DEVELOPMENT OF A PROPERTY MANAGEMENT, RENTAL PRICE PREDICTION AND RECOMMENDATION SYSTEM FOR SELANGOR (RentSel)

ZHANG CHENGJIN, Prof. Madya Dr. Dalbir Singh A/L Valbir Singh

Faculty of Information Science & Technology
Universiti Kebangsaan Malaysia
43600 Bangi, Selangor

ABSTRAK

Tahap pendidikan tinggi di Malaysia telah menunjukkan peningkatan yang berterusan dalam beberapa tahun kebelakangan ini, dan Selangor telah menghimpunkan sejumlah besar universiti berkualiti tinggi, menarik pelajar antarabangsa dari seluruh dunia. Penginapan merupakan salah satu kebimbangan dan keutamaan utama bagi pelajar antarabangsa yang tiba di Malaysia. Pelajar antarabangsa biasanya memilih untuk menyewa di luar kampus dan memilih rumah yang sesuai mengikut bajet mereka, terutamanya dalam kes penyewaan bersama, di mana kos biasanya lebih rendah. Dalam pasaran penyewaan, sejumlah besar maklumat sewaan telah muncul di laman web dan platform penyewaan utama. Disebabkan oleh kurangnya pengetahuan tentang persekitaran tempatan dan kerumitan pasaran sewaan, pelajar antarabangsa mungkin menghadapi beberapa cabaran dalam memilih tempat tinggal. Oleh itu, adalah sangat penting untuk mewujudkan model ramalan sewaan yang tepat dan boleh dipercayai bagi para peserta pasaran sewaan.

Kajian ini bertujuan untuk membangunkan satu sistem bagi pengurusan hartanah, ramalan harga sewaan dan cadangan rumah di Selangor. Sistem ini menggunakan Python untuk prapemprosesan data perumahan, menggunakan model Regresi Hutan Rawak (Random Forest Regression) untuk meramalkan kadar sewaan, dan membangunkan fungsi cadangan rumah melalui algoritma keserupaan kosinus (cosine similarity). Berdasarkan rangka kerja Flask dan pangkalan data MySQL, sistem ini membina modul pengurusan hartanah dalam bentuk aplikasi web, di mana pengguna boleh melihat maklumat rumah dan menghubungi tuan rumah. Tuan rumah pula boleh mengurus maklumat rumah dengan berkesan, termasuk fungsi menambah, menyunting dan memadam maklumat sewaan. Pelan integrasi ini bertujuan untuk menyediakan sokongan keputusan penyewaan yang lebih berinformasi kepada pelajar antarabangsa, penyewa tempatan dan pemilik hartanah, seterusnya mengurangkan ketidakpastian dalam pasaran Selangor dan meningkatkan kecekapan penyewaan di Selangor.

Kata kunci: Python, Regresi Hutan Rawak, Flask, MySQL

ABSTRACT

The level of higher education in Malaysia has continued to improve in recent years, and Selangor has gathered a large number of quality universities, attracting international students from all over the world. Accommodation is one of the top concerns and priorities for international students arriving in Malaysia. International students usually choose to rent outside and choose the right house according to their budget, especially in the case of shared accommodation, the cost is usually lower. In the rental market, a large number of rental information has emerged on major rental websites and platforms. Due to unfamiliarity with the local environment and the complexity of the rental market, international students may encounter some challenges when choosing accommodation. Therefore, it is of great significance to establish an accurate and reliable rent forecasting model for the participants of the rental market.

This study aims to develop a system for property management, rental price prediction and house recommendation in Selangor. The system uses Python to preprocess housing data, uses Random Forest Regression model to predict rent, and develops the housing recommendation function through cosine similarity algorithm. Based on the Flask framework and MySQL database to build a Web application property management module, users can view the housing information, contact the landlord. Landlords can effectively manage house information, including functions such as adding, editing, and deleting rental information. The integration plan aims to provide informed leasing decision support for international students, local tenants and property owners, ultimately reducing the uncertainty in the Selangor market and improving leasing efficiency in Selangor.

Keywords: Python, Random Forest Regression, Flask, MySQL

1.0 INTRODUCTION

The Malaysian government promotes international education and has attracted a large number of international students from China, India, Indonesia and other countries in recent years. As a popular country for studying abroad in Southeast Asia, the number of international students is expanding, and the rental population of apartments around the campus is growing rapidly. The fluctuation of housing rent is more and more concerned by international students, and the rent is determined by the decoration situation, location, housing pattern, transportation convenience, market supply and demand and other factors (Ogundunmade, 2023).

In the rental market, a large number of rental information has emerged on major rental websites and platforms. However, these massive amounts of data are often scattered and irregular, making it difficult to directly use them for analysis and prediction. Therefore, how to effectively extract valuable information from these data and then effectively predict the rental price has become an important issue in the field of rent research.

Many international students do not understand the rent level, the terms of the lease contract, and the suitable area for living, and are easy to choosing unreasonable prices or unsatisfactory environments. For the relatively traditional industry of rental housing, serious information asymmetry has always existed. Some agents tend to set rent prices at excessively high levels, divorcing them from prevailing market rates and the actual value of the properties. landlords face difficulties in renting out their properties, as they lack accurate information on the market price. Simultaneously, house hunters encounter challenges in finding housing options that offer genuine value for money amid inflated rent prices when renting and are prone to pay higher than the market rent due to insufficient information, which increases the rental cost (Ogundunmade, 2023).

The fluctuation of the rental price is often affected by many factors such as the location of the house, the size of the house, and the decoration situation. These factors interact with each other, making the forecast of rental prices complicated. Traditional forecasting methods are often based on experience or simple statistical analysis, it is difficult to accurately capture the law of price changes.

To address these issues, this project develops a web-based rental system (RentSel) that uses machine learning algorithms to build rental prediction models that provide reliable rental price predictions through key features such as location, region, and facility. In addition to house price prediction and recommendation, the system also contains a property management module, where users can view house information and contact landlords. Landlords can effectively manage housing information, including functions such as adding, editing, and deleting rental information. By combining data science, machine learning, and web application technologies,RentSel aims to increase transparency in the rental market, reduce uncertainty in decision-making, and provide a fairer and more efficient rental experience for international students, local tenants, and property owners.

2.0 LITERATURE REVIEW

At present, the research literature on predicting housing prices is quite extensive, most of the research focuses on the qualitative aspects such as influencing factors, pricing and process of rent. The determinants of housing rental prices vary by size. At the macro level, such as in the capital, housing rental prices are usually determined by economic fundamentals. Over the past few years, research using spatial econometrics to model house prices and rents has been at the forefront. Zheng Wenjuan studies China's urban housing price and housing rent from the perspective of "factor view", and uses panel data of 35 large and medium-sized cities in China to empirically study the key influencing factors of China's urban housing price and housing rent. Based on the perspective of asset pricing (Zheng, 2011). Chen Sichong and Chen Yingnan combined the standard dynamic Gordon growth model and the traditional housing use cost model to establish a dynamic housing use cost model about the rental yield of the housing market and used the variance decomposition method to investigate the factors affecting the dynamic fluctuations of the domestic housing market and their relative importance (Chen, 2015).

There has been a lot of research on using machine learning for real estate price prediction. Regression methods such as linear regression, Support Vector Regression(SVR), k-nearest neighbors(KNN), and random tree regression for real estate price predictions are widely used and have shown good results in the past. Shujia Zhao built a robust web-crawling spider that collects property information for real-time marketing and monitoring changes in property status, training a feature-weighted K-nearest neighbor (KNN) model to estimate home prices. Ferreira's research involved using multiple machine learning regression algorithms and feature engineering techniques to analyze and model the collected data to form the final model, XGBoost, as the most efficient model for accurate price predictions using parameters such as mean absolute error (MAE) and root mean square error (RMSE) (Ferreora, 2024). Li Jinze's research targets pain points in the rental market and is based on real rental market data after desensitization. Using the historical data of monthly rent labels, the LightGBM (Light Gradient Boosting) model based on machine learning is established to provide an accurate prediction of monthly housing rent based on basic housing information and provide an objective measure for the urban rental market. Based on the latest machine learning algorithm (Li, 2018). Ma Tao and Liu Ningning adopted the integrated strategy of model fusion (MA, 2019), and used some characteristic data of rental houses to predict their rents. Tomal's study attempts to identify the determinants that influence rental prices in Krakow. Obtained from the Internet platform otodom.pl using web scraping techniques. To determine the determinants of rent, ordinary least squares (OLS) regression and spatial econometrics methods were used. In particular, the traditional spatial autoregressive model (SAR) and spatial autoregressive weighted regression (GWR-SAR) are used to consider the spatial heterogeneity of the determinants and the spatial autocorrelation of the spatial changes of housing rents. An in-depth analysis of rental determinants using the GCR-SAR model reveals the complexity of the rental market in Kraków (Tomal, 2020).

Also worth noting is that LirongHu demonstrated how to monitor fine-scale housing rental prices based on OHRWs using the case of Shenzhen in China. Employing a hedonic model, a set of housing rental determinants is initially selected from three characteristics (neighborhood, location, and structure) and at three levels (nearest accessibility, 15-minute walking distance availability, and sub-district availability). Housing rent prediction models are then established (respectively for October 2017 and February 2018) using the training samples collected from the OHRWs and six machine-learning algorithms, including random forest regression (RFR), extra-trees regression (ETR), gradient-boosting regression (GBR), support vector regression (SVR), multi-layer perceptron neural network (MLP-NN) and k-nearest neighbor algorithm (k-NN). Thereafter, the relative importance of the determinants is calculated and visualized using partial dependence plots. Finally, the models are used to monitor housing rental price dynamics for all of the communities within Shenzhen (Hu, 2019).

3.0 METHODOLOGY

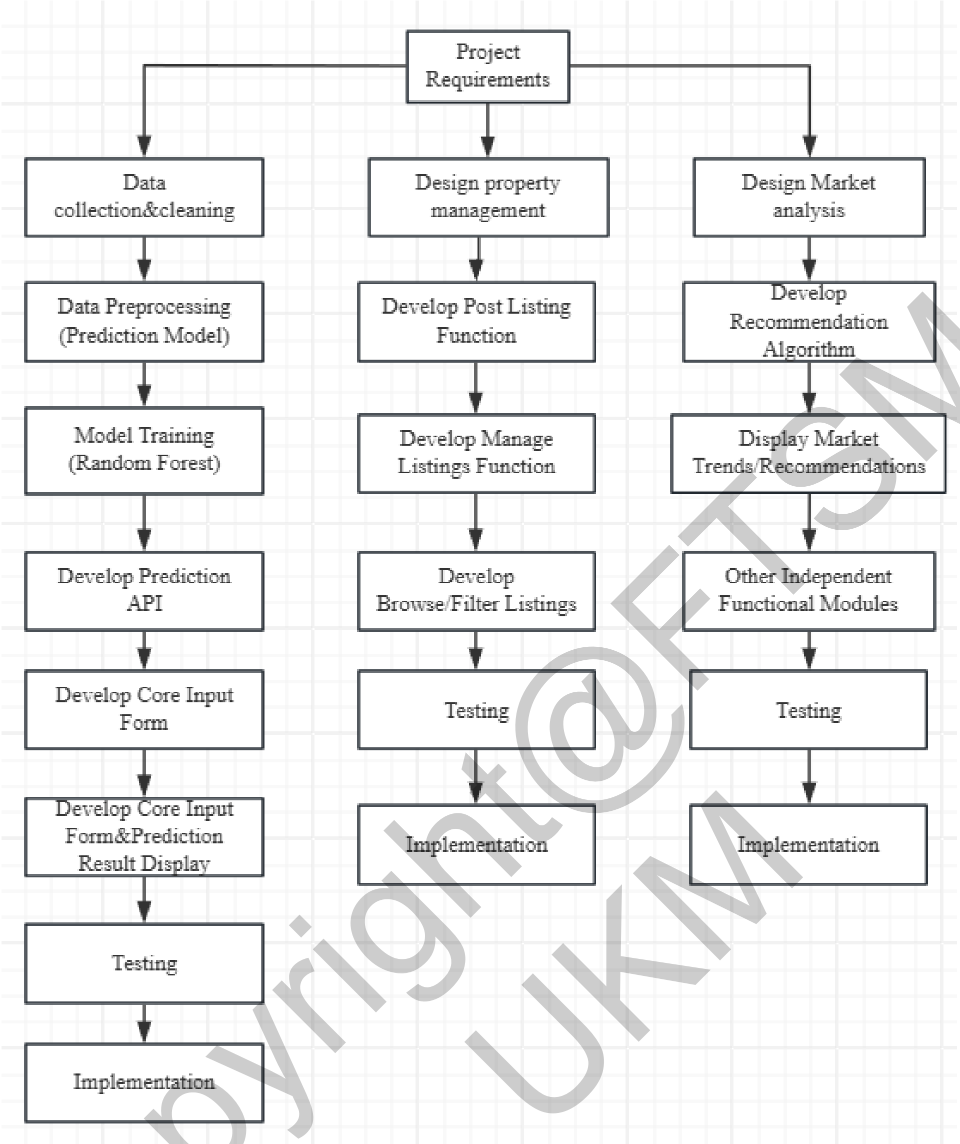

Figure 1.1 Incremental development process

Referring to Figure 1.1, for the rent price forecasting system, choose to use an incremental development model. The project is characterized by a number of relatively independent modules such as data acquisition and cleaning, predictive model construction and optimization, user interface and the overall integration of the system, so each module can be gradually developed and tested, and finally, the construction of the system can be completed. Rental price forecasting system involves data acquisition, data cleaning, model training, interface development, and other relatively independent functional modules, each module can be developed and tested separately. The incremental development model provides the flexibility to manage these modules and adjust them as each increment is completed. It can effectively reduce development risk and improve development efficiency while ensuring system functionality and flexibility.

3.1 System flow chart

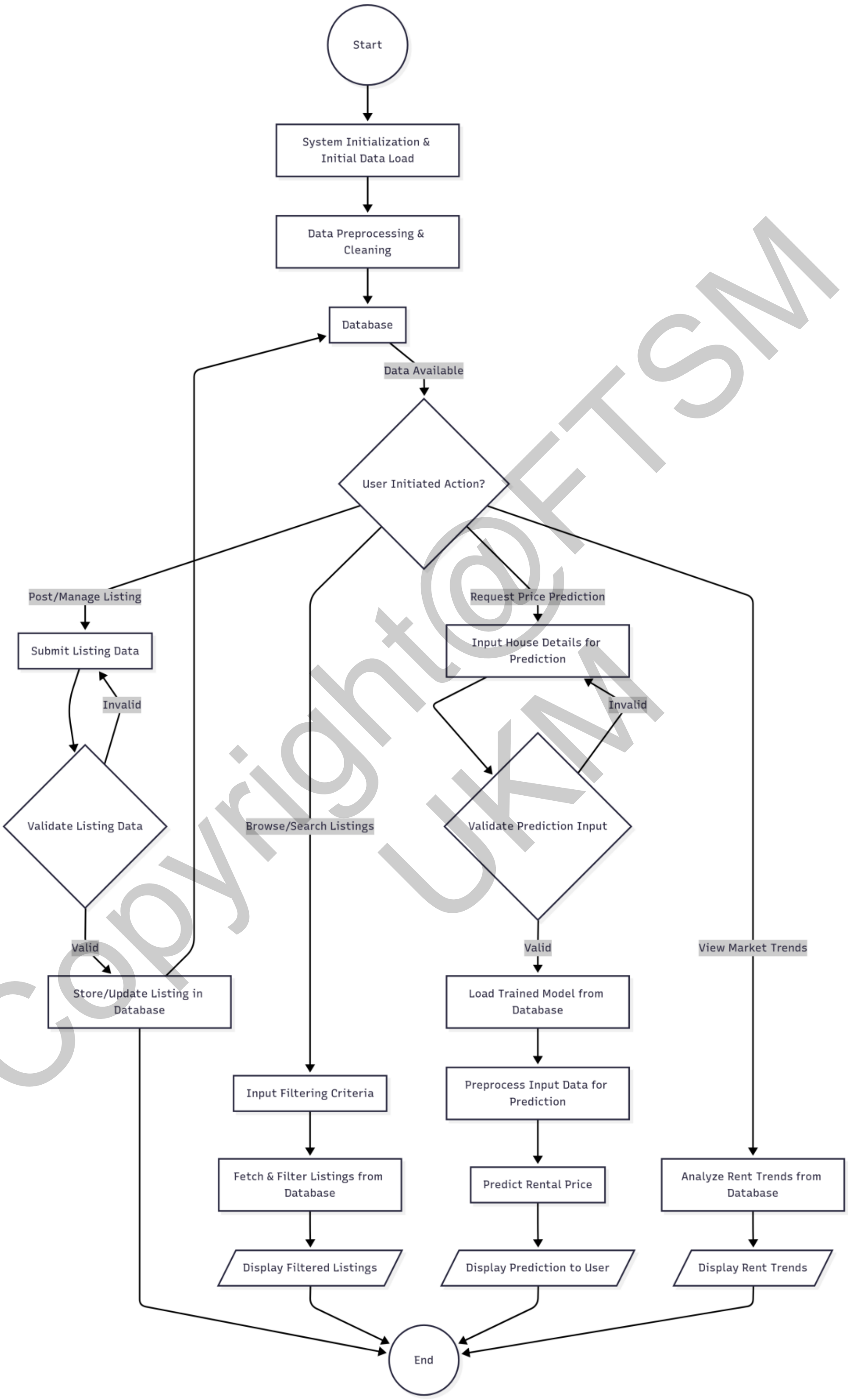

Figure 3.1 System flow chart

Figure 3.1 illustrates the technical architecture and data flow of the "RentSel" platform. The process begins in the system initialization stage, where the housing dataset is loaded, preprocessed, and cleaned, and then stored in the central database. Subsequently, the system is driven by user-initiated operations and branches into one of the four core modules. For property management, users can submit property data, which will be verified and stored in the database. To request price predictions, the system will verify the input data, load the trained random forest model, preprocess the data, and generate rental predictions. Users can also browse the property list using filtering conditions or view market trends derived from database analysis. Each path presents a complete and modular process, and it ends in the "end" state after meeting the user's request.

3.3 Architectural design

Figure 3.2, the system uses the Model-View-Controller (MVC) architecture, it supports application development based on user interaction well and is easy to extend and maintain. Models are the components in a system application that actually perform the work. The controller is used to send messages to the model and to provide an interface between the model and its associated views and the interactive user interface devices. Views handle all graphical things, they request data from the model and display the data.(Bucanek, 2009)

Model: Handles data logic, including data collection (crawling), cleaning, storage, and training and prediction of machine learning models.
View: Displays prediction results, provides a user interface, and supports interaction with the user.
Controller: Processes user input, calls the model to make data predictions and returns the results to the view.

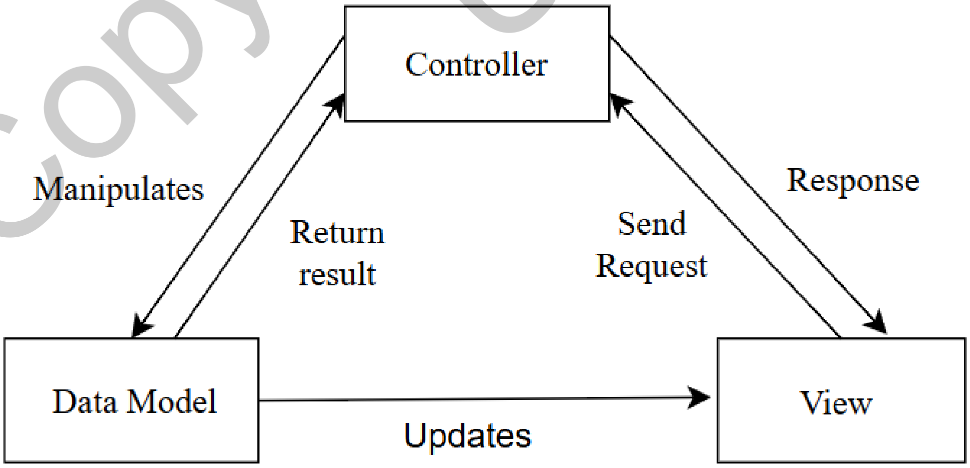

Figure3.2 Model-View-Controller State diagram

3.3 Dataset

To develop an accurate and reliable rental price prediction system, this project uses a real-world housing dataset collected from the Mudah.my platform, which was sourced via Kaggle. The dataset consists of 19,991 entries and 14 columns, covering residential rental listings in Kuala Lumpur and Selangor.

Referring to Figures 3.3, after data preprocessing, the dataset contains 13,533 records and 11 features, with the main attributes as follows:

1. prop_name: Property name (e.g. "Arte Mont Kiara")
2. monthly_rent: Monthly rent represented as a string (e.g. "Monthly 1500 US dollars")
3. location: Specific location (e.g. "Kuala Lumpur - Leksar")
4. rooms: Number of bedrooms
5. bathroom: Number of bathrooms
6. parking: Number of parking spaces (with missing values)
7. size: Building area (in square feet, stored as a string, e.g. "850 square feet")
8. furnished: Decoration status (e.g. "Fully furnished", "Partially furnished", or "Unfurnished")
9. facilities: Facilities provided (e.g. "Swimming pool", "Gym", "Security")
10. additional_facilities: Additional features such as air conditioning or proximity to a subway station
11. region: Main area where the property is located - Kuala Lumpur or Selangor

1	prop_name	monthly_re	location	rooms	parking	bathroom	size	furnished	facilities	additional	region
2	Segar Cour	2300	Kuala Lump	3	1	2	1170	Partially	Playground	Air-Cond,	Kuala Lumpur
3	Sentul Poi	1700	Kuala Lump	2	1	2	743	Partially	Parking, I	Cooking A	Kuala Lumpur
4	Arte Mont	1299	Kuala Lump	1	1	1	494	Not Furnis	Parking, S	Air-Cond	Kuala Lumpur
5	Residensi	1500	Kuala Lump	3	1	2	884	Partially	Parking, S	Cooking A	Kuala Lumpur
6	Sky Meridi	2900	Kuala Lump	3	2	2	982	Fully Furn	Parking, S	Air-Cond,	Kuala Lumpur
7	Arte Plus	1550	Kuala Lump	1	1	1	700	Fully Furn	Parking, (	Air-Cond,	Kuala Lumpur
8	Nova I	1400	Kuala Lump	2	1	1	750	Fully Furn	Playground	Air-Cond,	Kuala Lumpur
9	Sofiya Res	1350	Kuala Lump	3	1	2	862	Partially	Playground	Air-Cond,	Kuala Lumpur
10	PV9 Reside	2000	Kuala Lump	4	2	2	1100	Partially	Parking, S	Air-Cond,	Kuala Lumpur
11	Arte Plus	1500	Kuala Lump	1	1	1	700	Fully Furn	Gymnasium,	Air-Cond,	Kuala Lumpur
12	Maxim Citi	1300	Kuala Lump	3	1	2	1009	Fully Furn	Minimart,	Air-Cond,	Kuala Lumpur
13	Legasi Kar	3200	Kuala Lump	3	1	2	950	Fully Furn	Lift, Squa	Air-Cond,	Kuala Lumpur
14	Legasi Kar	3200	Kuala Lump	3	1	2	950	Fully Furn	Lift, Squa	Air-Cond,	Kuala Lumpur
15	Majestic M	1400	Kuala Lump	2	2	2	650	Not Furnis	Parking, (	Air-Cond,	Kuala Lumpur
16	Desa Villa	2500	Kuala Lump	4	2	2	1784	Not Furnis	Parking, S	Air-Cond,	Kuala Lumpur
17	East Side	1800	Kuala Lump	3	1	2	1100	Fully Furn	Gymnasium,	None	Kuala Lumpur
18	Majestic M	1099	Kuala Lump	2	1	2	650	Not Furnis	Playground	Air-Cond,	Kuala Lumpur
19	Majestic M	1099	Kuala Lump	2	1	2	650	Partially	Security,	Air-Cond,	Kuala Lumpur
20	Majestic M	1199	Kuala Lump	3	1	2	819	Not Furnis	Security,	Air-Cond,	Kuala Lumpur

Figure 3.3 l Selangor database

3.4 ALGORITHM DESIGN

To predict rental prices, this system implements a Random Forest Regressor, an ensemble machine learning algorithm known for its high accuracy and robustness against overfitting. The entire process, from data preparation to prediction, is encapsulated within a Scikit-learn Pipeline to ensure consistency and streamline deployment.

- Data Preparation and Preprocessing: The selangor_data.csv dataset is loaded, and the features (location, rooms, size, etc.) are separated from the target variable (monthly_rent).

The data is then split into an 80% training set and a 20% testing set. A ColumnTransformer is used to apply feature-specific preprocessing: numerical features like size and rooms are scaled using StandardScaler, while categorical features like location and furnished are transformed into a numerical format using OneHotEncoder.

- Model Training and Hyperparameter Optimization: A RandomForestRegressor is integrated into the pipeline. To achieve optimal performance, its hyperparameters are tuned using RandomizedSearchCV. This method systematically searches through 150 different combinations of parameters (such as the number of trees and their depth) using 5-fold cross-validation. This rigorous process identifies the best-performing model configuration, which is then trained on the full training dataset.
- Deployment and Prediction: The entire pipeline, including the preprocessing steps and the trained model, is serialized and saved into a single file (rent_model.joblib). For making a new prediction, this file is loaded, and the user's input is passed to it. The pipeline automatically handles all the necessary data transformations before feeding the data to the model, ensuring that predictions are both consistent and efficient.

```
Final Model Evaluation Results:
Mean Absolute Error (MAE): 208.07
Root Mean Squared Error (RMSE): 360.20
R-squared (R²): 0.7586
```

Figure 3.4 Model Evaluation

Model Evaluation: The performance of the final, optimized model is evaluated on the unseen test data. The results indicate a strong predictive capability, with a Mean Absolute Error (MAE) of RM 208.07 and an R-squared (R^2) value of 0.7586, signifying that the model can explain approximately 75.9% of the variance in rental prices.

4.0 RESULTS

4.1 Important modules

- app.py: The main entry point of the project, responsible for all the main business logic of Flask application initialization, route definition, user authentication, property management, history, recommendation and prediction.
- filter_selangor.py: The data cleaning and preprocessing module screens, cleans, handles outliers and standardizes the raw housing data to generate a high-quality dataset for modeling.
- rent_prediction.py: The Machine learning Modeling and prediction module is responsible for feature engineering, model training, model saving and loading, and renting the prediction interface.
- recommend_location.py: The recommendation algorithm module recommends suitable housing or locations based on user requirements such as budget, number of bedrooms and furniture.

- rent_visualization.py: Data visualization module, generate rent distribution and other charts to help users understand the market conditions.

4.2 Important components

- Flask router: Responsible for front-end and back-end interaction, including publishing, editing, deleting, displaying listings, registering users, managing history, etc.
- User authentication and session management: Achieve user login, registration, permission verification and other functions through Flask session.

4.3 Important libraries

- Flask: The core Web development framework that handles the application's URL routing, HTTP requests, and template rendering.
- Flask-SQLAlchemy: Object-Relational Mapping (ORM) library that simplifies database manipulation by mapping database tables to Python classes.
- Flask-Migrate: A database migration management tool for safe and orderly updates of database table structures during development and maintenance.
- Pandas: The main data processing and analysis library for performing critical tasks such as data cleaning, feature engineering, and reading and writing data.
- Scikit-learn: An open-source machine learning library written in Python that provides model training, feature normalization, data splitting, and prediction(Kramer, 2016).
- Joblib: Joblib is a library designed to provide lightweight pipelining in Python. It is particularly effective for tasks involving large arrays or datasets. Joblib's caching mechanism helps avoid recomputation, making it highly efficient for iterative algorithms(ZERBO, 2024).
- Matplotlib/Seaborn: Professional data visualization libraries, both used in combination to produce statistically significant plots.
- PyMySQL / mysql-connector-python: mysql-connector-python - Python-based client developed by Oracle. PyMySQL - Fully Python-based MySQL client(Ramler, 2020).

4.4 Test Results

In terms of functional testing, this system employs the black box testing method to verify whether all the functions related to rent can operate normally, ensuring that the system meets the user's requirements and identifying and resolving potential issues. A black box testing is a testing in which internal specifics and workings aren't known or accessible to its customer. It supports specifications and output needs. The fundamental purpose is to identify the requirements of the system(Anwar, 2019).The specific tests will be as follows：

- Functional testing: Verify whether each functional module is implemented according to the requirements.
- Interface testing: Checks that the page displays and interacts properly.
- Exception handling testing: Tests the ability of the system to handle illegal inputs and abnormal operations.

Table 4.4 below shows that all 22 test cases have been successfully completed, and the actual results are in line with the expectations. All the functional modules of the system run smoothly and meet the project requirements.

Test Area	Number of Cases	All Passed
Functional testing	9	Yes
Interface testing	6	Yes
Exception handling testing	7	Yes

Table 4.4 Testing module

5.0 CONCLUSION

The "RentSel" system successfully achieved key functions such as rental property management, price prediction, and personalized housing recommendations by combining machine learning with network development technology. It utilized a random forest regression model trained on the Kaggle real estate dataset to generate rental predictions and matched user preferences with suitable properties using cosine similarity.

System testing using the black-box testing method confirmed its reliability and usability, indicating that RentSel can effectively achieve its intended goals. Users benefit from a simple, responsive interface, interactive visualization tools such as rent trend display, and features like browsing history record tracking, which collectively enhance their decision-making process.

Although RentSel has achieved some accomplishments, it also has obvious limitations. Due to the technical challenges brought by real-time data extraction, it currently relies on static datasets and the adopted model is compared to more advanced integrated algorithms as a basis. These limiting factors affect the system's long-term adaptability and prediction accuracy.

Future plans include improvements in three main areas: real-time data collection through web crawlers, optimizing the machine learning model using more powerful algorithms (such as XGBoost or LightGBM), and expanding the system's geographical coverage to other cities in Malaysia. These improvements aim to enhance prediction accuracy, system response speed, and overall user experience, ensuring that RentSel remains a practical and scalable solution for the rental market.

Through its systematic development process and clear application goals, RentSel demonstrates the potential of intelligent systems in addressing rental issues for international students and other groups.

ACKNOWLEDGEMENT

I would like to sincerely thank my supervisor, Prof. Madya Dr. Dalbir Singh A/L Valbir Singh, for his continuous valuable guidance in the process of completing my thesis. It is my honor to study under his guidance. Thanks to his patience, support, tolerance, and encouragement, I have the confidence and courage to move forward.

6.0 REFERENCES

Ogundunmade, T. P., Abidoye, M., & Olunfunbi, O. M. (2023). Modelling Residential Housing Rent Price Using Machine Learning Models. Mod Econ Manag, 2.

Tomal M. Modelling Housing Rents Using Spatial Autoregressive Geographically Weighted Regression: A Case Study in Cracow, Poland. ISPRS International Journal of Geo-Information. 2020; 9(6):346.

Zheng W J. (2011). Research on the Influencing factors and Relationship between housing price and housing rent in Chinese cities (in Chinese). Hangzhou: PhD dissertation of Zhejiang University.

Ferreira, J. L. V. E. (2024). A Machine Learning Approach to Forecasting House Prices in the Portuguese Market.

Li, J. (2018). Monthly housing rent forecast based on lightgbm (light gradient boosting) model. International Journal of Intelligent Information and Management Science, 7(6), 09-03.

Ma, T., & Liu, N. (2019). Research of prediction on house rent based on intergration learning. Finance, 9, 586-594.

Chen, S., & Chen, Y. (2015). What Drives the Housing Markets in China: Rent, Cost of Capital, or Risk Premium of Owning Versus Renting?. Cost of Capital, or Risk Premium of Owning Versus Renting.

Hu, L., He, S., Han, Z., Xiao, H., Su, S., Weng, M., & Cai, Z. (2019). Monitoring housing rental prices based on social media: An integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies. Land use policy, 82, 657-673.

Zhao, S. (2018). Implementation and Study of K-Nearest Neighbour and Regression Algorithm for Real-time Housing Market Recommendation Application.

Ramler, P. H. (2020). Laimonas Suchockas.

Kramer, O., & Kramer, O. (2016). Scikit-learn. Machine learning for evolution strategies, 45-53.

ZERBO, A., OUEDRAOGO, M., SERE, A., & DIARRA, M. (2024). Comparison of Joblib and Pymp for Parallel Fingerprint Recognition.

Gupta, A., & Bhatia, R. (2010). Testing functional requirements using B model specifications. ACM SIGSOFT Software Engineering Notes, 35(2), 1-7.

Anwar, N., & Kar, S. (2019). Review paper on various software testing techniques & strategies. Global Journal of Computer Science and Technology, 19(2), 43-49.

ZHANG CHENGJIN(A197996)
Prof. Madya Dr. Dalbir Singh A/L Valbir Singh
Fakulti Teknologi & Sains Maklumat
Universiti Kebangsaan Malaysia