearly all of Python's other implementations. Python and CPython are managed by the non-profit Python Software Foundation.
* Python is a multi-paradigm programming language. Object-oriented programming and structured programming are fully supported, and many of its features support functional programming and aspect-oriented programming (including by metaprogramming and meta objects (magic methods)). Many other paradigms are supported via extensions, including design by contract and logic programming.
* Python uses dynamic typing, and a combination of reference counting and a cycle-detecting garbage collector for memory management. It also features dynamic name resolution (late binding), which binds method and variable names during program execution.
* Python's design offers some support for functional programming in the Lisp tradition. It has filter(), map(), and reduce() functions; list comprehensions, dictionaries, and sets; and generator expressions. The standard library has two modules (itertools and functools) that implement functional tools borrowed from Haskell and Standard ML.
* The language's core philosophy is summarized in the document The Zen of Python (PEP 20), which includes aphorisms such as:
* Beautiful is better than ugly
* Explicit is better than implicit
* Simple is better than complex
* Complex is better than complicated
* Readability counts

Rather than having all of its functionality built into its core, Python was designed to be highly [extensible](https://en.wikipedia.org/wiki/Extensibility). This compact modularity has made it particularly popular as a means of adding programmable interfaces to existing applications. Van Rossum's vision of a small core language with a large standard library and easily extensible interpreter stemmed from his frustrations with [ABC](https://en.wikipedia.org/wiki/ABC_(programming_language)), which espoused the opposite approach.

While offering choice in coding methodology, the Python philosophy rejects exuberant syntax (such as that of [Perl](https://en.wikipedia.org/wiki/Perl)) in favor of a simpler, less-cluttered grammar. As [Alex Martelli](https://en.wikipedia.org/wiki/Alex_Martelli) put it: "To describe something as 'clever' is not considered a compliment in the Python culture." Python's philosophy rejects the Perl "[there is more than one way to do it](https://en.wikipedia.org/wiki/There_is_more_than_one_way_to_do_it)" approach to language design in favor of "there should be one—and preferably only one—obvious way to do it".

Python's developers strive to avoid [premature optimization](https://en.wikipedia.org/wiki/Premature_optimization), and reject patches to non-critical parts of the [CPython](https://en.wikipedia.org/wiki/CPython) reference implementation that would offer marginal increases in speed at the cost of clarity. When speed is important, a Python programmer can move time-critical functions to extension modules written in languages such as C, or use [Py](https://en.wikipedia.org/wiki/PyPy), a [just-in-time compiler](https://en.wikipedia.org/wiki/Just-in-time_compilation). [CPython](https://en.wikipedia.org/wiki/Cython) is also available, which translates a Python script into C and makes direct C-level API calls into the Python interpreter.

An important goal of Python's developers is keeping it fun to use. This is reflected in the language's name—a tribute to the British comedy group [Monty Python](https://en.wikipedia.org/wiki/Monty_Python)—and in occasionally playful approaches to tutorials and reference materials, such as examples that refer to spam and eggs (from a [famous Monty Python sketch](https://en.wikipedia.org/wiki/Spam_(Monty_Python))) instead of the standard [foo and bar](https://en.wikipedia.org/wiki/Foobar).

Users and admirers of Python, especially those considered knowledgeable or experienced, are often referred to as Pythonists, Pythonistas, and Pythoneers.

**ADVANCED REGRESSION:**

Advanced regression is a predictive model used toevaluate the relationship between the dependentvariable (target) which is categorical data withnominal or ordinal scale and the independent variable(predictor) which is categorical data with interval orratio scale. This algorithm can also be used for timeseries modelling to find the relationship between thevariables involved. Advanced regression is an algorithmused to predict the probability of categoricaldependent variables. In advanced regression, the dependent variable is shown as a binary variable thatis valued at 1 (yes) Or 0 (no). The advanced regressionmodel predicts as a function of X. The assumptionsused in Advanced regression are as follows: binaryadvanced regression requires binary dependent variables,for binary regression, the factor 1 level of the

dependent variable must represent the desired result,independent variables must be independent of eachother. In this case, the model must have little or nomulticollinearity and be linearly related to logopportunities.Advanced regression used appropriate regressionanalysis to be performed when the dependent variableis dichotomous (binary). Advanced regression acts as apredictive analytical model.

Advanced regression isapplied to describe data and explain the relationshipbetween one dependent binary variable with one ormore independent variables at the nominal, ordinal,interval or ratio level. Advanced regression has severaladvantages and disadvantages. The benefits of advancedregression include the following. First, advancedregression can show a significant relationship betweenthe dependent variable and the independent variable.Second, advanced regression analysis can also be used tocompare the effect of variables measured at differentscales including the effect of price changes and thenumber of promotional activities.

This benefit helpsmarket researchers or data analysts to eliminate andevaluate the best set of variables that will be used tobuild predictive models. Third, the advanced regressionmodel is not only a classification model, but alsoprovides information related to probability. To achievea better result using Advanced Regression, first all

independent variable must contain their valid value.Secondly, advanced regression works well for predictingcategorical results and multinomial results. Third,there is no multicollinearity between variables in thedataset.

**5.2 Number of People**

Group of 4 people is required to complete the project with given time span successful. According to recommendation of person for development and development time consideration as number of persons of development is 4, so it leads to increase the development time from 2 months to 3 months.

**5.3 Risk Management w.r.t. NP Hard analysis:**

This section discusses Project risks and the approach to managing them.

5.3.1 Time Complexity

1. P:

If the running time is some polynomial function of size of the input, for instance is the algorithm runs in linear time or quadratic time or cubic time, then we can say that the algorithm runs in polynomial time and the problem it solves in class P.

2. NP:

Now there are a lot of programs that do not run-in polynomial time on regular computer, not only this but there are a few that do not run-in polynomial time on non-deterministic Turing machine as well. These programs have a tendency to solve problems in NP that stands for nondeterministic polynomial time. A nondeterministic Turing machine can do everything in regular computer can and more. This means that all problems in P also occur in NP. An equivalent way to define NP is by pointing to the problem that can be verified in polynomial time. This means that there does not necessarily have to be a way to solve the problem in polynomial time way to find solution, but once you have solution it only takes polynomial time to verify that it is correct.

3. NP-HARD:

A lot of times you can solve problem by reducing it to a different problem. I can reduce Problem B to problem A if, given a solution to problem A, I can easily construct a solution to Problem B, (In this case, “easily” meantime). If a problem is NP-HARD, this means i can reduce any problem in NP to that problem. This means if i can solve that problem, I can easily solve any problem in NP. If we could solve an NP-Hard problem in polynomial time, this would prove that P=NP. 8.1.4 NP-complete A problem is NP-complete if the problem isBoth NP-hard and in NP. Because the algorithms used to carry out our problemsStatement take polynomial time to execute, out problems can be classified as P.

**5.3.1Risk Identification**

For risks identification, review of scope document, requirements specifications and schedule are done. Answers to questionnaire revealed some risks. Each risk is categorized.

You can referee following risk identification questionnaire.

1. Have top software and customer managers formally committed to support the project?
2. Are end-users enthusiastically committed to the project and the system/product to be built?
3. Are requirements fully understood by the software engineering team and its customers?
4. Have customers been involved fully in the definition of requirements?
5. Do end-users have realistic expectations?
6. Does the software engineering team have the right mix of skills?
7. Are project requirements stable?
8. Is the number of people on the project team adequate to do the job?
9. Do all customer/user constituencies agree on the importance of the project and on the requirements for the system/product to be built?
   1. **DFD Diagram:**

DFD allows the software development team to depict flow of data from one process to another. In addition, the DFD accomplishes the following objectives:

* It represents system data in a hierarchical manner and with required levels of detail.
* It depicts processes according to defined user requirements and software scope.

A DFD depicts the flow of data within a system and considers a system as a transformation function that transforms the given inputs into desired outputs. When there is complexity in a system, data needs to be transformed using various steps to produce an output. These steps are required to refine the information. The objective of DFD is to provide an overview of transformations that occur in the input data within the system in order to produce an output.

The DFD consists of four basic notations (symbols), which help to depict notations in a system. These notations are:

|  |  |  |
| --- | --- | --- |
| Name | Notations | Description |
| External Entity |  | Represents the source or destination of data within the system. Each external entity is identified with a meaningful and unique name. |
| Data Flow |  | Represents the movement of data from its source to destination within the system. |
| Data Store |  | Indicates the place for storing information within the system. |
| Process |  | Shows a transformation or manipulation of data within the system. A process is also known as a bubble. |

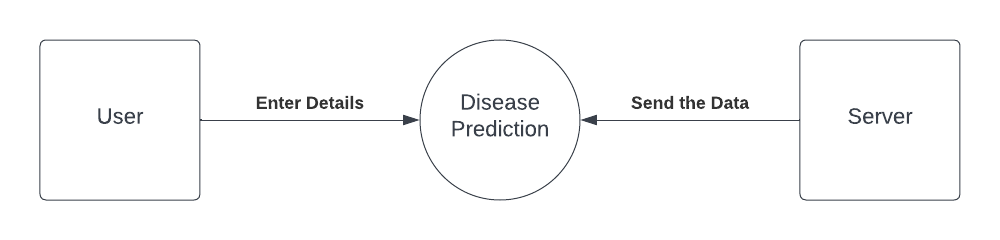
**T**able 6: DFD notations

There are various levels of DFD, which provides details about the input, processes and output of the system. Note that the level of detail of processes increases with increase in levels. However, these levels do not describe the system’s internal structure or behaviour. These levels are listed below:

* Level 0 DFD: this sho5.5ws an overall view of the system.
* Level 1 DFD: this elaborates the level 0 DFD and splits the process into detailed form.
* Level 2 DFD: this elaborates level 1 DFD and displays the processes in detailed form.
* Level 3 DFD: this elaborates level 2 DFD and displays the processes in detailed form.

5.4.1 DFD Level 0

The system designer makes context level DFD or Level 0, which shows only the interaction between system and system environment.



**5.4.2 DFD Level 1**

This elaborates the level 0 DFD and splits the process into detailed form.

Server

User

Input

Details

Send the details

**5.4.3 DFD Level 2**

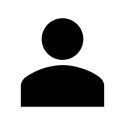
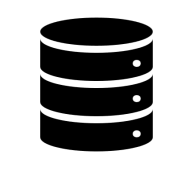
This elaborates level 1 DFD and displays the processes in detailed form.

Server

User

**5.5 Use case Diagram**

A use case diagram at its simplest is a representation of a user's interaction with the system that shows the relationship between the user and the different **use** cases in which the user is involved

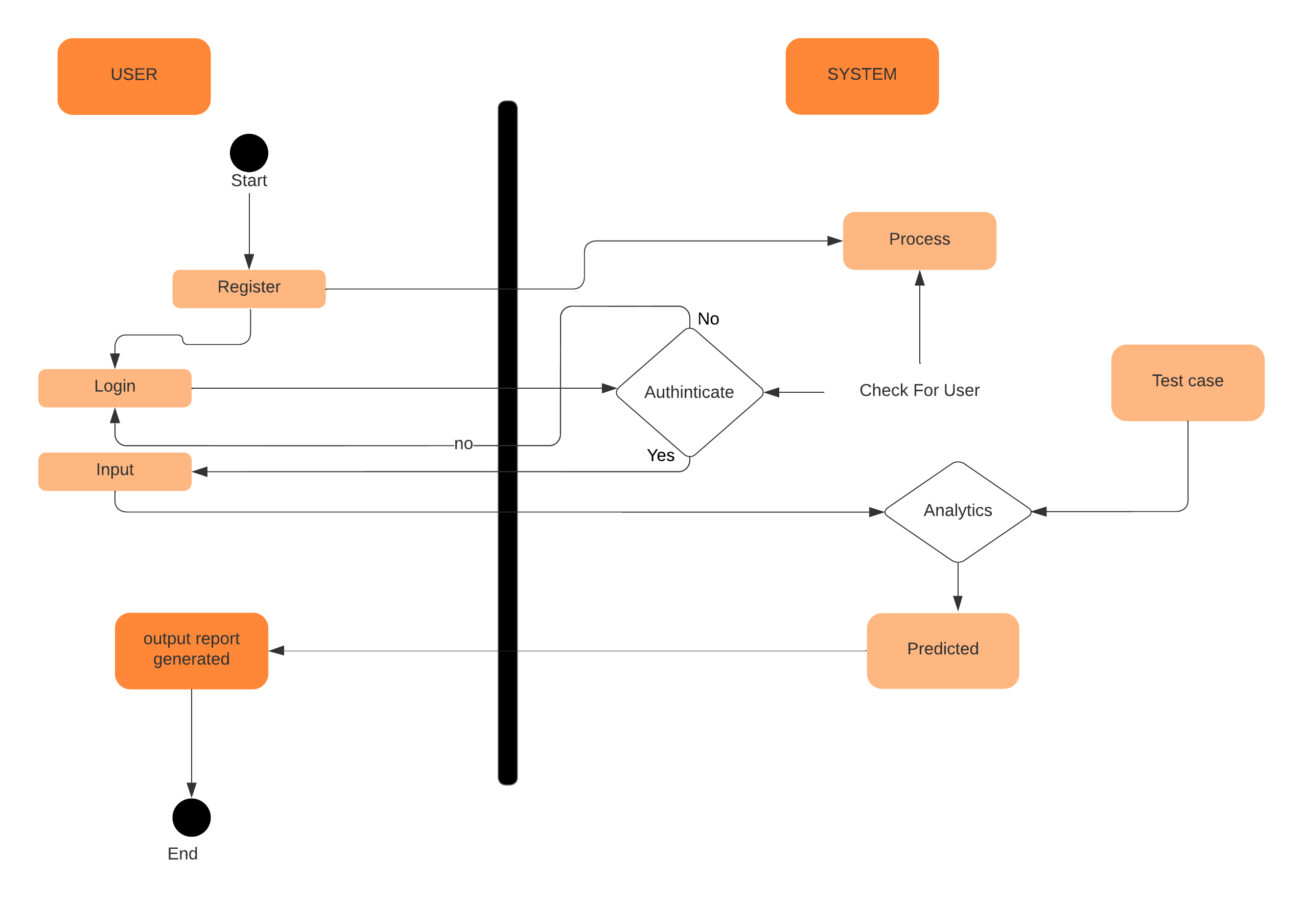
In this diagram, the user can perform multiple actions such as authentication, registering device, training, viewing remote control and testing. These actions lead to further breakups into other multiple functions provided in the system.

**5.6 Activity Diagram**

Activity diagram describes the flow of control in a system. So, it consists of activities and links. The flow can be sequential, concurrent or branched.

Activities are nothing but the functions of a system. Numbers of activity diagrams are prepared to capture the entire flow in a system.

Activity diagrams are used to visualize the flow of controls in a system. This is prepared to have an idea of how the system will work when executed.



1. **IMPLEMENTATION**
2. **Testing:**

Test Plan

Test Plan Identifier: Iterative filtering for Aggregation and trust assessment.

1. Purpose of the Test Plan Document

The main purpose of this document is to fit a particular project’s needs. It documents and tracks the necessary information required to effectively define the approach to be used in the testing of the project’s product. The Test Plan document is created during the Planning Phase of the project. Its intended audience is the project manager, project team, and testing team.

1. Objective of Test Panning

To find as many defects as possible and get them fix.

1. Items to be Tested OR Not to be Tested

Describe the items/features/functions to be tested that are within the scope of this test plan. Include a description of how they will be tested, when, by whom, and to what quality standards. Also include a description of those items agreed not to be tested.

1. Items to be tested

- Overall functionality of the application

* User Interface of the application

1. Not to be Testes

- Performance of the application

1. Test Approach

Describe the overall testing approach to be used to test the projects product. Provide an outline of any planned tests. There are many approaches like:

* Black Box Testing
* White Box Testing

Here we used Black Box Testing approach. In Black Box Testing we just give input to the system and check its output without checking how system processes it.

1. Test Pass OR Test Fail Criteria

When actual and expected results are same then test will be passed. When actual and expected results are different then test will be failed.

1. Test Entry OR Exit Criteria

Describe the entry and exit criteria used to start testing and determine when to stop testing.

* Entry criteria: As soon as we have requirement, we can start testing.

- Exit criteria: When bug rate falls below certain level, we can stop testing

1. Test Suspension OR Resumption Criteria describe the suspension criteria that may be used to suspend all or portions of testing. Also describe the resumption criteria that may be used to resume testing.

- Suspension criteria: if there is large change in application like change in requirements, we can suspend work for some time.

- Resumption criteria: after resolving the respective problem we can resume work.

1. Testing Type

It describes which testing types we are going to follow in our testing life cycle.

Here we are using:

* Black Box Testing
* Functional Testing
* UI Testing
* Integration Testing

|  |  |
| --- | --- |
| Risk ID | 1 |
| Risk Description | Wrong Input |
| Category | Run Time Environment |
| Source | This is identified during testing. |
| Probability | High |
| Impact | High |
| Response | Mitigate |
| Strategy | Testing will resolve this issue |
| Risk Status | Occurred |

Table: Test Case 1

|  |  |
| --- | --- |
| Risk ID | 2 |
| Risk Description | Server Error |
| Category | Run Time Environment |
| Source | Software Design Specification documentation review |
| Probability | High |
| Impact | Very High |
| Response | Mitigate |
| Strategy | Better deployment of network will resolve this issue. |
| Risk Status | Identified |

Table: Test Case 2

1. **Code Sample:**

**Frontend**

<!doctype html>

<html lang="en">

  <head>

    <!-- Required meta tags -->

    <meta charset="utf-8">

    <meta name="viewport" content="width=device-width, initial-scale=1, shrink-to-fit=no">

    <!-- Bootstrap CSS -->

    <link rel="stylesheet" href="https://stackpath.bootstrapcdn.com/bootstrap/4.3.1/css/bootstrap.min.css" integrity="sha384-ggOyR0iXCbMQv3Xipma34MD+dH/1fQ784/j6cY/iJTQUOhcWr7x9JvoRxT2MZw1T" crossorigin="anonymous">

    <link rel="stylesheet" href="../house\_price\_pridtion/css/style.css">

    <title>House Price Prediction</title>

  </head>

  <body class="bg-primary" >

    <div class="container">

      <div class="row">

        <div class="card"  >

          <div class="card-header">

              <h1>Bengaluru  House  Price  Prediction</h1>

          </div>

              <h6>want to predict the price of the house in Bengaluru? fill the below details!</h6>

              <div class="card-body">

                 <form method="post" accept-charset="utf-8">

                   <div class="row">

                     <div class="col-md-6 form-group" style="text-align: center " >

                       <label ><b style="color:white;">Select the location:</b></label>

                       <select style="border-radius:20px" class="selectpicker form-control" id="location" name="location"  required="1">

                          {% for location in locations %}

                            <option value="{{ location }}">{{ location }}</option>

                          {% endfor %}

                       </select>

                     </div>

                     <div class="col-md-6 form-group" style="text-align: center" >

                       <label><b style="color:white;" >Enter BHK:</b></label>

                       <input style="border-radius: 20px;" type="text" class="form-control" id="bhk" name="bhk" placeholder="Enter BHK">

                     </div>

                     <div class="col-md-6 form-group" style="text-align: center" >

                       <label><b style="color:white;">Enter the Number of Bathrooms:</b></label>

                       <input style="border-radius: 20px;" type="text" class="form-control" id="bath" name="bath" placeholder="Enter Number of Bathrooms">

                     </div>

                     <div class="col-md-6 form-group" style="text-align: center" >

                       <label><b style="color:white;">Enter Square feet:</b></label>

                       <input style="border-radius: 20px;" type="text" class="form-control" id="total\_sqft" name="total\_sqft" placeholder="Enter Square Feet">

                     </div>

                     <div class="col-md-12 form-group">

                       <button style="margin-left:448px; margin-top:30px; width: 200px; border-radius: 30px;" class="btn btn-primary btn-lg" onclick="send\_data()">Predict Price</button>

                     </div>

                   </div>

                 </form>

                <br>

                <div class="col-md-12 " style="text-align: center">

                  <h3><span id="prediction"> </span></h3>

                </div>

              </div>

            <br>

            <nav class="navbar navbar-expand-lg navbar-light bg-primary" style="border-radius: 40px;">

              <button class="navbar-toggler" type="button" data-toggle="collapse" data-target="#navbarSupportedContent" aria-controls="navbarSupportedContent" aria-expanded="false" aria-label="Toggle navigation">

                <span class="navbar-toggler-icon" style="border-radius: 40px;color :#000"></span>

              </button>

            </nav>

            </div>

    </div>

   <script>

            function form\_handler(event)

            {

                   event.preventDefault();

            }

            function send\_data()

            {

                    document.querySelector('form').addEventListener("submit",form\_handler);

                    var fd=new FormData(document.querySelector('form'));

                    var xhr= new XMLHttpRequest();

                    xhr.open('POST','/predict',true);

                    document.getElementById("prediction").innerHTML = "Wait Predicting Price...!";

                    xhr.onreadystatechange = function(){

                        if(xhr.readyState == XMLHttpRequest.DONE){

                            document.getElementById('prediction').innerHTML="Estimated Price is : ₹ "+xhr.responseText;

                        }

                    };

                    xhr.onload = function(){};

                    xhr.send(fd);

            }

   </script>

            </div>

        </div>

      </div>

    </div>

    <!-- Optional JavaScript -->

    <!-- jQuery first, then Popper.js, then Bootstrap JS -->

    <script src="https://code.jquery.com/jquery-3.3.1.slim.min.js" integrity="sha384-q8i/X+965DzO0rT7abK41JStQIAqVgRVzpbzo5smXKp4YfRvH+8abtTE1Pi6jizo" crossorigin="anonymous"></script>

    <script src="https://cdnjs.cloudflare.com/ajax/libs/popper.js/1.14.7/umd/popper.min.js" integrity="sha384-UO2eT0CpHqdSJQ6hJty5KVphtPhzWj9WO1clHTMGa3JDZwrnQq4sF86dIHNDz0W1" crossorigin="anonymous"></script>

    <script src="https://stackpath.bootstrapcdn.com/bootstrap/4.3.1/js/bootstrap.min.js" integrity="sha384-JjSmVgyd0p3pXB1rRibZUAYoIIy6OrQ6VrjIEaFf/nJGzIxFDsf4x0xIM+B07jRM" crossorigin="anonymous"></script>

  <br>

  </body>

</html>

**Home page:**

<!DOCTYPE html>

<html lang="en">

<head>

    <meta charset="UTF-8">

    <meta http-equiv="X-UA-Compatible" content="IE=edge">

    <meta name="viewport" content="width=device-width, initial-scale=1.0">

    <link rel="stylesheet" href="../house\_price\_pridtion/css/style.css">

    <title>Seaduck Properties</title>

</head>

<body>

    <div class="header">

        <div class="logo"><img class="logo-img"src="../house\_price\_pridtion/image/companyname.png"></div>

        <div class="navagation">

            <a class="nav" href="/house\_price\_pridtion/index.html">Home</a>

            <a class="nav" href="/house\_price\_pridtion/main.html">Pridiction</a>

            <a class="nav" href="https://www.kaggle.com/amitabhajoy/bengaluru-house-price-data">Reviews</a>

            <a class="nav" href="#">New Sells</a>

            <a class="nav" href="">Sign In</a>

        </div>

    </div>

    <div class="main">

        <div class="banner">

            <aside></aside>

            <div class="banner\_text">

                <span>Best Selling And Buyer Properties with 97.66 accurccy</span>

                <a class="overlay" href="#signup">Sell Now & Buy</a>

            </div>

        </div>

        <div class="main\_wrapper">

            <h1>Some House Selling</h1>

            <div class="main\_wrapper\_content">

                <div class="page\_compent">

                    <img src="./image/h.jpg">

                    <div class="page\_compent\_element">

                        <h6>Flat 3bhk</h6>

                        <h6>40.99L</h6>

                    </div>

                </div>

                <div class="page\_compent">

                    <img src="./image/h.jpg">

                    <div class="page\_compent\_element">

                        <h6>Flat 3bhk</h6>

                        <h6>40.99L</h6>

                    </div>

                </div>

                <div class="page\_compent">

                    <img src="./image/h.jpg">

                    <div class="page\_compent\_element">

                        <h6>Flat 3bhk</h6>

                        <h6>40.99L</h6>

                    </div>

                </div>

                <div class="page\_compent">

                    <img src="./image/h.jpg">

                    <div class="page\_compent\_element">

                        <h6>40.99L</h6>

                    </div>

                </div>

            </div>

        </div>

        <div class="user\_info">

            <div id ="#login"class="login\_interface">

                <form  action="login.php">

                    <h4> User Login</h4>

                    <input type="email" name="email"placeholder="Enter your Email Id">

                    <input type="password"name=" password" placeholder="Enter your Password">

                    <button type="submit">Login Here</button>

                </form>

            </div>

        </div>

    </div>

</body>

</html>

**Css Code:**

\*{

    box-sizing: border-box;

    margin: 0;

    padding: 0;

}

body{

    font-family: Arial, "Helvetica Neue", Helvetica, sans-serif;

    color:#53525200

}

.header{

    padding: 10px 20px;

    background: #9cd6f1;

    display: table;

}

.logo{

    width: 30%;

    float: left;

}

.logo-img{

    width: 50%;

    text-align: center;

    padding-top: 2px;

}

.navagation{

    float: left;

    width: 70%;

    padding :19px 0px 0px 17%;

}

.nav{

    padding: 10px 20px;

    text-align: center;

    text-decoration: none;

    background: #5fbce7;;

    color: #000;

    display: inline-table;

    overflow: hidden;

    font-weight: 500;

    font-size: 16px;

    border-radius: 25px;

}

.sell\_buuton{

    background: #2fed0cfa;

    border-radius: 28px;

    margin-left: 65px;

}

.navagation a:hover{

    background: #fff;

    opacity: 0.5;

    border-radius: 20px;

    padding: 12px 25px;

    font-size: 16px;

}

.banner{

    background: url("../image/banner.jpg") no-repeat  center;

    background-size: cover;

    height: 450px ;

    width: 100%;

    position: relative;

}

.banner aside{

    background-color: #d8d5d5;

    opacity:50%;

    height: 450px ;

    width: 100%;

    position: absolute;

}

.banner\_text{

    position: absolute;

    width: 100%;

    text-align: center;

    color: #0aeb4d;

    height:450px

}

.overlay{

    top: 50%;

    background: #f7ef00;

    opacity: 1;

    border-radius: 15px;

    text-align: center;

    text-decoration: none;

    left: 45%;

    display: block;

    font-size: 20px;

    padding: 12px 20px;

    box-sizing: border-box;

    position: absolute;

}

.banner\_text span{

    top: 40%;

    font-size: 25px;

    font-weight: bold;

    color: #000;

    left: 27%;

    display: block;

    position: absolute;

}

.main\_wrapper{

    background: #ededed;

    width: 100%;

    display: table;

    padding: 12px 40px;

}

.main\_wrapper h1{

    text-align: center;

    color: black;

    padding:12px 20px;

    left: 43%;

    font-size: 25px;

}

.main\_wrapper\_content{

    display: table;

    box-shadow: #dddddd 7px 6px 1px 2px, #ddd -6px 5px 1px 2px

}

.page\_compent{

    float: left;

    padding: 12px;

}

.page\_compent img{

    height: 194px;

}

.page\_compent\_element h6{

    color: #000;

    text-align: center;

    font-size: 14px;

}

.user\_info{

    text-align: center;

    padding: 40px 0px;

    background: #eaee12fa;

}

.login\_interface{

    display: table;

    width: 400px;

    margin: 0 auto;

    box-shadow: #000 3px 4px 21px 2px;

    text-align: center;

    height: 270px;

    padding: 10px 20px;

    border-radius: 25px;

    background: #9de90a;

}

.login\_interface h4,.sign\_in h4{

    color: #000;

    font-size: 20px;

    padding: 12px 10px;

    text-align: center;

}

.login\_interface input{

    padding: 10px;

    border-radius: 20px;

    line-height: 15px;

    text-transform: capitalize;

    display: inline-block;

    cursor: text;

    width: 100%;

    border: none;

    margin: 10px 0px;

}

.login\_interface button ,.sign\_in button{

    padding:12px 20px ;

    width: 100%;

    border: 1px solid;

    border-radius: 25px;

    background: #efff01;

}

.sign\_in{

    width: 400px;

    margin: 0 auto;

    display: table;

    color: #000;

    padding: 12px 20px;

    border-radius: 24px;

    background: #9de90a;

}

.sign\_in label{

    display: inline-table;

    padding: 10px 20px;

    font-size: 14px;

    text-align: center;

}

.sign\_in input{

    width: 100%;

    display: inline-table;

    padding: 10px 20px;

    border: none;

    border-radius: 24px;

    font-size: 14px;

}

.sign\_in button{

    margin: 10px 0px;

}

   .container {

    position: relative;

    max-width: 1200px;

    display: flex;

    align-items: center;

    justify-content: center;

    flex-wrap: wrap;

    z-index: 1;

  }

  .card {

    margin-top: 50px;

    width: 300px;

    height: 380px;

    position: relative;

    background: radial-gradient(

      140% 140% at 0% 0%,

      rgba(255, 255, 255, 0.4) 0%,

      rgba(255, 255, 255, 0) 100%

    );

    border-top: 1px solid #eeeded70;

    border-left: 1px solid #eeeded70;

    border-radius: 50px;

    box-shadow: inset -5px -5px 250px rgba(255, 255, 255, 0.02);

    backdrop-filter: blur(30px);

    display: table;

    align-items: center;

    justify-content: center;

  }

  .card-header{

    padding-top: 10px;

    padding-bottom: 0px;

    padding-right: 0px;

    padding-left: 0px;

    border-bottom-width: 0px;

    text-align:center

  }

  .card-header h1 , .card h6{

     text-align: center;

    color: #fff;

    font-weight: 800;

  }

<!DOCTYPE html>

<html lang="en">

<head>

    <meta charset="UTF-8">

    <meta http-equiv="X-UA-Compatible" content="IE=edge">

    <meta name="viewport" content="width=device-width, initial-scale=1.0">

    <link rel="stylesheet" href="https://maxcdn.bootstrapcdn.com/bootstrap/3.4.1/css/bootstrap.min.css">

    <title>Document</title>

</head>

<body>

    <div class="container">

        <form class="well form-horizontal" action=" " method="post"  id="contact\_form">

    <fieldset>

    <!-- Form Name -->

    <legend><center><h2><b>Registration Form</b></h2></center></legend><br>

    <!-- Text input-->

    <div class="form-group">

      <label class="col-md-4 control-label">First Name</label>

      <div class="col-md-4 inputGroupContainer">

      <div class="input-group">

      <span class="input-group-addon"><i class="glyphicon glyphicon-user"></i></span>

      <input  name="first\_name" placeholder="First Name" class="form-control"  type="text">

        </div>

      </div>

    </div>

    <!-- Text input-->

    <div class="form-group">

      <label class="col-md-4 control-label" >Last Name</label>

        <div class="col-md-4 inputGroupContainer">

        <div class="input-group">

      <span class="input-group-addon"><i class="glyphicon glyphicon-user"></i></span>

      <input name="last\_name" placeholder="Last Name" class="form-control"  type="text">

        </div>

      </div>

    </div>

      <div class="form-group">

      <label class="col-md-4 control-label">Department / Office</label>

        <div class="col-md-4 selectContainer">

        <div class="input-group">

            <span class="input-group-addon"><i class="glyphicon glyphicon-list"></i></span>

        <select name="department" class="form-control selectpicker">

          <option value="">Select your Department/Office</option>

          <option>Department of Engineering</option>

          <option>Department of Agriculture</option>

          <option >Accounting Office</option>

          <option >Tresurer's Office</option>

          <option >MPDC</option>

          <option >MCTC</option>

          <option >MCR</option>

          <option >Mayor's Office</option>

          <option >Tourism Office</option>

        </select>

      </div>

    </div>

    </div>

    <!-- Text input-->

    <div class="form-group">

      <label class="col-md-4 control-label">Username</label>

      <div class="col-md-4 inputGroupContainer">

      <div class="input-group">

      <span class="input-group-addon"><i class="glyphicon glyphicon-user"></i></span>

      <input  name="user\_name" placeholder="Username" class="form-control"  type="text">

        </div>

      </div>

    </div>

    <!-- Text input-->

    <div class="form-group">

      <label class="col-md-4 control-label" >Password</label>

        <div class="col-md-4 inputGroupContainer">

        <div class="input-group">

      <span class="input-group-addon"><i class="glyphicon glyphicon-user"></i></span>

      <input name="user\_password" placeholder="Password" class="form-control"  type="password">

        </div>

      </div>

    </div>

    <!-- Text input-->

    <div class="form-group">

      <label class="col-md-4 control-label" >Confirm Password</label>

        <div class="col-md-4 inputGroupContainer">

        <div class="input-group">

      <span class="input-group-addon"><i class="glyphicon glyphicon-user"></i></span>

      <input name="confirm\_password" placeholder="Confirm Password" class="form-control"  type="password">

        </div>

      </div>

    </div>

    <!-- Text input-->

           <div class="form-group">

      <label class="col-md-4 control-label">E-Mail</label>

        <div class="col-md-4 inputGroupContainer">

        <div class="input-group">

            <span class="input-group-addon"><i class="glyphicon glyphicon-envelope"></i></span>

      <input name="email" placeholder="E-Mail Address" class="form-control"  type="text">

        </div>

      </div>

    </div>

    <!-- Text input-->

    <div class="form-group">

      <label class="col-md-4 control-label">Contact No.</label>

        <div class="col-md-4 inputGroupContainer">

        <div class="input-group">

            <span class="input-group-addon"><i class="glyphicon glyphicon-earphone"></i></span>

      <input name="contact\_no" placeholder="(639)" class="form-control" type="text">

        </div>

      </div>

    </div>

    <!-- Select Basic -->

    <!-- Success message -->

    <div class="alert alert-success" role="alert" id="success\_message">Success <i class="glyphicon glyphicon-thumbs-up"></i> Success!.</div>

    <!-- Button -->

    <div class="form-group">

      <label class="col-md-4 control-label"></label>

      <div class="col-md-4"><br>

        &nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp<button type="submit" class="btn btn-warning" >&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbspSUBMIT <span class="glyphicon glyphicon-send"></span>&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp</button>

      </div>

    </div>

    </fieldset>

    </form>

    </div>

        </div><!-- /.container -->

        <script>

            $(document).ready(function() {

                $('#contact\_form').bootstrapValidator({

                    // To use feedback icons, ensure that you use Bootstrap v3.1.0 or later

                    feedbackIcons: {

                        valid: 'glyphicon glyphicon-ok',

                        invalid: 'glyphicon glyphicon-remove',

                        validating: 'glyphicon glyphicon-refresh'

                    },

                    fields: {

                        first\_name: {

                            validators: {

                                    stringLength: {

                                    min: 2,

                                },

                                    notEmpty: {

                                    message: 'Please enter your First Name'

                                }

                            }

                        },

                         last\_name: {

                            validators: {

                                 stringLength: {

                                    min: 2,

                                },

                                notEmpty: {

                                    message: 'Please enter your Last Name'

                                }

                            }

                        },

                         user\_name: {

                            validators: {

                                 stringLength: {

                                    min: 8,

                                },

                                notEmpty: {

                                    message: 'Please enter your Username'

                                }

                            }

                        },

                         user\_password: {

                            validators: {

                                 stringLength: {

                                    min: 8,

                                },

                                notEmpty: {

                                    message: 'Please enter your Password'

                                }

                            }

                        },

                        confirm\_password: {

                            validators: {

                                 stringLength: {

                                    min: 8,

                                },

                                notEmpty: {

                                    message: 'Please confirm your Password'

                                }

                            }

                        },

                        email: {

                            validators: {

                                notEmpty: {

                                    message: 'Please enter your Email Address'

                                },

                                emailAddress: {

                                    message: 'Please enter a valid Email Address'

                                }

                            }

                        },

                        contact\_no: {

                            validators: {

                              stringLength: {

                                    min: 12,

                                    max: 12,

                                notEmpty: {

                                    message: 'Please enter your Contact No.'

                                 }

                            }

                        },

                         department: {

                            validators: {

                                notEmpty: {

                                    message: 'Please select your Department/Office'

                                }

                            }

                        },

                            }

                        }

                    })

                    .on('success.form.bv', function(e) {

                        $('#success\_message').slideDown({ opacity: "show" }, "slow") // Do something ...

                            $('#contact\_form').data('bootstrapValidator').resetForm();

                        // Prevent form submission

                        e.preventDefault();

                        // Get the form instance

                        var $form = $(e.target);

                        // Get the BootstrapValidator instance

                        var bv = $form.data('bootstrapValidator');

                        // Use Ajax to submit form data

                        $.post($form.attr('action'), $form.serialize(), function(result) {

                            console.log(result);

                        }, 'json');

                    });

            });

        </script>

</body>

</html>

**Backend:**

## import package

In [1]:

**import** pandas **as** pd

**import** numpy **as** np

## import dataset file that we gone train

In [2]:

data**=**pd**.**read\_csv('Bengaluru\_House\_Data.csv')

## head() descdribe the features headings

In [3]:

data**.**head()

Out[3]:

|  | **area\_type** | **availability** | **location** | **size** | **society** | **total\_sqft** | **bath** | **balcony** | **price** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | Super built-up Area | 19-Dec | Electronic City Phase II | 2 BHK | Coomee | 1056 | 2.0 | 1.0 | 39.07 |
| **1** | Plot Area | Ready To Move | Chikka Tirupathi | 4 Bedroom | Theanmp | 2600 | 5.0 | 3.0 | 120.00 |
| **2** | Built-up Area | Ready To Move | Uttarahalli | 3 BHK | NaN | 1440 | 2.0 | 3.0 | 62.00 |
| **3** | Super built-up Area | Ready To Move | Lingadheeranahalli | 3 BHK | Soiewre | 1521 | 3.0 | 1.0 | 95.00 |
| **4** | Super built-up Area | Ready To Move | Kothanur | 2 BHK | NaN | 1200 | 2.0 | 1.0 | 51.00 |

## it will describe how many data set we have

In [4]:

data**.**shape

Out[4]:

(13320, 9)

## Using info we can check the datatype of the data

In [5]:

data**.**info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 13320 entries, 0 to 13319

Data columns (total 9 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 area\_type 13320 non-null object

1 availability 13320 non-null object

2 location 13319 non-null object

3 size 13304 non-null object

4 society 7818 non-null object

5 total\_sqft 13320 non-null object

6 bath 13247 non-null float64

7 balcony 12711 non-null float64

8 price 13320 non-null float64

dtypes: float64(3), object(6)

memory usage: 936.7+ KB

## here we put value\_count on every features

In [6]:

**for** column **in** data**.**columns:

print(data[column]**.**value\_counts())

print("\*"**\***20)

Super built-up Area 8790

Built-up Area 2418

Plot Area 2025

Carpet Area 87

Name: area\_type, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Ready To Move 10581

18-Dec 307

18-May 295

18-Apr 271

18-Aug 200

...

15-Jun 1

15-Dec 1

16-Jan 1

15-Aug 1

16-Nov 1

Name: availability, Length: 81, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Whitefield 540

Sarjapur Road 399

Electronic City 302

Kanakpura Road 273

Thanisandra 234

...

sankeswari 1

2nd phase jp nagar, jp nagar 1

Electronic City Phase 1, 1

Yediyur 1

Anathanagar 1

Name: location, Length: 1305, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

2 BHK 5199

3 BHK 4310

4 Bedroom 826

4 BHK 591

3 Bedroom 547

1 BHK 538

2 Bedroom 329

5 Bedroom 297

6 Bedroom 191

1 Bedroom 105

8 Bedroom 84

7 Bedroom 83

5 BHK 59

9 Bedroom 46

6 BHK 30

7 BHK 17

1 RK 13

10 Bedroom 12

9 BHK 8

8 BHK 5

11 BHK 2

10 BHK 2

11 Bedroom 2

18 Bedroom 1

27 BHK 1

13 BHK 1

19 BHK 1

14 BHK 1

12 Bedroom 1

43 Bedroom 1

16 BHK 1

Name: size, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

GrrvaGr 80

PrarePa 76

Sryalan 59

Prtates 59

GMown E 56

..

Ocezes 1

Naenti 1

Skesta 1

Vithm R 1

Cotaror 1

Name: society, Length: 2688, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

1200 843

1100 221

1500 205

2400 196

600 180

...

763 - 805 1

132Sq. Yards 1

456 1

668 1

2.09Acres 1

Name: total\_sqft, Length: 2117, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

2.0 6908

3.0 3286

4.0 1226

1.0 788

5.0 524

6.0 273

7.0 102

8.0 64

9.0 43

10.0 13

12.0 7

11.0 3

13.0 3

16.0 2

14.0 1

40.0 1

18.0 1

27.0 1

15.0 1

Name: bath, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

2.0 5113

1.0 4897

3.0 1672

0.0 1029

Name: balcony, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

75.00 310

65.00 302

55.00 275

60.00 270

45.00 240

...

45.69 1

85.41 1

91.31 1

102.92 1

62.28 1

Name: price, Length: 1994, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

## check how many null value each feature have

In [7]:

data**.**isna()**.**sum()

Out[7]:

area\_type 0

availability 0

location 1

size 16

society 5502

total\_sqft 0

bath 73

balcony 609

price 0

dtype: int64

## here we drop the Society and balcony (contains so many null value)

if we dont't drop that data set it will cause our prediction process at the end

In [8]:

data**.**drop(columns**=**['area\_type','availability','society','balcony'],inplace**=True**)

## now we check the min max average number of the features

In [9]:

data**.**describe()

Out[9]:

|  | **bath** | **price** |
| --- | --- | --- |
| **count** | 13247.000000 | 13320.000000 |
| **mean** | 2.692610 | 112.565627 |
| **std** | 1.341458 | 148.971674 |
| **min** | 1.000000 | 8.000000 |
| **25%** | 2.000000 | 50.000000 |
| **50%** | 2.000000 | 72.000000 |
| **75%** | 3.000000 | 120.000000 |
| **max** | 40.000000 | 3600.000000 |

## recheck info() after dropping the some features

here we see some features haev some null values like in location 13319 we want 13320

In [10]:

data**.**info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 13320 entries, 0 to 13319

Data columns (total 5 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 location 13319 non-null object

1 size 13304 non-null object

2 total\_sqft 13320 non-null object

3 bath 13247 non-null float64

4 price 13320 non-null float64

dtypes: float64(2), object(3)

memory usage: 520.4+ KB

## In (locaion) we have 1 missing values

here we gone a add sarjarpur road as location in empty cell of data set

In [11]:

data['location']**.**value\_counts()

Out[11]:

Whitefield 540

Sarjapur Road 399

Electronic City 302

Kanakpura Road 273

Thanisandra 234

...

sankeswari 1

2nd phase jp nagar, jp nagar 1

Electronic City Phase 1, 1

Yediyur 1

Anathanagar 1

Name: location, Length: 1305, dtype: int64

In [12]:

data['location'] **=** data['location']**.**fillna('Sarjapur Road')

## In (size) we have 16 missing values

here we have different name for the size so we gone replace Bedroom to Bhk also 16 missing value so we do same as the location

In [13]:

data['size']**.**value\_counts()

Out[13]:

2 BHK 5199

3 BHK 4310

4 Bedroom 826

4 BHK 591

3 Bedroom 547

1 BHK 538

2 Bedroom 329

5 Bedroom 297

6 Bedroom 191

1 Bedroom 105

8 Bedroom 84

7 Bedroom 83

5 BHK 59

9 Bedroom 46

6 BHK 30

7 BHK 17

1 RK 13

10 Bedroom 12

9 BHK 8

8 BHK 5

11 BHK 2

10 BHK 2

11 Bedroom 2

18 Bedroom 1

27 BHK 1

13 BHK 1

19 BHK 1

14 BHK 1

12 Bedroom 1

43 Bedroom 1

16 BHK 1

Name: size, dtype: int64

In [14]:

data['size'] **=** data['size']**.**fillna('2 BHK')

## in (bath) we have 73 missing values

In [ ]:

data['bath'] **=** data['bath']**.**fillna(data['bath']**.**median())

In [16]:

data**.**info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 13320 entries, 0 to 13319

Data columns (total 5 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 location 13320 non-null object

1 size 13320 non-null object

2 total\_sqft 13320 non-null object

3 bath 13320 non-null float64

4 price 13320 non-null float64

dtypes: float64(2), object(3)

memory usage: 520.4+ KB

## we are now convert all bedroom name into BHK

In [17]:

data['bhk']**=**data['size']**.**str**.**split()**.**str**.**get(0)**.**astype(int)

In [18]:

data[data**.**bhk **>** 20]

Out[18]:

|  | **location** | **size** | **total\_sqft** | **bath** | **price** | **bhk** |
| --- | --- | --- | --- | --- | --- | --- |
| **1718** | 2Electronic City Phase II | 27 BHK | 8000 | 27.0 | 230.0 | 27 |
| **4684** | Munnekollal | 43 Bedroom | 2400 | 40.0 | 660.0 | 43 |

# Now we gone resolve the range query of (total\_sqft) using one function and store it in our funtion

In [19]:

data['total\_sqft']**.**unique()

Out[19]:

array(['1056', '2600', '1440', ..., '1133 - 1384', '774', '4689'],

dtype=object)

In [20]:

**def** convertRange(x):

temp **=** x**.**split('-')

**if** len(temp) **==** 2:

**return** (float(temp[0]) **+** float(temp[1]))**/**2

**try**:

**return** float(x)

**except**:

**return** **None**

In [21]:

data['total\_sqft']**=**data['total\_sqft']**.**apply(convertRange)

In [22]:

data**.**head()

Out[22]:

|  | **location** | **size** | **total\_sqft** | **bath** | **price** | **bhk** |
| --- | --- | --- | --- | --- | --- | --- |
| **0** | Electronic City Phase II | 2 BHK | 1056.0 | 2.0 | 39.07 | 2 |
| **1** | Chikka Tirupathi | 4 Bedroom | 2600.0 | 5.0 | 120.00 | 4 |
| **2** | Uttarahalli | 3 BHK | 1440.0 | 2.0 | 62.00 | 3 |
| **3** | Lingadheeranahalli | 3 BHK | 1521.0 | 3.0 | 95.00 | 3 |
| **4** | Kothanur | 2 BHK | 1200.0 | 2.0 | 51.00 | 2 |

# Price Per Square feet

In [23]:

data['price\_per\_sqft'] **=** data['price'] **\***100000 **/** data['total\_sqft']

In [24]:

data['price\_per\_sqft']

Out[24]:

0 3699.810606

1 4615.384615

2 4305.555556

3 6245.890861

4 4250.000000

...

13315 6689.834926

13316 11111.111111

13317 5258.545136

13318 10407.336319

13319 3090.909091

Name: price\_per\_sqft, Length: 13320, dtype: float64

In [25]:

data**.**describe()

Out[25]:

|  | **total\_sqft** | **bath** | **price** | **bhk** | **price\_per\_sqft** |
| --- | --- | --- | --- | --- | --- |
| **count** | 13274.000000 | 13320.000000 | 13320.000000 | 13320.000000 | 1.327400e+04 |
| **mean** | 1559.626694 | 2.688814 | 112.565627 | 2.802778 | 7.907501e+03 |
| **std** | 1238.405258 | 1.338754 | 148.971674 | 1.294496 | 1.064296e+05 |
| **min** | 1.000000 | 1.000000 | 8.000000 | 1.000000 | 2.678298e+02 |
| **25%** | 1100.000000 | 2.000000 | 50.000000 | 2.000000 | 4.266865e+03 |
| **50%** | 1276.000000 | 2.000000 | 72.000000 | 3.000000 | 5.434306e+03 |
| **75%** | 1680.000000 | 3.000000 | 120.000000 | 3.000000 | 7.311746e+03 |
| **max** | 52272.000000 | 40.000000 | 3600.000000 | 43.000000 | 1.200000e+07 |

## Now we gone reduce the outlier value from the loaction and white space

In [26]:

data['location']**.**value\_counts()

Out[26]:

Whitefield 540

Sarjapur Road 399

Electronic City 302

Kanakpura Road 273

Thanisandra 234

...

sankeswari 1

2nd phase jp nagar, jp nagar 1

Electronic City Phase 1, 1

Yediyur 1

Anathanagar 1

Name: location, Length: 1306, dtype: int64

## remove the white spaces from the starting & ending from the location values

In [27]:

data['location'] **=** data['location']**.**apply(**lambda** x: x**.**strip())

location\_count **=** data['location']**.**value\_counts()

In [28]:

location\_count

Out[28]:

Whitefield 541

Sarjapur Road 399

Electronic City 304

Kanakpura Road 273

Thanisandra 237

...

Yediyur 1

Somanna Garden 1

Byrasandra Extension 1

Jakkasandra 1

Anathanagar 1

Name: location, Length: 1295, dtype: int64

## Now we gone Find out how many data in location are come 10 or less are replace by others

In [29]:

location\_count\_less\_10 **=** location\_count[location\_count**<=**10]

location\_count\_less\_10

Out[29]:

Sadashiva Nagar 10

Kalkere 10

Naganathapura 10

Gunjur Palya 10

Sector 1 HSR Layout 10

..

Yediyur 1

Somanna Garden 1

Byrasandra Extension 1

Jakkasandra 1

Anathanagar 1

Name: location, Length: 1054, dtype: int64

In [32]:

data['location'] **=** data['location']**.**apply(**lambda** x: 'other' **if** x **in** location\_count\_less\_10 **else** x)

In [34]:

data['location']**.**value\_counts()

Out[34]:

other 2886

Whitefield 541

Sarjapur Road 399

Electronic City 304

Kanakpura Road 273

...

Tindlu 11

Pattandur Agrahara 11

HAL 2nd Stage 11

Kodigehalli 11

Marsur 11

Name: location, Length: 242, dtype: int64

# Outlier detection and removal (totalsqft)

In [35]:

data**.**describe()

Out[35]:

|  | **total\_sqft** | **bath** | **price** | **bhk** | **price\_per\_sqft** |
| --- | --- | --- | --- | --- | --- |
| **count** | 13274.000000 | 13320.000000 | 13320.000000 | 13320.000000 | 1.327400e+04 |
| **mean** | 1559.626694 | 2.688814 | 112.565627 | 2.802778 | 7.907501e+03 |
| **std** | 1238.405258 | 1.338754 | 148.971674 | 1.294496 | 1.064296e+05 |
| **min** | 1.000000 | 1.000000 | 8.000000 | 1.000000 | 2.678298e+02 |
| **25%** | 1100.000000 | 2.000000 | 50.000000 | 2.000000 | 4.266865e+03 |
| **50%** | 1276.000000 | 2.000000 | 72.000000 | 3.000000 | 5.434306e+03 |
| **75%** | 1680.000000 | 3.000000 | 120.000000 | 3.000000 | 7.311746e+03 |
| **max** | 52272.000000 | 40.000000 | 3600.000000 | 43.000000 | 1.200000e+07 |

In [36]:

(data['total\_sqft']**/**data['bhk'])**.**describe()

Out[36]:

count 13274.000000

mean 575.074878

std 388.205175

min 0.250000

25% 473.333333

50% 552.500000

75% 625.000000

max 26136.000000

dtype: float64

##### here minimum square feet value is 1 which wrong so we gone convert it in 300

In [37]:

data **=** data[((data['total\_sqft']**/**data['bhk']) **>=** 300)]

data**.**describe()

Out[37]:

|  | **total\_sqft** | **bath** | **price** | **bhk** | **price\_per\_sqft** |
| --- | --- | --- | --- | --- | --- |
| **count** | 12530.000000 | 12530.000000 | 12530.000000 | 12530.000000 | 12530.000000 |
| **mean** | 1594.564544 | 2.559537 | 111.382401 | 2.650838 | 6303.979357 |
| **std** | 1261.271296 | 1.077938 | 152.077329 | 0.976678 | 4162.237981 |
| **min** | 300.000000 | 1.000000 | 8.440000 | 1.000000 | 267.829813 |
| **25%** | 1116.000000 | 2.000000 | 49.000000 | 2.000000 | 4210.526316 |
| **50%** | 1300.000000 | 2.000000 | 70.000000 | 3.000000 | 5294.117647 |
| **75%** | 1700.000000 | 3.000000 | 115.000000 | 3.000000 | 6916.666667 |
| **max** | 52272.000000 | 16.000000 | 3600.000000 | 16.000000 | 176470.588235 |

In [38]:

data**.**shape

Out[38]:

(12530, 7)

## Outlier detection and removal (price\_per\_sqft)

In [39]:

data**.**price\_per\_sqft**.**describe()

Out[39]:

count 12530.000000

mean 6303.979357

std 4162.237981

min 267.829813

25% 4210.526316

50% 5294.117647

75% 6916.666667

max 176470.588235

Name: price\_per\_sqft, dtype: float64

In [43]:

**def** remove\_outliers\_sqft(df):

df\_output **=** pd**.**DataFrame()

*#groupby gives us the key and subdf value*

**for** key,subdf **in** df**.**groupby('location'):

*#find mean value of every price\_per\_sqft data*

m **=** np**.**mean(subdf**.**price\_per\_sqft)

*#find standard deviation value of every price\_per\_sqft data*

st **=** np**.**std(subdf**.**price\_per\_sqft)

*#now we drop the data which are outlier & store data*

*#which are in between 1 mean and standard deviation*

*#here we have( 23 45 67 87 98 ) and suppose that std = 20 so m-st = 3 and m+st= 43*

gen\_df **=** subdf[(subdf**.**price\_per\_sqft **>** (m**-**st)) **&** (subdf**.**price\_per\_sqft **<=** (m**+**st))]

df\_output **=** pd**.**concat([df\_output,gen\_df],ignore\_index **=** **True**)

**return** df\_output

data **=** remove\_outliers\_sqft(data)

data**.**describe()

Out[43]:

|  | **total\_sqft** | **bath** | **price** | **bhk** | **price\_per\_sqft** |
| --- | --- | --- | --- | --- | --- |
| **count** | 10301.000000 | 10301.000000 | 10301.000000 | 10301.000000 | 10301.000000 |
| **mean** | 1508.440608 | 2.471702 | 91.286372 | 2.574896 | 5659.062876 |
| **std** | 880.694214 | 0.979449 | 86.342786 | 0.897649 | 2265.774749 |
| **min** | 300.000000 | 1.000000 | 10.000000 | 1.000000 | 1250.000000 |
| **25%** | 1110.000000 | 2.000000 | 49.000000 | 2.000000 | 4244.897959 |
| **50%** | 1286.000000 | 2.000000 | 67.000000 | 2.000000 | 5175.600739 |
| **75%** | 1650.000000 | 3.000000 | 100.000000 | 3.000000 | 6428.571429 |
| **max** | 30400.000000 | 16.000000 | 2200.000000 | 16.000000 | 24509.803922 |

## Outlier detection and removal (bhk)

here we gone run the for loop where if 3: s mean is greater then 2: s mean then we will keep it otherwise we drop it

In [50]:

**def** bhk\_outlier\_remover(df):

*#here we gone a store that indices which we want to exclude in this array*

exclude\_indices **=** np**.**array([])

*#on df we group by on basis of location and got*

*#location and it's sub data frame*

**for** loacation, location\_df **in** df**.**groupby('location'):

bhk\_stats **=** {}

**for** bhk, bhk\_df **in** location\_df**.**groupby('bhk'):

bhk\_stats[bhk] **=** {

'mean': np**.**mean(bhk\_df**.**price\_per\_sqft),

'std': np**.**std(bhk\_df**.**price\_per\_sqft),

'count': bhk\_df**.**shape[0]

}

*#print(location, bhk\_stats)*

**for** bhk, bhk\_df **in** location\_df**.**groupby('bhk'):

stats **=** bhk\_stats**.**get(bhk**-**1)

**if** stats **and** stats['count']**>**5:

exclude\_indices **=** np**.**append(exclude\_indices, bhk\_df[bhk\_df**.**price\_per\_sqft**<**(stats['mean'])]**.**index**.**values)

**return** df**.**drop(exclude\_indices,axis**=**'index')

In [51]:

data**=**bhk\_outlier\_remover(data)

In [52]:

data**.**shape

Out[52]:

(7361, 7)

In [53]:

data

Out[53]:

|  | **location** | **size** | **total\_sqft** | **bath** | **price** | **bhk** | **price\_per\_sqft** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 1st Block Jayanagar | 4 BHK | 2850.0 | 4.0 | 428.0 | 4 | 15017.543860 |
| **1** | 1st Block Jayanagar | 3 BHK | 1630.0 | 3.0 | 194.0 | 3 | 11901.840491 |
| **2** | 1st Block Jayanagar | 3 BHK | 1875.0 | 2.0 | 235.0 | 3 | 12533.333333 |
| **3** | 1st Block Jayanagar | 3 BHK | 1200.0 | 2.0 | 130.0 | 3 | 10833.333333 |
| **4** | 1st Block Jayanagar | 2 BHK | 1235.0 | 2.0 | 148.0 | 2 | 11983.805668 |
| **...** | ... | ... | ... | ... | ... | ... | ... |
| **10292** | other | 2 BHK | 1200.0 | 2.0 | 70.0 | 2 | 5833.333333 |
| **10293** | other | 1 BHK | 1800.0 | 1.0 | 200.0 | 1 | 11111.111111 |
| **10296** | other | 2 BHK | 1353.0 | 2.0 | 110.0 | 2 | 8130.081301 |
| **10297** | other | 1 Bedroom | 812.0 | 1.0 | 26.0 | 1 | 3201.970443 |
| **10300** | other | 4 BHK | 3600.0 | 5.0 | 400.0 | 4 | 11111.111111 |

7361 rows × 7 columns

In [55]:

data**.**drop(columns**=**['size','price\_per\_sqft'],inplace**=True**)

## Cleaned Data

In [56]:

data**.**head()

Out[56]:

|  | **location** | **total\_sqft** | **bath** | **price** | **Bhk** |
| --- | --- | --- | --- | --- | --- |
| **0** | 1st Block Jayanagar | 2850.0 | 4.0 | 428.0 | 4 |
| **1** | 1st Block Jayanagar | 1630.0 | 3.0 | 194.0 | 3 |
| **2** | 1st Block Jayanagar | 1875.0 | 2.0 | 235.0 | 3 |
| **3** | 1st Block Jayanagar | 1200.0 | 2.0 | 130.0 | 3 |
| **4** | 1st Block Jayanagar | 1235.0 | 2.0 | 148.0 | 2 |

In [65]:

data**.**to\_csv("Cleaned\_data.csv")

cmd -> conda activate sklearn-env

In [68]:

X**=**data**.**drop(columns**=**['price'])

y**=**data['price']

In [69]:

**from** sklearn.model\_selection **import** train\_test\_split

**from** sklearn.linear\_model **import** LinearRegression,Lasso,Ridge

**from** sklearn.preprocessing **import** OneHotEncoder, StandardScaler

**from** sklearn.compose **import** make\_column\_transformer

**from** sklearn.pipeline **import** make\_pipeline

**from** sklearn.metrics **import** r2\_score

In [70]:

X\_train,X\_test,y\_train,y\_test **=** train\_test\_split(X,y, test\_size**=**0.2, random\_state**=**0)

In [71]:

print(X\_train**.**shape)

print(X\_test**.**shape)

(5888, 4)

(1473, 4)

## Applying Linear Regression

Summary :

first we apply OneHotEncoder location after that we create the standard scaler to scale the things and then we apply linearregression and then on pipeline we apply (column\_trans,scaler,lr) and then when we run X\_train it will first go to columns\_trans run then apply scaler and give the output

In [73]:

column\_trans **=** make\_column\_transformer((OneHotEncoder(sparse**=False**), ['location']), remainder**=**'passthrough')

In [74]:

scaler **=** StandardScaler()

In [75]:

lr **=** LinearRegression(normalize**=True**)

In [76]:

pipe **=** make\_pipeline(column\_trans,scaler,lr)

In [77]:

pipe**.**fit(X\_train,y\_train)

Out[77]:

Pipeline(steps=[('columntransformer',

ColumnTransformer(remainder='passthrough',

transformers=[('onehotencoder',

OneHotEncoder(sparse=False),

['location'])])),

('standardscaler', StandardScaler()),

('linearregression', LinearRegression(normalize=True))])

In [90]:

y\_pred\_lr **=** pipe**.**predict(X\_test)

r2\_score(y\_test, y\_pred\_lr)

Out[90]:

0.8234146633312647

## Applying Lasso

In [81]:

lasso **=** Lasso()

In [82]:

pipe **=** make\_pipeline(column\_trans,scaler, lasso)

In [83]:

pipe**.**fit(X\_train,y\_train)

Out[83]:

Pipeline(steps=[('columntransformer',

ColumnTransformer(remainder='passthrough',

transformers=[('onehotencoder',

OneHotEncoder(sparse=False),

['location'])])),

('standardscaler', StandardScaler()), ('lasso', Lasso())])

In [84]:

y\_pred\_lasso **=** pipe**.**predict(X\_test)

r2\_score(y\_test, y\_pred\_lasso)

Out[84]:

0.8128285650772719

## Appling Ridge

In [85]:

ridge **=** Ridge()

In [86]:

pipe **=** make\_pipeline(column\_trans,scaler, ridge)

In [87]:

pipe**.**fit(X\_train,y\_train)

Out[87]:

Pipeline(steps=[('columntransformer',

ColumnTransformer(remainder='passthrough',

transformers=[('onehotencoder',

OneHotEncoder(sparse=False),

['location'])])),

('standardscaler', StandardScaler()), ('ridge', Ridge())])

In [89]:

y\_pred\_ridge **=** pipe**.**predict(X\_test)

r2\_score(y\_test, y\_pred\_ridge)

Out[89]:

0.8234146633312647

In [92]:

print("No Regularization: ", r2\_score(y\_test, y\_pred\_lr))

print("Lasso:", r2\_score(y\_test, y\_pred\_lasso))

print("Ridge", r2\_score(y\_test, y\_pred\_ridge))

No Regularization: 0.8234146633312647

Lasso: 0.8128285650772719

Ridge 0.8234146633312647

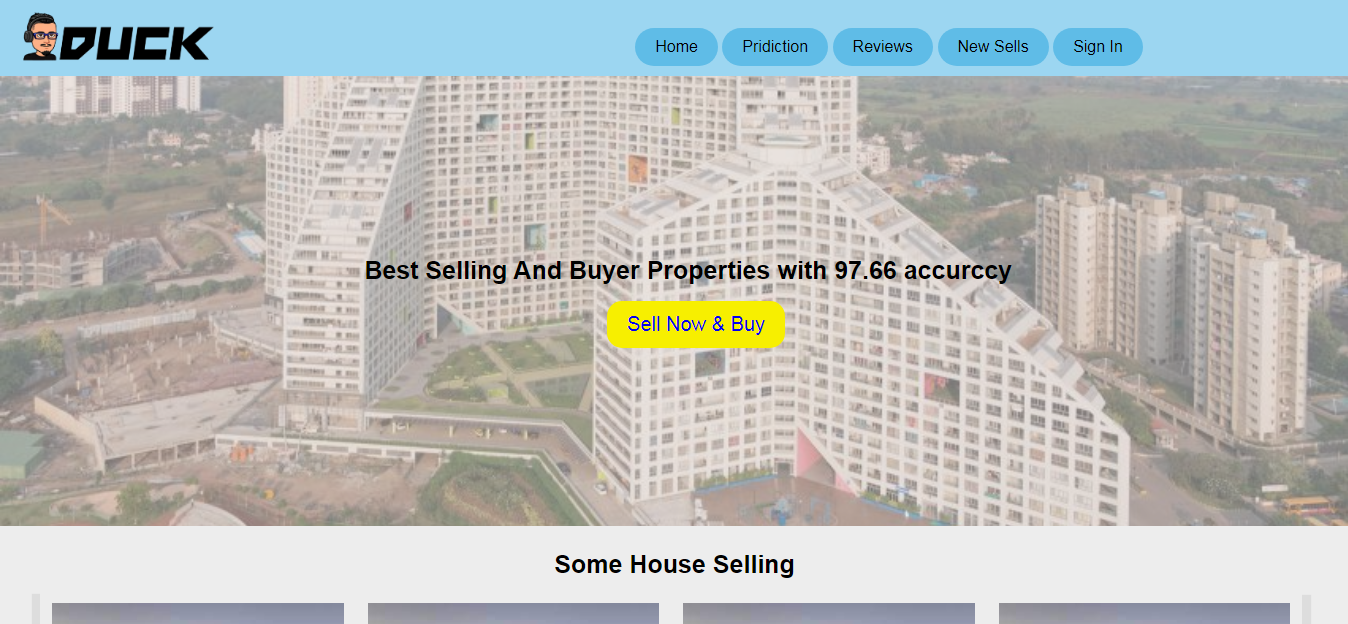
In [93]:

**import** pickle

In [94]:

pickle**.**dump(pipe, open('RidgeModel.pk1', 'wb'))

|  |
| --- |
|  |
|  | import pandas as pd  Import numpy as np |
|  | from flask import Flask, render\_template, request |
|  | import pickle |
|  |  |
|  |  |
|  | app = Flask(\_\_name\_\_) |
|  | data= pd.read\_csv('Cleaned\_data.csv') |
|  | pipe = pickle.load(open("RidgeModel.pk1","rb")) |
|  |  |
|  |  |
|  | @app.route('/') |
|  | def index(): |
|  |  |
|  | locations = sorted(data['location'].unique()) |
|  | return render\_template('index.html', locations=locations) |
|  |  |
|  | @app.route('/predict', methods=['POST']) |
|  | def predict(): |
|  | location = request.form.get('location') |
|  | bhk = request.form.get('bhk') |
|  | bath = request.form.get('bath') |
|  | sqft = request.form.get('total\_sqft') |
|  |  |
|  | print(location, bhk, bath, sqft) |
|  | input = pd.DataFrame([[location,sqft,bath,bhk]],columns=['location', 'total\_sqft', 'bath', 'bhk']) |
|  | prediction = pipe.predict(input)[0] \* 1e5 |
|  |  |
|  | return str(np.round(prediction,2)) |
|  |  |
|  | if \_\_name\_\_=="\_\_main\_\_": |
|  | app.run(debug=True, port=5001) |

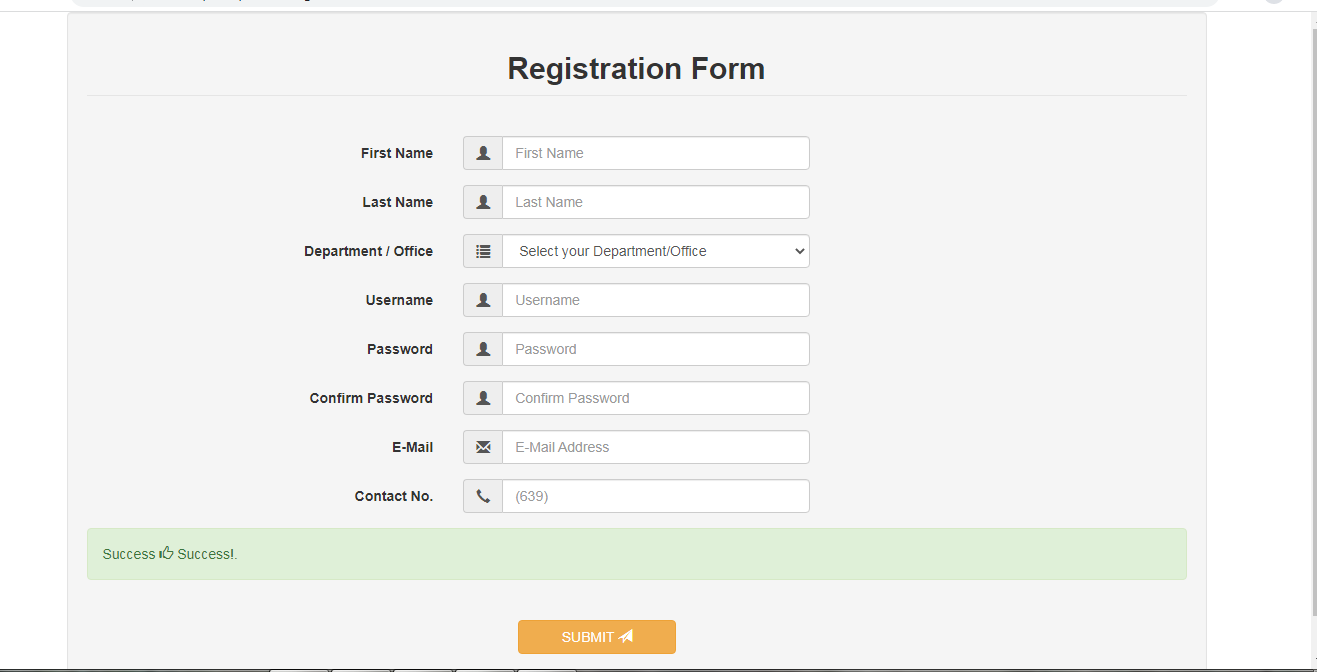


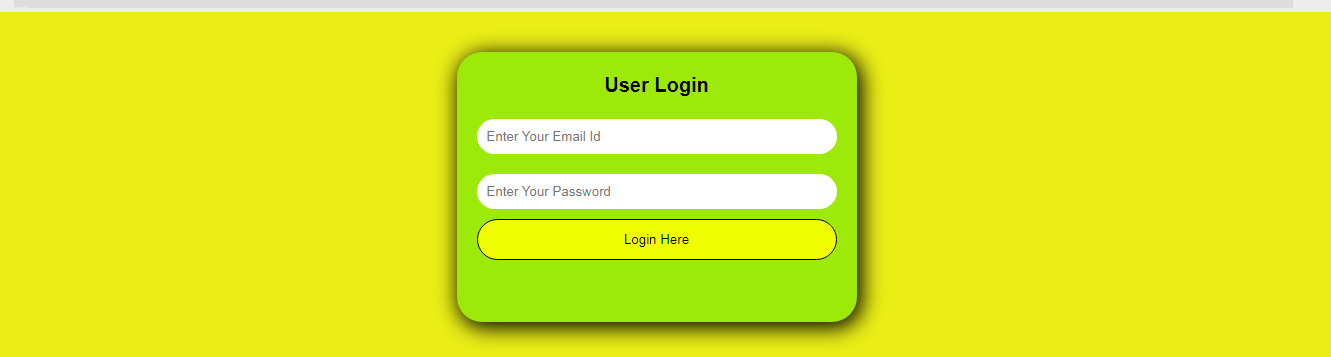
**Screenshots:**

Below is the Login Screen. A l**ogin page**is an entry page to app that requires user identification and authentication,regularly performed by entering a username and password combination. If user don’t have account, then he/she can create account for the app by clicking sign up button. And if user enter wrong or less than 6-character password then it will show wrong or password should be more than 6 characters.

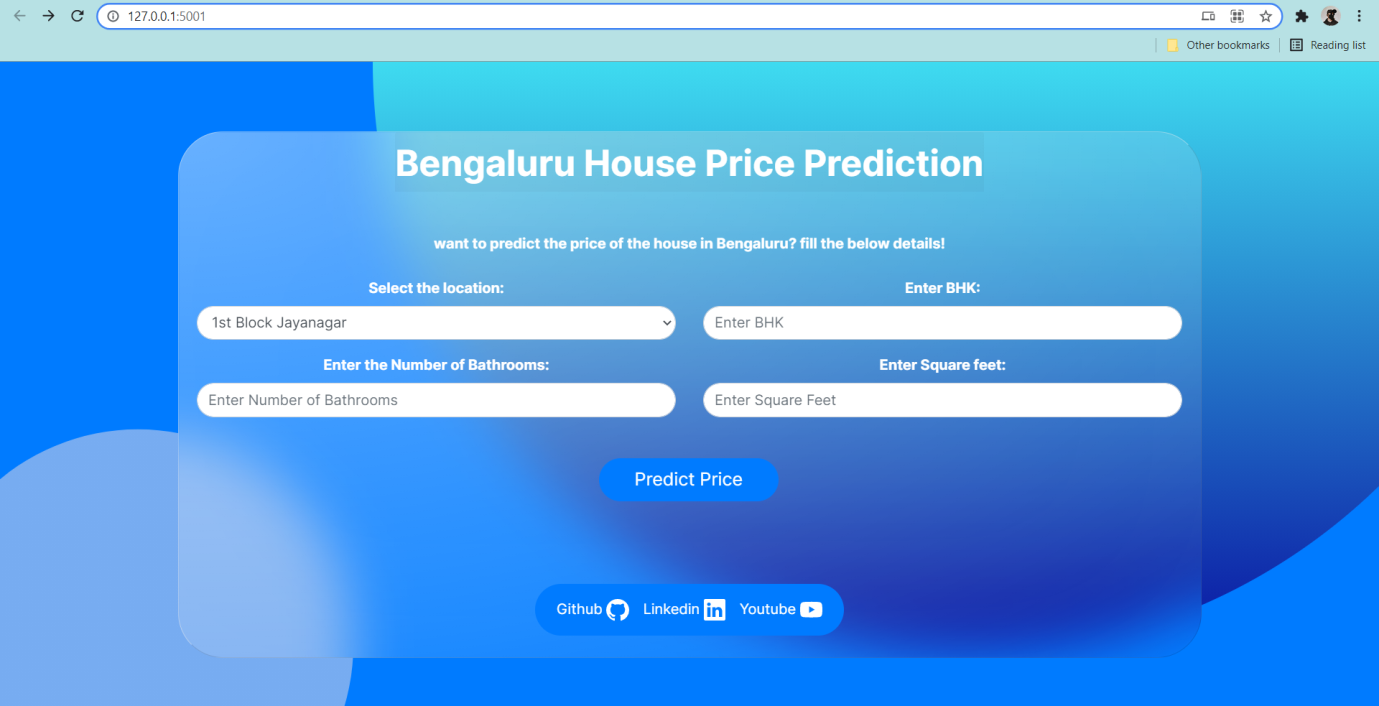
Below is a Sign-up Screen (also known as Registration page). This enables user to register and gain access to our application. Here user has to provide full name, email, password &If users mistype their password, they won't recognize it. The confirm password catches typos by prompting users to type their password twice.

Below is the Info Screen. Here detail information regarding Heart Chamber, Heart Condition, Heart Tests, Heart Treatments is provided. Heart chamber provides physical information about heart and how the heart works. Heart condition shows how our heart get defected, in heart test user will know what are the various test should be done to get result related to heart disease.





Below is the Prediction Screen. Here user has to provide thirteen attributes in numeric format,including Age, gender, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, oldpeak,exercise relative to rest, the slope of the peak, number of major vessels,thal. After Entering those thirteen attributes in numeric format, below there is submit button. By clicking on submit button we proceed for the predication of Cardiovascular heart disease.



Below is Dashboard screen, were all the previous prediction records are saved. User entered values are stored in firebase database including predicted result. And in this dashboard user can access all his/her previous record by date and time.

1. **CONCLUSION**

**Conclusion:**

The paper entitled “House Price Prediction Using Machine Learning” has presented to predict house price based on various features on given data. From our analysis we set value of RMSE as 2.9131889. In this model we have to add additional features like tax, air quality so it become different from other prediction system. It helps people to buy house in budget and reduce loss of money.

.

1. **REFERENCE**

**Reference:**

1. Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh**, A Hybrid Regression Technique for House Prices Prediction**, Institute of High Performance Computing (IHPC), Agency for Science Technology and Research (A\*STAR), Singapore.
2. Ayush Varma, Abhijit Sarma, Sagar Doshi, Rohini Nair, **House Price Prediction Using Machine Learning And Neural Networks.**
3. Adyan Nur Alfiyatin, Hilman Taufiq, Ruth Ema Febrita, Wayan Firdaus Mahmudy, **Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization**.
4. G. Naga Satish, Ch. V. Raghavendran, M.D.Sugnana Rao, Ch.Srinivasulu, **House Price Prediction Using Machine Learning.**
5. Nehal N Ghosalkar, Sudhir N Dhage, **Real Estate Value Prediction Using Linear Regression.**
6. D. Banerjee and S. Dutta, **Predicting the housing price direction using machine learning techniques**, 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), Chennai, 2017, pp. 2998-3000.

https://doi.org/10.1109/ICPCSI.2017.8392275

1. Bruno Klausde Aquino Afonso1, Luckeciano Carvalho Melo2, Willian Dihanster Gomesde Oliveira1, Samuel Brunoda Silva Sousa1, Lilian Berton1,**Housing Prices Prediction with a Deep Learning and Random Forest Ensemble**, Institute of Science and Technology – Federal University of S˜ao Paulo (UNIFESP)
2. T. D. Phan, **Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia**, 2018 International Conference on Machine Learning and DataEngineering (iCMLDE), Sydney, Australia, 2018, pp.35-42.

https://doi.org/10.1109/iCMLDE.2018.00017

1. Sayan Putatunda, **PropTech for Proactive Pricing of Houses in Classified Advertisements in the Indian Real Estate Market**