

**MULTIMEDIA UNIVERSITY OF KENYA**

FACULTY OF COMPUTING & INFORMATION TECHNOLOGY

**CALIFORNIA HOUSE PRICE PREDICTOR**

BY

EMANUEL OKOTH

CIT-222-021/2018

**SUPERVISOR:**

JULY 2022

Submitted in partial fulfillment of the requirements of Bachelor of Science in Computer Science.

# ACKNOWLEDGEMENTS

I extend my sincere appreciation to any lecturers who have ever educated me. The foundation they laid in me aided me in developing a well-structured study project. Additionally, I would like to thank my classmates and friends for the support that has helped me improve this project. Finally, I want to express my gratitude to my parents for their encouragement and for being there for me every step of the journey.

# DECLARATION

I hereby declare that this Proposal is my work and has, to the best of my knowledge, not been submitted to any other institution of higher learning.

Student: Registration Number:

Signature: ............................................... Date: .....................................................

This Proposal has been submitted as a partial fulfillment of requirements for the Bachelor of Science in Computer Technology of Multimedia University of Kenya with my approval as the University supervisor.

Supervisor:

Signature: ..................................................... Date: ..................................................

# ABSTRACT

Real Estate Business has been on the Rise in California State the past few years; that has made the market sector a lucrative business to venture in. Real Estate companies however, need to know the various prices of houses in each location in California, so that they can be able to make data guided decision on their investment plan and locations of interest within the state.

To solve their problem, I have built a Machine Learning model that will help them predict the value of any house in California based on the house features. Using the application is simple they will insert the relevant house features to the Machine Learning App and it will predict and output the house value.

Table of Contents

[ACKNOWLEDGEMENTS 2](#_TOC_250055)

[DECLARATION 3](#_TOC_250054)

[ABSTRACT 4](#_TOC_250053)

[CHAPTER 1 - INTRODUCTION 7](#_TOC_250052)

* 1. [BACKGROUND STUDY 7](#_TOC_250051)
  2. [PROBLEM STATEMENT 7](#_TOC_250050)
  3. [AIM OF THE STUDY 7](#_TOC_250049)
     1. [RESEARCH OBJECTIVES 8](#_TOC_250048)
  4. [SIGNIFICANCE OF THE RESEARCH 8](#_TOC_250047)
  5. [SCOPE 8](#_TOC_250046)
  6. [ASSUMPTIONS 8](#_TOC_250045)
  7. [LIMITATIONS 9](#_TOC_250044)

[CHAPTER 2 - LITERATURE REVIEW 9](#_TOC_250043)

* 1. [INTRODUCTION 9](#_TOC_250042)
  2. [RELATED SYSTEMS 9](#_TOC_250041)
     1. [THE INSTITUTION WEBSITE 10](#_TOC_250040)
  3. [LIMITATIONS OF RELATED SYSTEMS 10](#_TOC_250039)
     1. [THE INSTITUTION WEBSITE 10](#_TOC_250038)
  4. [HOW THE PROPOSED SOLUTIONS WILL HANDLE THE AFOREMENTIONED WEAKNESS 10](#_TOC_250037)

[CHAPTER 3 – METHODOLOGY 11](#_TOC_250036)

* 1. [INTRODUCTION 11](#_TOC_250035)
  2. [METHODOLOGY 11](#_TOC_250034)
     1. [AGILE METHODOLOGY 11](#_TOC_250033)
     2. [ADVANTAGES OF AGILE MODEL 12](#_TOC_250032)
     3. [JUSTIFICATION FOR USING AGILE MODEL 13](#_TOC_250031)
  3. [DATA COLLECTION METHODS AND TOOLS 13](#_TOC_250030)
     1. [INTERVIEWS 13](#_TOC_250029)

[CHAPTER 4 – SYSTEM ANALYSIS 14](#_TOC_250028)

* 1. [INTRODUCTION 14](#_TOC_250027)
  2. [DETAILED ANALYSIS 14](#_TOC_250026)
     1. [FLOW CHART 14](#_TOC_250025)
     2. [DATA FLOW DIAGRAM 15](#_TOC_250024)
  3. [SYSTEM REQUIREMENTS 15](#_TOC_250023)
     1. [FUNCTIONAL REQUIREMENTS 16](#_TOC_250022)
     2. [NON-FUNCTIONAL REQUIREMENTS 16](#_TOC_250021)

[CHAPTER 5 – SYSTEM DESIGN 17](#_TOC_250020)

* 1. [ARCHITECTURAL DESIGN 17](#_TOC_250019)
  2. [DATABASE DESIGN 18](#_TOC_250018)
     1. [ENTITY DESCRIPTION 18](#_TOC_250017)
     2. [USER INTERFACE DESIGN 19](#_TOC_250016)

[CHAPTER 6 - IMPLEMENTATION AND TESTING 24](#_TOC_250015)

* 1. [DEVELOPMENT AND ENVIRONMENT 24](#_TOC_250014)
  2. [SYSTEM COMPONENTS 24](#_TOC_250013)
     1. [MODELS 24](#_TOC_250012)
     2. [TEMPLATES 24](#_TOC_250011)
  3. [TEST PLAN 24](#_TOC_250010)
     1. [UNIT TESTING 25](#_TOC_250009)
     2. [INTERGRATION TESTING 25](#_TOC_250008)
     3. [REGRESSION TESTING 25](#_TOC_250007)
     4. [SYSTEM TESTING 25](#_TOC_250006)

[CHAPTER 7 – CRITICAL APPRAISAL 27](#_TOC_250005)

* 1. [ACHIEVEMENTS AND LESSONS LEARNT 27](#_TOC_250004)
  2. [CHALLENGES 27](#_TOC_250003)
  3. [RECOMMENDATIONS 27](#_TOC_250002)
  4. [CONCLUSIONS 28](#_TOC_250001)

[REFERENCES 28](#_TOC_250000)

# CHAPTER 1 - INTRODUCTION

## BACKGROUND STUDY

Machine Learning exist to ease the way we do business by predicting outcomes in business, automating repetitive task, provides a different view in customer trends etc. Machine Learning has provided reliable insights to our data before and has helped many organizations improve their operations and enhance their business models. These Machine Learning models have enable companies to improve their decision-making process and hence increase business profits. This makes Machine Learning the ideal shot for this problem, since it has proven track record and trusted by many companies across the globe.

## PROBLEM STATEMENT

Real Estate business has been on the rise and hence the hasty rush by many enterprises to catch the business wave. Investing is not an easy task and making investment decision is critical since it might be the bloodline of the business or its demise. Market research will be very important and time consuming too, since you will need troops on the ground to do the work. This will slow the entire process since it will take days to gather data and analyze it. This will slow the investing process just in case it will be done by a human, due to the vast data that might be collected. To solve this whole tedious process, I am going to use my Data Science skills to analyze and build a great house predictor based on the previous California Census data that has all the houses, their features and prices.

## AIM OF THE STUDY

The purpose of this research is to come up with a Machine Learning system to help Real Estate enterprises to accurately predict house prices and make data guided decision based on my data analysis and Machine Learning Model.

### RESEARCH OBJECTIVES

1. Develop a Machine Learning Model that discover patterns in the house industry data and then make prediction based on these and intricate patterns for answering business questions and solving business problems.

## SIGNIFICANCE OF THE RESEARCH

Provide an easier option for companies to find the right answers about their business and industry. Provide a quick automated option to doing all that tedious data analysis work, giving the enterprise more time to work on other business problems.

## SCOPE

The system maintains two levels of users; the admin level and the user level. The administrator level can only be accessed by the club administrator or patron. The administrator can be able to create update or delete an article or post. The system allows the manager to login and retrieve or update information. The user need not necessarily log in unless they want to comment on a post but they can browse through the various categories and read through their posts without logging in.

## ASSUMPTIONS

The following assumptions were made in the creation of the Model:

1. It assumes the observations are independent of each other to predict the house value.
2. The users are computer literate.
3. The data collected was enough to build an accurate model.

## LIMITATIONS

* Lack of data; Many Machine Learning algorithms require large amount of data before they begin to give useful results.

# CHAPTER 2 - LITERATURE REVIEW

## INTRODUCTION

The Real Estate enterprise business is growing at high rate in California. To make better business decisions on that vast amount of that industry data you will need statistics for both analysis and prediction. Using old statistical ways and a human labor will take time and consume a lot of resources to achieve the objective. They might even end up not giving the required insight and predictions due to the big data collected. This method will prove to not be effective to answer the business crucial questions.

Therefore, building an automated Machine Learning Model will provide a timely and accurate predictions compared to old statistical model.

## RELATED SYSTEMS

* + 1. HUMAN LABOR AND STATISTICS

Most companies depend on old statistics method for market analysis and predictions to make well informed data guided decisions. This has been used for such a long time and proved to be effective.

## LIMITATIONS OF RELATED SYSTEMS

### THE INSTITUTION WEBSITE

* The old statistics methods cannot be used for big data analyis.
* The old statistics methods are slow.
* The old statistics methods are prone to inaccuracies.

## HOW THE PROPOSED SOLUTIONS WILL HANDLE THE AFOREMENTIONED WEAKNESS

* The Model is able to handle big data very well.
* The Model is quick and automated in doing the data analysis and data prediction.
* The Model provides more accurate insight and prediction the statistical method.

# CHAPTER 3 – METHODOLOGY

## INTRODUCTION

In this chapter, I will describe the research methodology used for this project. I will explain how to design the proposed application, elaborate the procedures and process used in designing the application and data collection as well as provide methods on how to analyze the collected data.

## METHODOLOGY

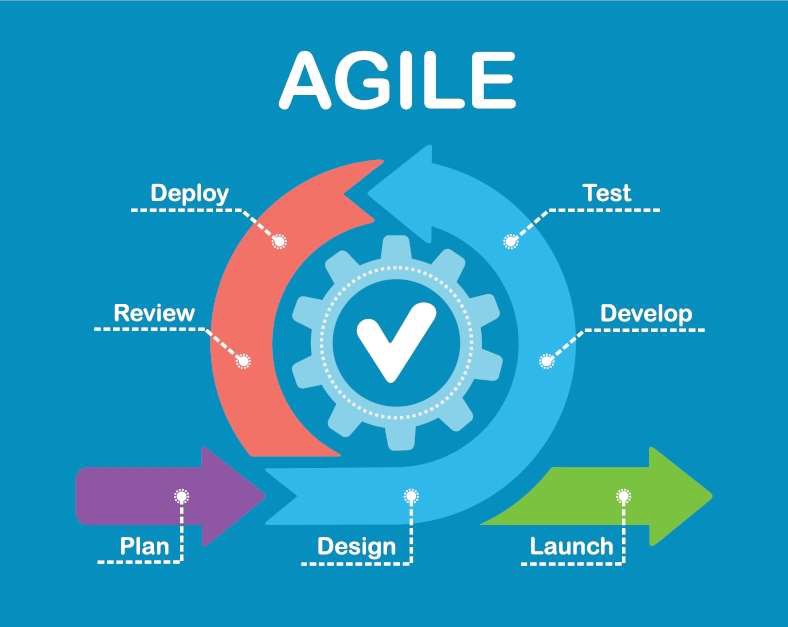
In this project, we will use Agile methodology. This choice is greatly influenced by the need to develop a working software with high speed, high quality and within limited time period.

### AGILE METHODOLOGY

This model is a combination of iterative and incremental process models with focus on adaptability and customer satisfaction by rapid delivery of working software project.

Agile methods break the product into small incremental builds which include the following: Planning, Requirements Analysis, Design, Coding, Unit Testing and Acceptance Testing. The major advantage of this model is that at the end of each iteration, I can be able to display the working product to my supervisor and any other stakeholder involved.

The graphical representation of agile model is as shown below: -



*Figure 1 Agile Model*

### ADVANTAGES OF AGILE MODEL

The advantages of Agile Model are as follows-

* + Good model for environments that change steadily.
  + Minimal rules, documentation easily employed.
  + Enables concurrent development and delivery within an overall planned context.
  + Little or no planning required.
  + Easy to manage.
  + Customer satisfaction.
  + Is a very realistic approach to software development
  + Promotes teamwork and cross training.
  + Functionality can be developed rapidly and demonstrated.
  + Resource requirements are minimum
  + Suitable for fixed or changing requirements
  + Delivers early partial working solutions.

### JUSTIFICATION FOR USING AGILE MODEL

**The following are the specific reasons as to why I chose agile model:**

* Adapting to change- Agile Development is focused on quick responses to change and continuous development making it easier to factor in new requirements. This is very vital since my Machine Learning app will be prone to changes.
* High Product Quality- Regular testing to see that the product is working during the development.

## DATA COLLECTION METHODS AND TOOLS

### WEBSITE

I as the researcher, interviewed some startup founders regarding their Data Science problems and their current methods and implementation s to curb their problems. Their responses were quite helpful in developing the system.

Justification

1. The Machine Learning Model was very accurate.
2. It allowed me to get real-time feedback and allowed cross-examination thus all the information I got was well verified.

# CHAPTER 4 – SYSTEM ANALYSIS

## INTRODUCTION

System analysis is a process of collecting and interpreting facts, identifying the problems and decomposition of a system into components.

System analysis is conducted for the purpose of studying the system or its parts in order to identify its objectives. It is a problem-solving technique that improves the system and ensures that all the components of the system work efficiently to accomplish their purpose.

## DETAILED ANALYSIS

### FLOW CHART

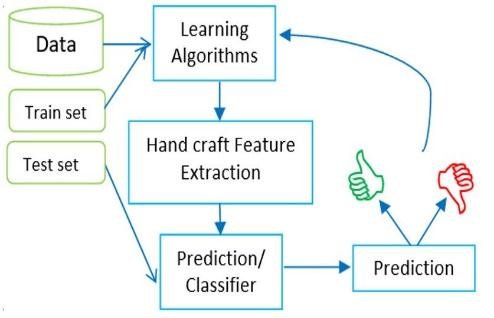
### 

### DATA FLOW DIAGRAM

Data Flow Diagram (DFD) provides a visual representation of the flow of information within a system. Using the level 0 DFD diagram shown below, you can easily tell the information provided by and delivered to someone who takes part in system processes, the information needed to complete the processes and the information needed to be stored and accessed.

Some of the importance of using a DFD include:

* + - 1. The diagram provided a basic understanding of how the system works.
      2. DFD simplifies the problem so as to make design stage easier.
      3. DFD may be drawn to represent different levels of details.



## SYSTEM REQUIREMENTS

System requirements are all of the requirements at the system level that describe the functions which the system as a whole should fulfill to satisfy the stakeholder needs and requirements, and are expressed in an appropriate combination of functional and non- functional requirements.

System requirements play major roles in systems engineering, as they:

* Act as reference for validation and stakeholder acceptance.
* Provide a means of communication between the various technical staff that interact throughout the project.
* Form the basis of system architecture and design activities.

I have classified the system requirements into functional and non-functional system requirements.

### FUNCTIONAL REQUIREMENTS

Functional requirements are product features or functions, that developers must implement to enable users to accomplish their tasks.

The following are the functional requirements of the proposed system:

* + The Application should be able to accept user input and pass it to the model.
  + The Application should be able to be able to effectively display the predicted value of the house based on the input entered by user.
  + The Application should be able to refresh for new user input after displaying the previous output.

### NON-FUNCTIONAL REQUIREMENTS

Non-functional requirements, as the name implies, are those that are not directly related to the specific services provided by the system to its users. They describe system aspects concerned with how the system meets functional requirements. Non-functional requirements are frequently referred to as a system's characteristics. Non-functional requirements pertain to the system's evolution through time rather than its execution.

Non-functional requirements include the following:

EASE OF USE: Given the level of knowledge possessed by the system's users, a simple but high-quality user interface should be designed that is easy to understand and require minimal training.

AVAILABILITY: The system should always be accessible 24 hours a day, seven days a week in the existence of an internet connection or a networked environment and electric power unless the network or internet connection fails or the electric power goes out.

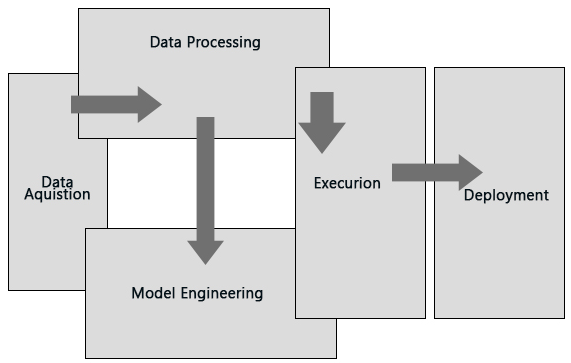
ACCURACY: The system should provide only relevant answers to the user based on his inputs.

# CHAPTER 5 – SYSTEM DESIGN

System design is the process of designing the elements of a system such as the architecture, modules and components, the different interfaces of those components and the data that goes through that system.

## ARCHITECTURAL DESIGN

Architectural design is a process for identifying the sub-systems making up a system and the framework for sub-system control and communication. The output of this design process is a description of the software architecture.



## DATABASE DESIGN

The database system implemented in the system is purely for storing inputs features that will be later sent to the model backend for prediction.

The proposed system works in real-time, so the necessity to have a database.

### INPUTS DESCRIPTION

This section contains a description of inputs to be entered in the system.

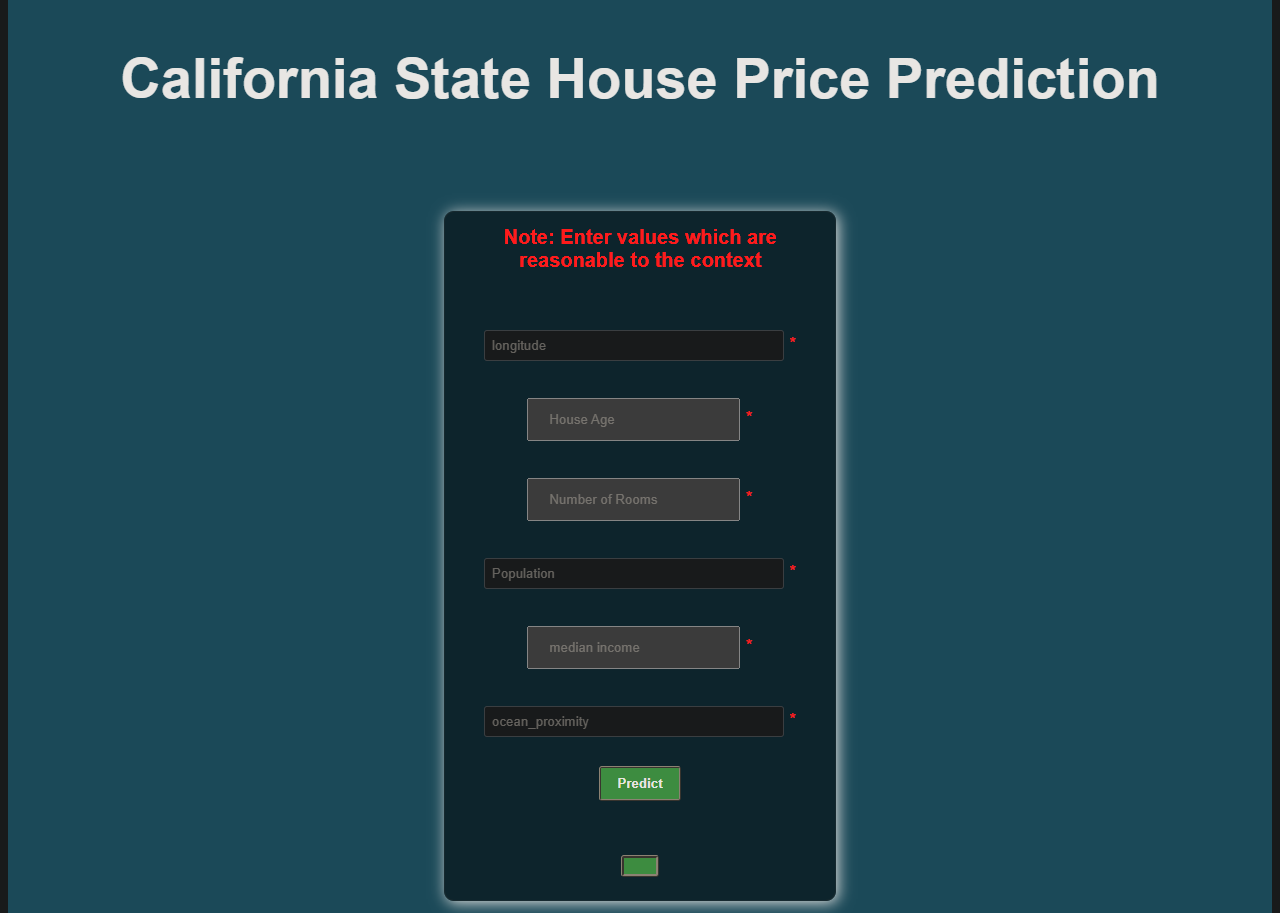
|  |  |
| --- | --- |
| Entity Name | Business Definition |
| longitude | A measure of how far west a house is; a higher value is farther west |
| House Median Age | Median age of a house within a block; a lower number is a newer building |
| Total Rooms | Total number of rooms within a block |
| Population | Total number of people residing within a block |
| Median Income | Median income for households within a block of houses (measured in tens of thousands of US Dollars) |
| Ocean Proximity | Location of the house with reference to ocean/sea |

|  |  |
| --- | --- |
|  |  |

## USER INTERFACE DESIGN

The user interface (UI) design process sits at the core of any human-centered product design project. From it, designers bring to life a page or product that stands out from the rest. UI design, involves the stylistic choices a designer makes when creating a product, e.g., an image, button, menu bar, or footer. All of these elements will affect the user’s interaction, and so, must be planned accordingly.

**HOME PAGE**



*Fig 1.Home Page*

# CHAPTER 6 - IMPLEMENTATION AND TESTING

## DEVELOPMENT AND ENVIRONMENT

A development environment is a collection of tools and procedures for developing testing and debugging an application or program. It helps developers to develop the application or product using a set of processes and programming tools.

The development environment normally has three server tiers, development server, staging server, production server.

California House Price Predictor was developed using Visual Studio Code.

## SYSTEM COMPONENTS

Every system comprises of basic components which come together to form the system, as a whole. Components involves input elements, processes, control mechanism, feedback system and objective.

The proposed system is to be developed using Python, HTML, CSS , Flask

### Data

Data is very important since it’s the primary building block of our project. Without the data we cannot train machine learning model to be able to make predictions

### Model

Model is a file that has been trained to recognize certain types of patter in data. This will be created from training our data through a machine learning algorithm. This file will be henceforth what will be used for making the valid predictions.

## 6.2.3 Code

This will be the full data analysis, model and data prediction code. It contains the full machine learning code that will be used to save the model and pickle it to be used in the flask App.

* + 1. Flask Application

This will enable us to build the back end of the application that will be used to connect the model pickle file and the UI used by the enterprise users to enter and output the information.

* + 1. UI

This will be the front-end side where the user will be able to interact with the machine learning model. I am going to implement html and CSS for this system requirement.

### SYSTEM TESTING

System Testing is a black box testing technique performed to evaluate the complete system the system's compliance against specified requirements. In System testing, the

functionalities of the system are tested from an end-to-end perspective. System Testing is usually carried out by a team that is independent of the development team in order to measure the quality of the system unbiased. It includes both functional and non-Functional testing.

The tested areas in the system include:

Model Performance and Accuracy

front end and back-end integration

# CHAPTER 7 – CRITICAL APPRAISAL

### ACHIEVEMENTS AND LESSONS LEARNT

The following are the lessons learnt during the implementation of the project:

* + - * 1. It aided in the development and polishing of my research techniques, report-writing, and communication abilities.
        2. I learned a variety of technical skills as a result of the project, including web design and development, Python programming, data science and coding.
        3. Application of knowledge acquired in school practically.

### CHALLENGES

During the development, the following challenges were encountered:

1. Building the Flask App to connect both front and back end was a challenge.
2. Building the Machine Learning model and testing it to check its accuracy really drained me but I learnt a lot in the process.

### RECOMMENDATIONS

The following are some of the extensions that can be made to this project:

1. Creation of Power BI for the dataset to give real time visualizations on data collected.
2. Building a separate model to cluster the data point and integrate the functionality in the current Application to improve further data mining of the dataset.

### CONCLUSIONS

The Houe Predicting Model system help in giving a quick prediction on house prices and data analysis. This improves enterprises decision making process which mean they will be able to make investment decisions.

## REFERENCES

1. Official Python documentation: <https://docs.python.org/3/>

2. Scikit learn documentation: <https://scikit-learn.org/stable/>

3. Flask documentation: <https://flask.palletsprojects.com/en/2.1.x/>