**Fundamentals of Data Mining – IT3051**

**Statement of Work (SOW) Report**

**USA Housing Listing**

A picture containing text, clipart, vector graphics

Description automatically generated

**Group Number: 25**

**Group Members:**

|  |  |
| --- | --- |
| **Name** | **Registration Number** |
| Biyanwila B. D. V. J | IT20212490 |
| Maddumage P. W | IT21007538 |
| Hewage R. P | IT21054686 |
| Kiriella K. G. A. K | IT21035876 |

Table of Contents

[1.Background 2](#_Toc146103506)

[2.Scope of Work 3](#_Toc146103507)

[3.Activities 4](#_Toc146103508)

[**1.Dataset Selection:** 4](#_Toc146103509)

[**2. Data Understanding:** 4](#_Toc146103510)

[**3. Identify the Problem:** 4](#_Toc146103511)

[**4. Data Preparation:** 4](#_Toc146103512)

[**5. Data Preprocessing:** 4](#_Toc146103513)

[**6. Data Visualization:** 4](#_Toc146103514)

[**7. Model Building:** 4](#_Toc146103515)

[**8. Front End Development:** 5](#_Toc146103516)

[**9. Combine the Model and UI:** 5](#_Toc146103517)

[**10. Deployment:** 5](#_Toc146103518)

[**11. Testing the Data:** 5](#_Toc146103519)

[**12. Documentation:** 5](#_Toc146103520)

[4.Approach 6](#_Toc146103521)

[5.Deliverables 7](#_Toc146103522)

[6. Project Plan and Timeline 8](#_Toc146103523)

[7.Assumptions 9](#_Toc146103524)

[8. Roles and Responsibilities 10](#_Toc146103525)

# 

# **1.Background**

**Dataset Name - USA – Housing Listings**

Craigslist is the world's largest online classifieds site, and it is a popular platform for finding apartments and houses for rent. This dataset was created to analyze short-term rental prices in the United States. The data was collected in January 2020 and contains over 300,000 housing listings. The dataset includes information such as region, rent per month, housing type, square footage, number of bedrooms, and number of bathrooms.

We are using this information to create a sample dataset and build a classification model that can suggest the type of housing that best meets a customer's needs based on their requirements. This will help customers who are unable to decide which type of housing is right for them.

**Identified Problem -** People are often overwhelmed by the number of housing options available to them. It can be difficult to know which type of housing is right for their needs, and it can be time-consuming to go through each property listing individually.

We want to build a system that can help people choose the right type of housing for their needs. This system would consider factors such as budget, desired location, and amenities. It would then generate a list of recommended properties that meet the user's criteria.

This system would make it easier and faster for people to find their dream home.

Dataset Link - <https://www.kaggle.com/austinreese/usa-housing-listings>

# 

# **2.Scope of Work**

For prospective buyers and renters, finding the ideal property has grown to be a difficult undertaking in the constantly changing world of real estate. Due to the wide range of housing options, individual tastes, and financial limitations, people are frequently confused about the kind of home that will best meet their needs. Our team has started working on creating a housing recommendation system to address this issue from data set.

Our goal is to use data and technology to streamline the house search process, and this is our clear aim. We have defined a thorough scope that details our group's activities and responsibilities in order to accomplish this. First and foremost, the acquisition of a home dataset from Craigslist that was gathered in January 2020 forms the basis of our study.

The foundation of our system is this information, which includes more than 300,000 housing listings from across the country. It gives us the data we need to build a classification model and make helpful recommendations. Data preparation is an essential phase that comes after data gathering. The dataset must be cleaned and transformed to ensure its quality and usefulness for analysis. This stage has significance for dealing with missing numbers, fixing errors, and getting the data ready for our model. In addition, we commit to handle sensitive information carefully and in keeping with data privacy laws. After we have cleaned up our data, we enter the data set we take for Finding patterns, trends, and insights within the dataset is the goal of this stage. Our goal is to have an in-depth understanding of the housing market, from regional pricing variances to the variables guiding preferences for particular living types.

For our classification model to be accurate, essential features must be carefully chosen and prepared. We create the necessary conditions for the effectiveness of our recommendation system by translating data into a modeling-friendly format. Our project's main focus is the creation of a living classification model. Based on user-specified variables, such as budget, square footage, bedrooms, and bathrooms, this model will predict the best form of housing. To guarantee the accuracy and dependability of the model, accurate development and validation procedures are essential. Our efforts ended up resulting in the development of a user-friendly recommendation system.

Users will be able to input their preferences and limits through this interface, giving them the ability to make well-informed judgements. Our system will be continuously improved by using user testing and feedback to improve its usability and accuracy. We want to enable consumers to easily and confidently find their perfect houses by combining data and technology. With this project, we hope to close the gap between housing seekers and their ideal homes, making the search of the perfect place to live a practical reality.

# **3.Activities**

## **1.Dataset Selection:**

The Craigslist housing dataset, which acts as the primary data source for our project, was carefully chosen at the beginning of the project.

## **2. Data Understanding:**

Analyzing the Raw Data: We explore the dataset to get an initial understanding of its variables, format, and contents.

Creating a suitable Scenario: Based on our understanding of the raw data, we create a suitable scenario that is in line with the goals of our project.

## **3. Identify the Problem:**

We recognize and describe the primary issue, which is to help people locate the best possible home option within their financial and other constraints.

## **4. Data Preparation:**

The dataset was divided into the training set and the testing set in order to make the process simpler to create models.

## **5. Data Preprocessing:**

Data Split: To ensure that we have separate datasets for training and testing our models, the data is split into subsets.

Data cleansing unwanted in order to improve the dataset's quality, unnecessary or irrelevant data is systematically removed.

Add New Data: To improve our analysis, we may add fresh data or features that we have obtained.

## **6. Data Visualization:**

We use data visualization tools to make insights and patterns easier to understand graphically, which improves our understanding of the housing consumers.

## **7. Model Building:**

Choose the modelling techniques from the list: We pick the best modelling techniques based on the problem statement.

Create the model by: We create the model for classifying house types.

Model Training: To enhance its performance, the model goes through extensive training.

Model assessment: Using a variety of evaluation measures, we systematically evaluate the model's effectiveness.

Choose the Best Model: The candidate with the best accuracy and performance is chosen.

## **8. Front End Development:**

Identify the Requirements: In order to provide an excellent user experience, we establish the needs for the user interface.

Design the UI (User Interface): An attractive UI is carefully created.

Implementation: The created user interface is implemented.

Deployment: The user interface is made available for general usage as the result of front-end development.

## **9. Combine the Model and UI:**

The user interface and the methodology for categorizing home types are merged, resulting in a full and useful system for recommending homes.

## **10. Deployment:**

The deployment of the system of suggestions makes it available to people looking for housing recommendations.

## **11. Testing the Data:**

A lot of testing is done to ensure the accuracy and dependability of the system's predictions.

## **12. Documentation:**

We carefully record every aspect of the project, including the data sources, preprocessing procedures, model architecture, user interface design, and system deployment information

# **4.Approach**

According to the Knowledge Discovery in Databases (KDD) principle in data mining, after data selection, data preprocessing, data transformation, data mining, and evaluation and interpretation steps are followed. For the selected ‘USA Housing Listings’ dataset, the classification method is used. And determine the specific datasets needed, including information on housing prices, property details, and geographical data.

After the selection step is done, the data set must be cleaned by handling missing values, inconsistencies, and outliers in the dataset integration by combining data from multiple sources, data transformation by converting data types, performing feature engineering, and encoding categorical variables and data reduction by reducing dimensionality if needed through techniques like PCA (Principal Component Analysis). By doing these, the data preprocessing and transformation parts were completed.

Perform exploratory data analysis (EDA) to gain a preliminary understanding of the data. Utilize data mining techniques to discover patterns, relationships, and insights within the dataset. This can include statistical analysis, clustering, classification, and association rule mining. Specific analyses could include determining common property types, identifying geographical clusters with similar price trends, or finding correlations between property features and prices. The class field ‘house type', which is a categorical variable, is chosen to be added as the label. In this step, supervised learning is used. The classification method is a machine-learning technique used to predict future values according to historical data.

Data evaluation assesses the quality and validity of the patterns and insights discovered during data mining. Use appropriate evaluation metrics and statistical tests to validate the results. This step helps ensure that the knowledge extracted from the data is reliable and meaningful. Represent the discovered patterns and insights in a structured and understandable form. This could include visualizations, statistical summaries, or models.

Interpret the knowledge and insights gained from the data mining process. And evaluate the relevance and significance of these findings in the context of the USA housing market. Identify actionable insights that could be valuable for various stakeholders, such as buyers, sellers, or real estate professionals. Evaluation and interpretation are done by developing a web application. For the implementation of the web application for backend Python language and its flask, framework is used for frontend Angular, and bootstrap is used. The application will be deployed using the Heroku cloud platform.

# **5.Deliverables**

We can use the Housing List dataset to build a classification model that can distinguish between different types of housing, such as apartments, houses, condominiums, duplexes, and townhouses. This model can be used to help users find the right type of housing for their needs.

To build the model, we will use a data mining technique called classification. Classification is a machine learning technique that can be used to predict the category that a new data point belongs to. In this case, we will train the model to predict the housing type of a new property based on its features, such as square footage, number of bedrooms and bathrooms, and location.

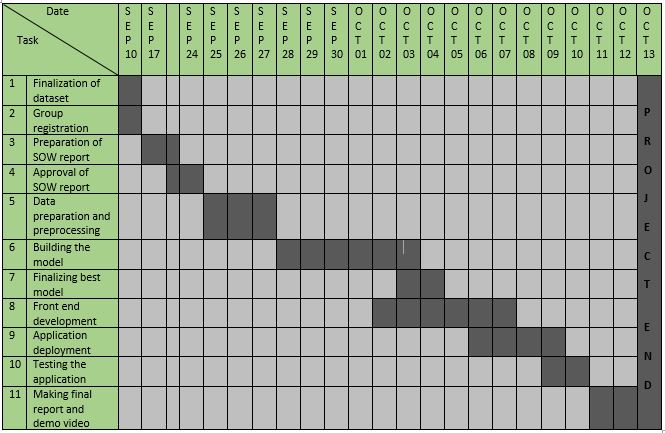
Once the model is trained, we can create a user-friendly interface that allows users to input their requirements. The model will then generate a list of recommended housing types that meet the user's criteria.

This system will make it easier for people to find the right type of housing for their needs. It will also save them time and effort by eliminating the need to go through each property listing individually.

# **6. Project Plan and Timeline**

|  |  |
| --- | --- |
| Start Date | September 10, 2023 |

|  |  |
| --- | --- |
| End Date | October 15, 2023 |



# **7.Assumptions**

* Assume that the selected dataset is an actual dataset
* Assume relationships which are discovered exist within the dataset
* Assume that the data are related to each other
* Assume that the given prediction column of the training data is correct
* Assume that the selected data are suitable for predicting frequent item sets and the best types of housing to rent
* Assume whether a person’s decision to rent a housing type is depending on the attributes used for the classification purposes

# **8.** **Roles and Responsibilities**

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Registration Number | Role | Responsibility |
| Biyanwila B.D.V.J. | IT20212490 | Data engineer | * Scope planning * Integrate * Testing performance * Documentation |
| Maddumage P.W. | IT21007538 | Data scientist | * Scope planning * Implement the model * Testing * Documentation |
| Hewage R.P. | IT21054686 | Data scientist | * Scope planning * Implement the model * Testing * Documentation |
| Kiriella K.G.A.K. | IT21035876 | Data analyst | * Scope planning * Analyzing and preprocessing data * Visualization * Documentation |

# 