REAL ESTATE INVESTMENT SAFETY PREDICTION USING MACHINE LEARNING

CS19643-FOUNDATIONS OF MACHINE LEARNING

Submitted by

SANJANA SHREE S (220701245)

In partial fulfilment of the award of the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

ANNA UNIVERSITY, CHENNAI

MAY 2025

RAJALAKSHMI ENGINEERING COLLEGE

CHENNAI - 602105

BONAFIDE CERTIFICATE

Certified that this Report titled "**REAL ESTATE INVESTMENT SAFETY PREDICTION USING MACHINE LEARNING"** is the bonafide work of **SANJANA SHREE S (220701245)** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Mrs. M. Divya M.E.

Supervisor

Assistant Professor

Department of Computer Science and Engineering

Rajalakshmi Engineering College,

Chennai – 602105

Submitted to Project Viva-Voce Examination held on _____

Internal Examiner **External Examiner**

TABLE OF CONTENTS

ACKNOWLEDGEMENT

Initially we thank the Almighty for being with us through every walk of ourlife and showering his blessings through the endeavor to put forth this report. Our sincere thanks to our Chairman **Mr. S. MEGANATHAN, B.E,F.I.E.**, our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN**, **Ph.D.,** for providing us with the requisite infrastructure and sincere endeavoring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.,** our belovedPrincipal for his kind support and facilities provided to complete our workin time. We express our sincere thanks to **Dr.P.KUMAR, Ph.D.**, Professorand Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey oursincere and deepest gratitude to our internal guide, **Mrs. DIVYA M**, **M.E.,** Department of Computer Science and Engineering. Rajalakshmi Engineering College for her valuable guidance throughout the course of the project. We are very glad to thank our Project Coordinator, **Dr.K.ANATHAJOTHI, M.E, Ph.D.,** Department of Computer Science and Engineering for his useful tips during our review to build our project.

SANJANA SHREE S 220701245

ABSTRACT

In the real estate sector, making safe and profitable investment decisions is a challenge due to the multitude of factors influencing property value and risk. This project, titled **"Real Estate Investment Safety Predictor"**, aims to leverage machine learning techniques to predict whether a real estate property is a safe investment based on various input features such as location characteristics, market trends, and property attributes. A synthetic dataset was used for analysis, and the workflow included data preprocessing, exploratory data analysis (EDA), feature scaling, and model training using multiple classification algorithms including Logistic Regression, Decision Tree, and Random Forest. Evaluation was conducted using metrics such as accuracy, precision, recall, and F1-score. Among the models tested, the Random Forest Classifier emerged as the best-performing model. The final model was integrated into a user-friendly Streamlit web application, allowing users to input property details and receive an instant investment safety prediction. The project not only demonstrates the power of machine learning in risk analysis but also offers a scalable solution for real-world deployment. It bridges the gap between data science and real estate investment through interactive, intelligent decision support. This solution has potential for extension using real-world datasets and additional economic indicators.

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 GENERAL

Real estate investments play a crucial role in wealth creation and economic development, but they are also accompanied by significant financial risks due to market volatility, regional disparities, and fluctuating economic conditions. Factors such as property location, infrastructure development, neighborhood safety, historical price trends, and economic indicators greatly influence the potential success or failure of an investment. Therefore, accurately assessing whether a property is a safe investment is essential for minimizing risks and maximizing returns for both individual buyers and institutional investors.

Traditionally, investment decisions in the real estate sector have relied heavily on expert opinions and manual analysis, which can be time-consuming, subjective, and prone to errors. With the advent of machine learning, it is now possible to analyze vast amounts of structured and unstructured data to uncover hidden patterns and make data-driven predictions. This project leverages machine learning techniques to build a predictive model that evaluates the safety of real estate investments based on key features from a curated dataset.

By integrating automated data preprocessing, model selection, and evaluation into a streamlined pipeline, this solution offers a practical and intelligent decision-support system for investors. The final model is deployed as an interactive web application using Streamlit, making advanced predictive analytics accessible to users in real-time. This end-to-end system not only enhances investment decision accuracy but also serves as a scalable foundation for future real estate intelligence tools.

1.2 OBJECTIVE

The primary objective of the proposed project, **"Real Estate Investment Safety Predictor using Machine Learning"**, is to develop an end-to-end predictive framework that accurately determines the investment safety of real estate properties using classification algorithms. The model aims to evaluate various factors such as location features, economic indicators, and market trends to identify whether a given property is a safe investment or not. Through comprehensive data preprocessing, feature analysis, and model comparison, the project seeks to enhance the accuracy and reliability of investment predictions. To ensure practical usability, the best-performing model—Random Forest in this case—is integrated into a user-friendly web application that allows real-time predictions based on user input. This interactive tool supports informed decision-making for both individual and institutional investors. Additionally, the project framework is designed to be scalable and adaptable for integration with real-world datasets, enabling future expansion into intelligent real estate analytics platforms. The system's effectiveness will be assessed based on predictive accuracy, and its potential to minimize investment risks in the real estate domain.

1.3 EXISTING SYSTEM

Traditional real estate investment analysis primarily relies on manual evaluation, domain expertise, and basic statistical methods to assess the safety and profitability of a property. These conventional approaches often involve analyzing historical market trends, location-specific factors, and economic indicators, typically through spreadsheets or static dashboards. While useful, these systems are limited by human subjectivity, scalability issues, and their inability to process large and complex datasets effectively. In recent years, basic machine learning models such as Linear Regression or Decision Trees have been applied to real estate pricing and trend prediction. However, many of these models are focused solely on price estimation rather than evaluating investment risk or safety. Additionally, these models often lack robust preprocessing pipelines, and they do not handle feature scaling, outlier detection, or categorical encoding efficiently.

Moreover, existing systems frequently overlook the importance of a user-friendly interface, which limits their accessibility to non-technical users such as individual investors or small real estate firms. Finally, many of these models are not deployed in a practical form (e.g., web or mobile apps), making real-time predictions unavailable to end users. These shortcomings highlight the need for a comprehensive, interactive, and accurate machine learning solution that can predict real estate investment safety and be easily integrated into real-world decision-making processes.

1.4 PROPOSED SYSTEM

The main goal of the proposed system is to develop an intelligent framework for predicting the safety of real estate investments using machine learning techniques, specifically leveraging a Random Forest Classifier. This system aims to assist investors in making informed decisions about property investments based on a variety of factors, including location, price, property age, area, and other relevant features. The system evaluates the safety of investments by predicting whether a real estate property is a “Safe” or “Not Safe” investment.

The proposed model integrates a Random Forest-based approach, which is a powerful ensemble learning method known for its robustness and ability to handle high-dimensional data. Random Forest is particularly suited for this application due to its ability to manage complex, non-linear relationships between input features and target outcomes, providing a high level of accuracy in classification tasks.

To evaluate the system’s performance, the model is trained and validated on a synthetic real estate dataset, and key performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC are used to compare its effectiveness. The system's goal is to ensure that it can make reliable

predictions based on historical data, thereby reducing the risks associated with real estate investments.

For an interactive user experience, the system includes a user-friendly interface built using Streamlit. This interface allows users to input features such as price, area, and location, and instantly receive a prediction about whether the property is a safe investment or not. The web app is designed to be intuitive and accessible to both experienced and novice investors, providing immediate feedback based on the trained model's predictions.

In addition to the machine learning model, the system employs secure data handling protocols to ensure the privacy of user information. The application is deployed on Streamlit Cloud, allowing real-time access to the model via a secure and scalable web platform.

The motivation for this project is to create a tool that helps potential real estate investors make data-driven decisions. By utilizing machine learning, the system can predict the safety of an investment, thus offering a reliable way to minimize risks and maximize returns in the real estate market. Furthermore, the deployment of the system on the cloud ensures that users can access it from anywhere, promoting flexibility and ease of use. Ultimately, this project seeks to empower investors with the tools to analyze and assess the risk factors associated with real estate investments effectively, thus improving decision-making and financial outcomes.

Furthermore, the system can be extended to incorporate real-time market trends and economic indicators to further enhance prediction accuracy. Future enhancements may also include integration with property listing platforms, enabling seamless evaluation of live investment opportunities.

CHAPTER 2

LITERATURE SURVEY

[1] Babu and Chandran (2019), This study presents a comprehensive review of machine learning techniques for predicting real estate prices and investment safety. It emphasizes how algorithms like Random Forest and Decision Trees can capture non-linear relationships between property attributes and market value. The paper highlights the importance of feature engineering and model interpretability in real estate forecasting. It also notes limitations in handling sparse and noisy data in traditional models**.**

[2] Jha et al. (2020), The authors compare multiple machine learning algorithms (Linear Regression, Decision Trees, Random Forest) for real estate price prediction. Their experimental analysis shows Random Forest provides superior accuracy due to its ensemble nature. The paper underlines the challenge of selecting the most influential features and emphasizes the value of combining location, area, and historical pricing in predictive models.

[3] Al-Qawasmi (2022), This paper reviews the evolution of AI applications in the real estate domain, focusing on how intelligent systems enhance investment safety analysis. It discusses the role of feature extraction and automated decision support in reducing human bias. The study identifies gaps in current models, particularly in incorporating real-world uncertainties and financial risk.

[4] Yazdani et al. (2021), The authors present a comparative analysis of deep learning, hedonic, and machine learning methods for real estate price prediction. Their findings show that traditional ML algorithms like Gradient

Boosting and XGBoost outperform deep learning on smaller datasets. The paper also identifies issues of overfitting and stresses the importance of model validation in safety-critical applications like investment forecasting.

[5] Zhang et al. (2024), This paper explores the use of time series and machine learning models in forecasting real estate price trends. It integrates socioeconomic variables and location-based metrics to improve investment predictions. The study demonstrates how feature importance techniques help investors understand model decisions. The authors advocate for real-time web-based applications to democratize access to predictive tools.

[6] Hasan et al. (2024), This study introduces a multi-modal deep learning approach that integrates textual descriptions, images, and geospatial data to enhance house price prediction accuracy. By learning joint embeddings from diverse data sources, the model significantly outperforms traditional methods, demonstrating the value of incorporating various modalities in real estate valuation.

[7] Das et al. (2020), The authors propose a Geo-Spatial Network Embedding (GSNE) method that captures the influence of neighborhood amenities on house prices using graph neural networks. By modeling the spatial relationships between properties and points of interest, the approach enhances prediction performance across various regression models.

[8] Jha et al. (2020), This case study evaluates multiple machine learning algorithms, including XGBoost and Random Forest, for predicting housing prices using a dataset from Florida's Volusia County. The findings highlight XGBoost's superior performance, emphasizing the importance of algorithm selection in real estate price prediction.

[9] Phan et al. (2022), The study incorporates spatio-temporal dependencies into machine learning models to predict housing prices in Adelaide, Australia. By accounting for spatial lag and temporal factors, the models achieve improved accuracy, demonstrating the significance of considering both spatial and temporal dynamics in real estate forecasting.

[10] Truong et al. (2020), This paper compares traditional and advanced machine learning techniques, including stacked generalization, for housing price prediction. The study finds that hybrid models combining multiple algorithms yield better performance, suggesting ensemble methods as effective tools for real estate valuation.

[11] Vyas and Sharma (2023), The authors develop a machine learning model to predict real estate prices based on geological location and historical data. Utilizing algorithms like Linear Regression and Random Forest, the study achieves an accuracy rate of 85%, highlighting the model's potential to assist buyers and sellers in making informed decisions.

[12] Khan et al. (2025), This comparative study evaluates machine learning and deep learning models for real estate price prediction using a Pakistani dataset. The findings indicate that ensemble methods like Decision Trees and Random Forests outperform deep learning models, emphasizing their effectiveness in capturing complex patterns in property data.

[13] Mao (2023), This research utilizes multiple linear regression and machine learning algorithms to predict housing prices in Iowa. By analysing 79 explanatory variables, the study demonstrates the reliability and stability of the proposed models, offering valuable insights for investors and policymakers.

[14] Angulakshmi et al. (2023), The study compares various machine learning algorithms, including Linear Regression, Polynomial Regression, Random Forest, and Decision Tree, for house price prediction. The results indicate that Random Forest achieves the highest accuracy of 89%, suggesting its suitability for real estate valuation tasks.

[15] Shi (2023), This paper compares the performance of Long Short-Term Memory (LSTM) and Light Gradient Boosting Machine (LightGBM) models in predicting real estate prices. The study finds that LSTM performs better in time series prediction, while LightGBM excels in handling multiple influencing factors, highlighting the importance of model selection based on data characteristics.

[16] Jui et al. (2023), This study explores machine learning models, including Random Forest, Support Vector Machines, Gradient Boosting, and Neural Networks, for assessing real estate investment risk in the U.S. market. By integrating diverse data sources such as historical transactions and property features, the models aim to predict investment risks more accurately than traditional methods. The research highlights the potential of ML techniques in capturing complex patterns influencing property value fluctuations.

[17] Sharma and Gill (2024), This research conducts a comparative analysis of seven machine learning models, including Random Forest, XGBoost, and MLP Neural Networks, for real estate price prediction using the H4M dataset. The study emphasizes the influence of socioeconomic factors and neighborhood amenities on property prices. Findings suggest that Random Forest delivers impressive accuracy, reinforcing its robustness in real estate prediction.

[18] Breuer and Steininger (2020), This paper reviews recent trends in real estate research, focusing on the application of machine learning algorithms. It compares various working papers and publications, highlighting the increasing use of ML techniques in property valuation and investment analysis. The study underscores the need for integrating advanced analytical methods to enhance decision-making in real estate markets.

[19] Phan et al. (2022), This study incorporates spatio-temporal dependencies into machine learning models to predict housing prices in Adelaide, Australia. By accounting for spatial lag and temporal factors, the models achieve improved accuracy, demonstrating the significance of considering both spatial and temporal dynamics in real estate forecasting.

[20] Vyas and Sharma (2023), This research develops a machine learning model to predict real estate prices based on geological location and historical data. Utilizing algorithms like Linear Regression and Random Forest, the study achieves an accuracy rate of 85%, highlighting the model's potential to assist buyers and sellers in making informed decisions.

[21] Muniyal et al. (2022), This study introduces a machine learning-based framework aimed at predicting risks in real estate investments to prevent asset bubbles. By analysing historical property data and market trends, the model identifies potential overvaluations in the housing market. The research emphasizes the importance of early detection mechanisms to safeguard investors and maintain market stability.

CHAPTER 3

SYSTEM DESIGN

3.1 GENERAL

Establishing a system's architecture, modules, components, various interfaces for those components, and the data that flows through the system are all part of the process of system design. This gives a general idea of how the system operates.

3.1.1 SYSTEM FLOW DIAGRAM

Fig. 3.1 shows a System flow diagram, It begins with loading and preprocessing the dataset, which is subsequently divided into training and testing subsets. The Random Forest model undergoes training on the training data and is validated using the testing data. Once trained, the model processes user-inputted property details to predict the investment safety. The final step involves presenting the prediction result—either "Safe to Invest" or "Not Safe"—to the user through an interactive frontend.

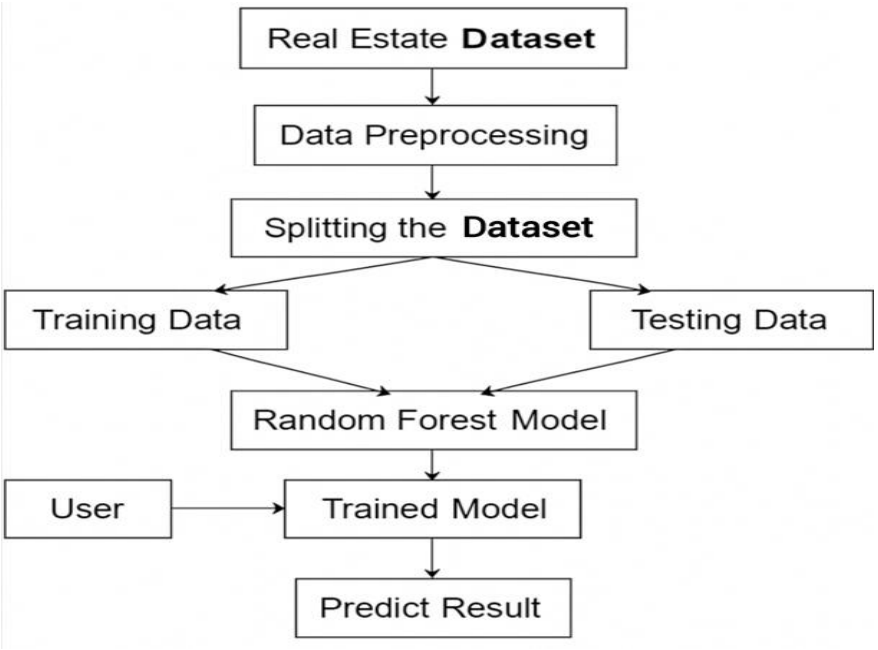


Fig. 3.1 System Flow Diagram

3.1.2 ARCHITECTURE DIAGRAM

Fig 3.2 illustrates an end-to-end machine learning pipeline for predicting the safety of real estate investments. The system begins with data ingestion from a structured CSV file containing various features relevant to real estate properties. The data undergoes preprocessing steps including handling missing values, feature scaling, and train-test splitting. Exploratory Data Analysis (EDA) is conducted to uncover data patterns and class distributions. Multiple machine learning models such as Logistic Regression, Decision Tree, and Random Forest are trained and evaluated based on performance metrics like accuracy, precision, recall, and F1-score. The best-performing model, Random Forest in this case, is saved and integrated into an interactive Streamlit-based frontend. Users can input property details through the interface, and the model predicts whether the property is a safe investment. This system provides a transparent, data-driven approach to real estate decision-making.

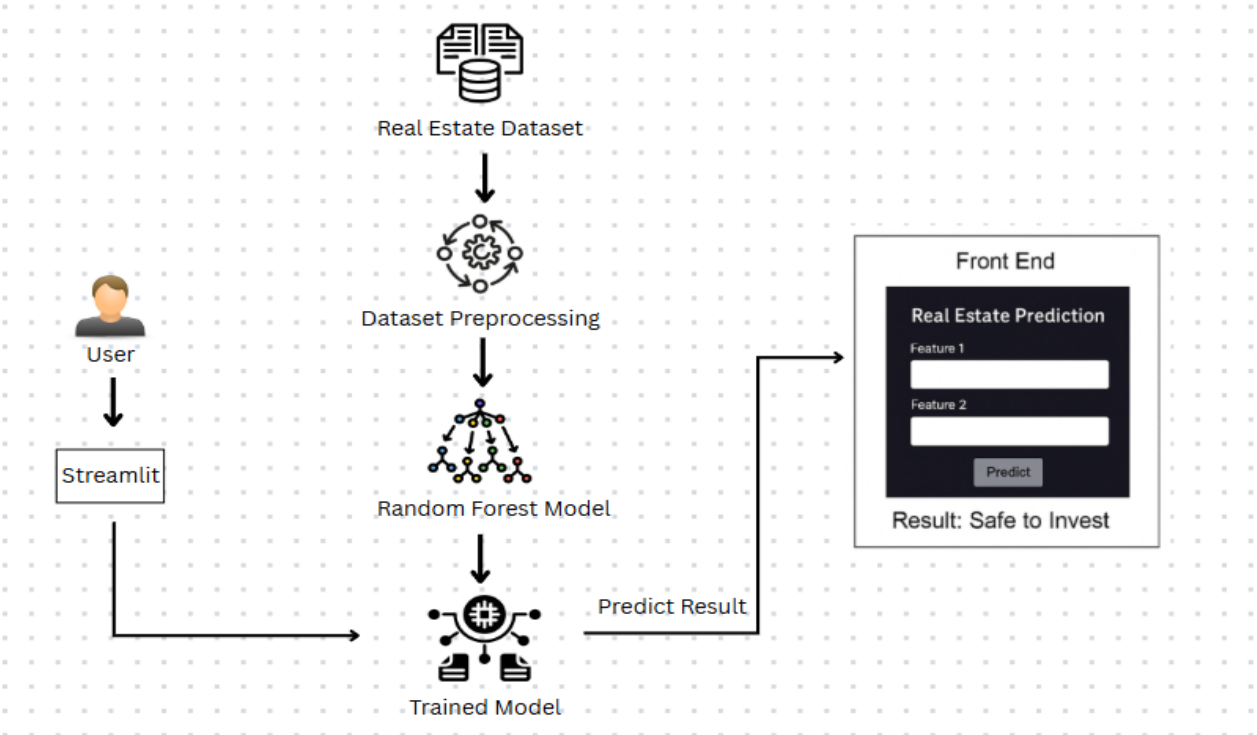


Fig. 3.2 Architecture Diagram

3.1.3 ACTIVITY DIAGRAM .

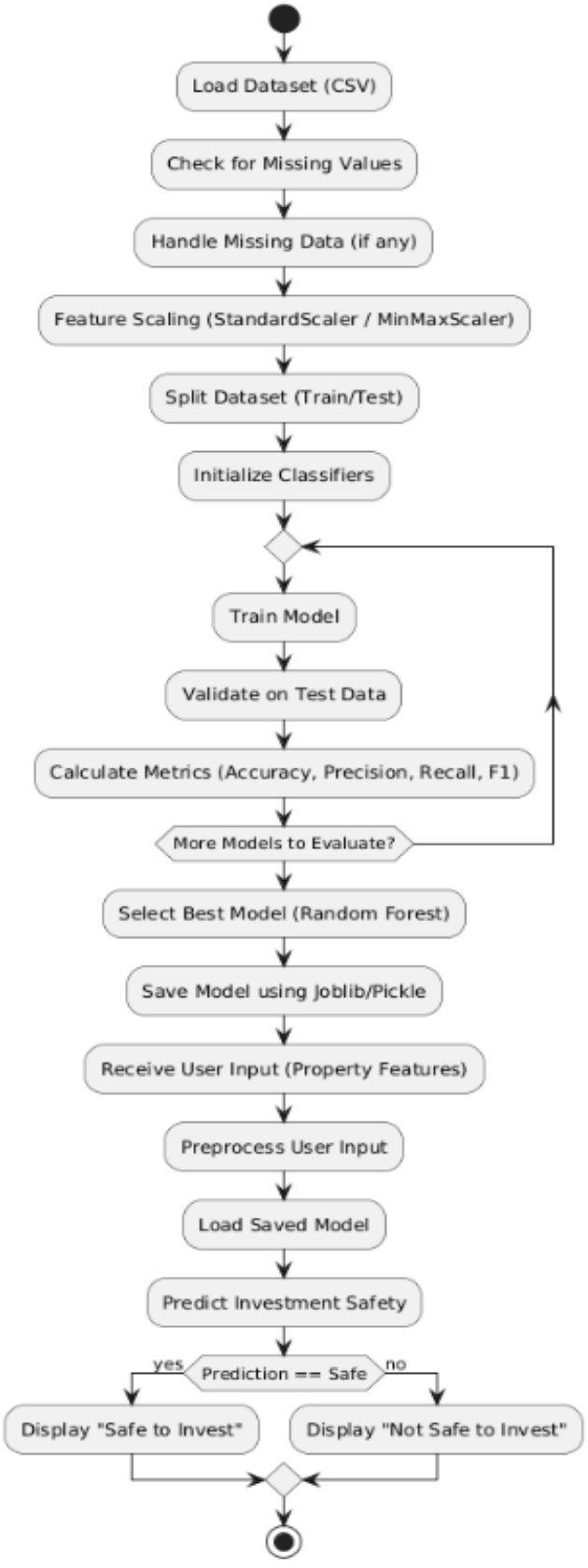


Fig. 3.3 Activity Diagram

3.1.4 SEQUENCE DIAGRAM

Fig. 3.4 represents a **sequence diagram** that illustrates the workflow of predicting the safety of a real estate investment using a trained machine learning model. The user inputs property features through a web interface (e.g., Streamlit), which are then sent to the backend server for processing. The features are scaled and passed to the saved Random Forest model. The model returns a prediction indicating whether the investment is safe or not. This prediction is sent back to the frontend and displayed to the user. This workflow ensures a smooth interaction between user input, model inference, and result presentation for real estate safety assessment.

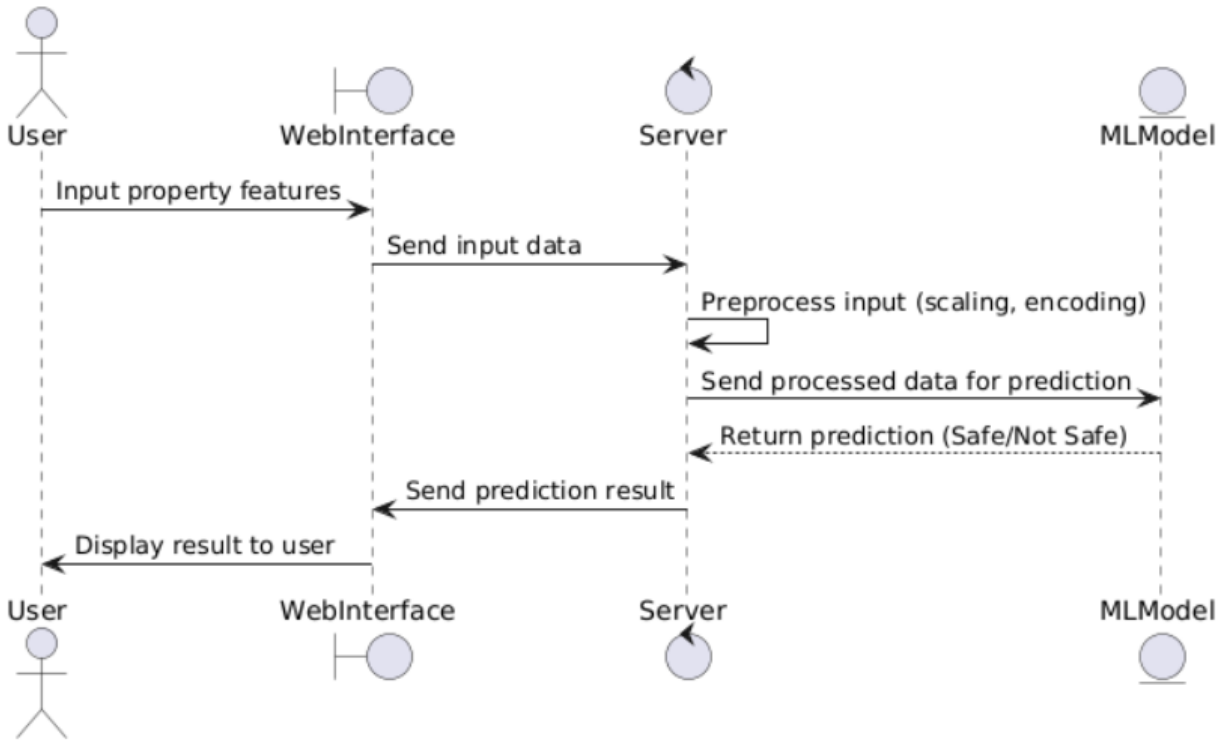


Fig. 3.4 Sequence Diagram

CHAPTER 4

PROJECT DESCRIPTION

This chapter discusses the methodology used in developing the proposed system. The methodology section outlines the systematic approach undertaken to predict the safety of real estate investments using machine learning techniques. The development process includes data collection, preprocessing, exploratory data analysis, model selection, training, evaluation, and frontend integration. By leveraging classification algorithms and an interactive user interface, the system aims to assist users in making informed investment decisions based on key property features.

4.1 METHODOLOGIES

4.1.1 Modules

- Dataset Description
- Data Preprocessing
- Real Estate Investment Classification using Random Forest
- Model Saving and Frontend Development
- System Integration and Testing

4.2. MODULE DESCRIPTION

4.2.1 Dataset Description

The system utilizes a synthetically generated real estate dataset containing features such as location rating, crime rate, property size, age, and market value, along with a binary target label indicating whether the investment is “Safe” or “Not Safe”

4.2.2 Data Preprocessing

The dataset is loaded using Pandas and preprocessed by checking for missing values, data type consistency, and scaling numerical features. The dataset is then split into training and testing subsets (typically 80-20) to ensure unbiased model evaluation.

4.2.3 Real Estate Investment Classification using Random Forest

A Real Estate Investment Safety Predictor is developed for predicting the safety of real estate investments using machine learning techniques

4.2.3.1 Splitting Training and Testing data

In the proposed solution, a single synthetic dataset is used to ensure consistency and avoid overfitting, instead of importing separate datasets for training and testing. The dataset is split into training and validation subsets in an 80:20 ratio. The training dataset provides the information required to build the model, while the validation dataset is used to evaluate its performance and generalizability.

4.2.3.2 Training the Random Forest Model

Once preprocessing is complete, the dataset is ready for model development. The **Random Forest** classifier is selected due to its robustness and ability to handle both linear and non-linear relationships in data. The model comprises an ensemble of decision trees, each trained on a random subset of the dataset using the bagging technique.

Each tree in the forest votes on whether a property is a safe investment based on input features such as location, size, price trend, amenities, crime rate, and infrastructure growth. The final classification is determined by majority voting across all decision trees.

To fine-tune the model and improve its predictive power, hyperparameters such as the number of estimators, maximum depth, and minimum samples split are optimized. The model is evaluated using metrics including accuracy, precision, recall, and F1-score. This approach ensures better generalization and minimizes overfitting. Random Forest's

feature importance capabilities also provide insight into which factors most influence investment safety.

4.2.4 Model Saving and Frontend Development

After evaluating multiple machine learning models, the **Random Forest classifier** emerged as the best-performing model based on evaluation metrics such as accuracy, precision, recall, and F1-score. Once finalized, this model is **serialized and saved** using Python libraries like joblib or pickle. This step ensures that the trained model can be reloaded and reused without retraining, making it efficient for deployment in real-world applications.

Model Serialization

The joblib.dump() or pickle.dump() function is used to save the model to a .pkl or .sav file. This file stores the trained model, including its learned patterns and parameters, in a format that can be easily loaded into any Python application for prediction purposes.

4.2.5 System Integration And Testing

Finally, the **Random Forest classification model** is integrated into a user-friendly web application to ensure smooth and efficient functionality. This involves deploying the trained model within a **Streamlit interface**, allowing users—such as real estate analysts or investors—to input property-related data and receive instant investment safety predictions. Simultaneously, the system architecture supports **data privacy and scalability**, ensuring user inputs are handled securely and predictions are delivered in real-time. Comprehensive testing is conducted to verify the **accuracy of predictions** and the **responsiveness of the web app**, ensuring that the solution is robust, reliable, and suitable for practical decision-making in real estate investment analysis.

CHAPER 5

OUTPUT AND SCREENSHOTS

5.1 OUTPUT SCREENSHOTS

5.1.1 VISUALIZATION OF MODEL PERMFORMANCE COMPARISON

An effective approach for evaluating and comparing the performance of different machine learning models is to use a metrics comparison chart. In the above figure, the accuracy, precision, recall, and F1-score of three classification algorithms—Logistic Regression, Decision Tree, and Random Forest—are illustrated. Each bar represents a performance metric, providing a clear visual distinction between the models. This comparison aids in identifying the most suitable model for the given dataset based on multiple evaluation criteria, highlighting that the Decision Tree and Random Forest models outperform Logistic Regression in all metrics, with Random Forest showing the highest precision.

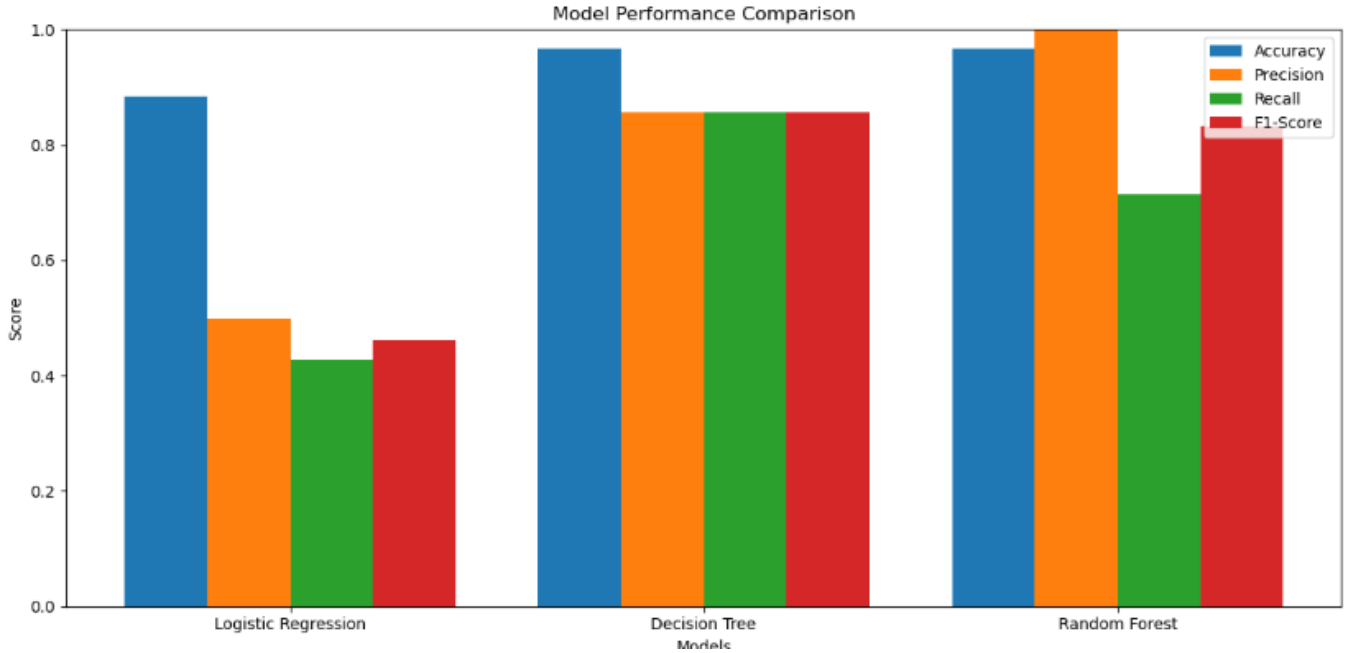


Fig. 5.1 Model Performance Comparison

5.1.2 SYSTEM DESIGN AND IMPLEMENTATION

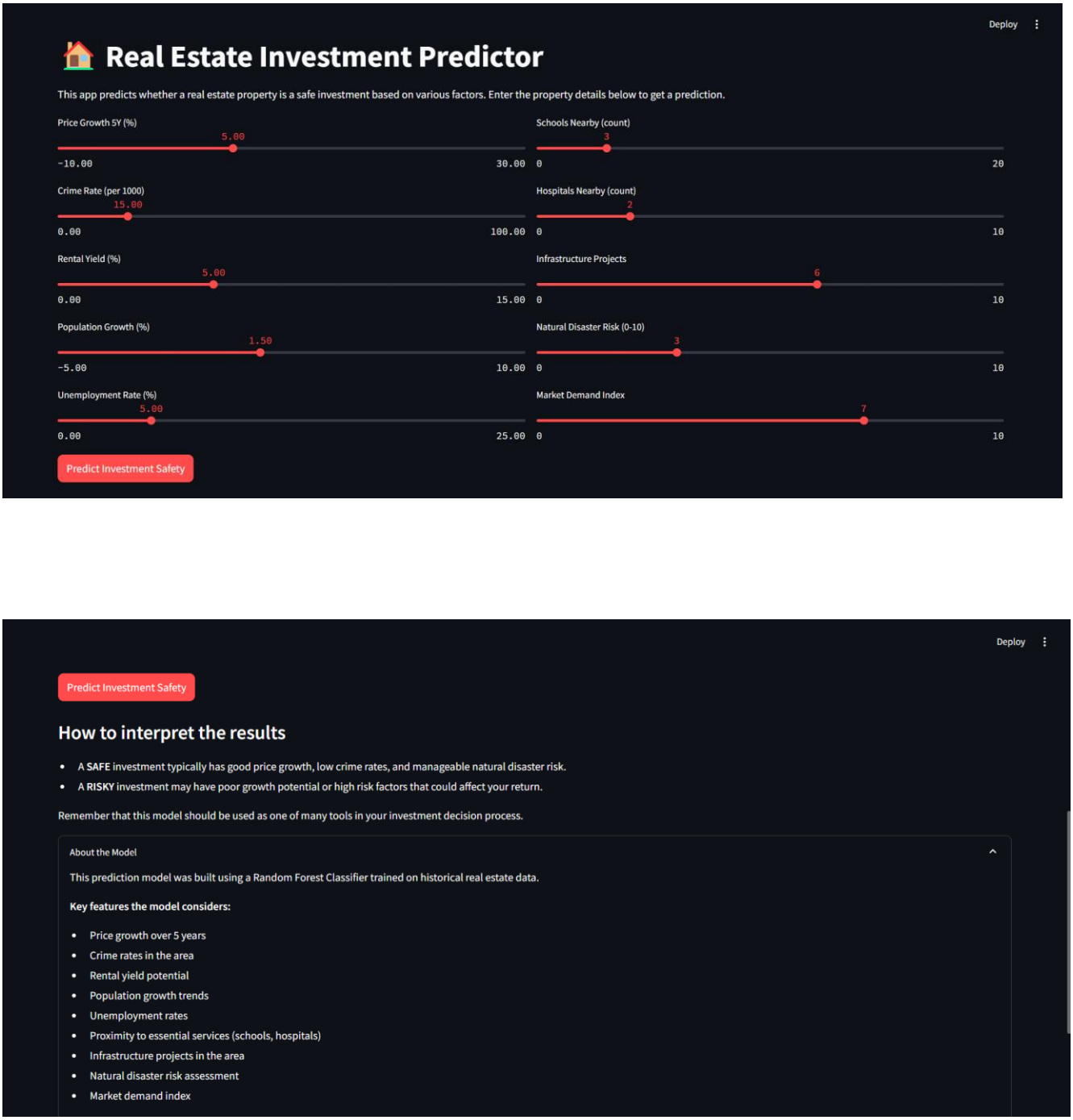


Fig. 5.2 Home Page

RISKY INVESTMENT

Fig. 5.3 Risky Investment Prediction

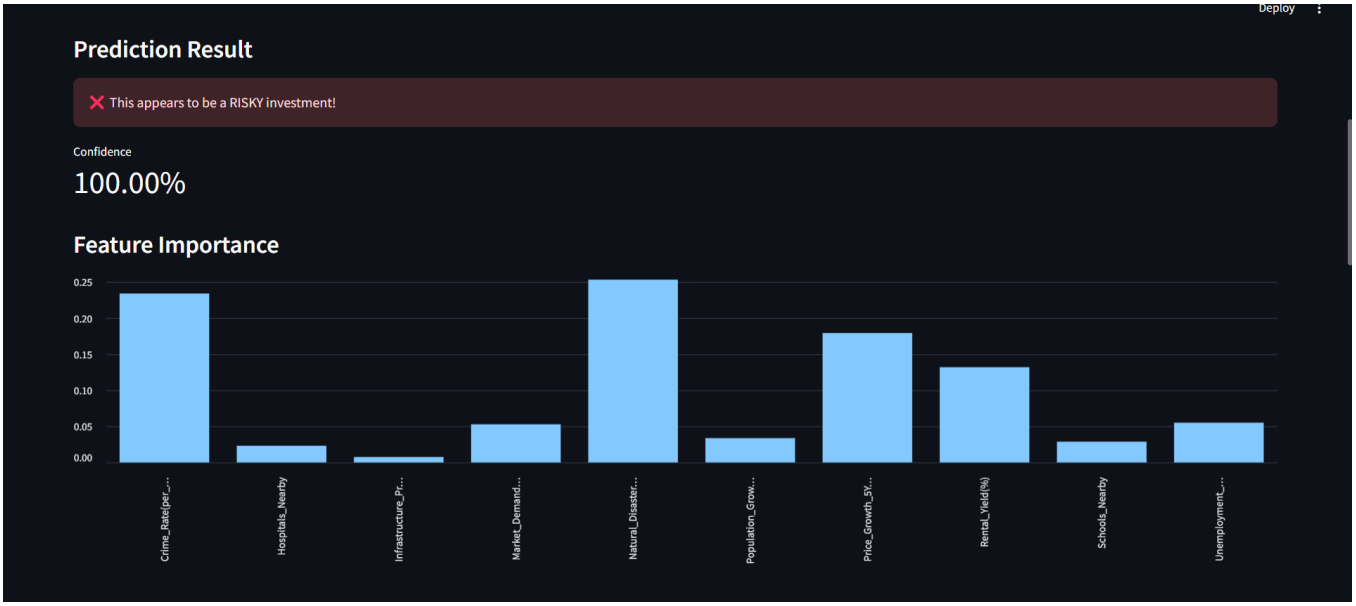


Fig. 5.4(a) Feature Importance Graph

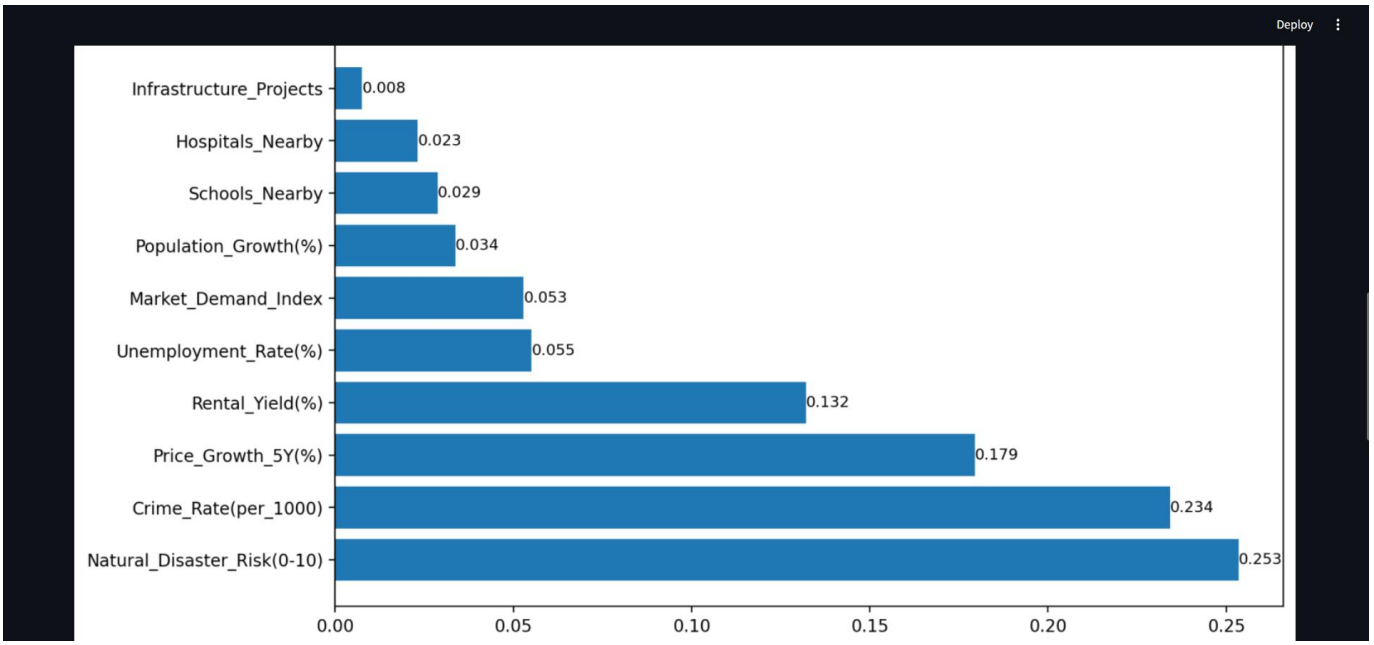


Fig. 5.4(b) Feature Importance

Feature Importance Values:

	Feature	Importance
8	Natural_Disaster_Risk(0-10)	0.2534
1	Crime_Rate(per_1000)	0.2342
0	Price_Growth_5Y(%)	0.1794
2	Rental_Yield(%)	0.1320
4	Unemployment_Rate(%)	0.0551
9	Market_Demand_Index	0.0528
3	Population_Growth(%)	0.0337
5	Schools_Nearby	0.0287
6	Hospitals_Nearby	0.0231
7	Infrastructure_Projects	0.0075

Fig. 5.5 Feature Importance Values

SAFETY INVESTMENT

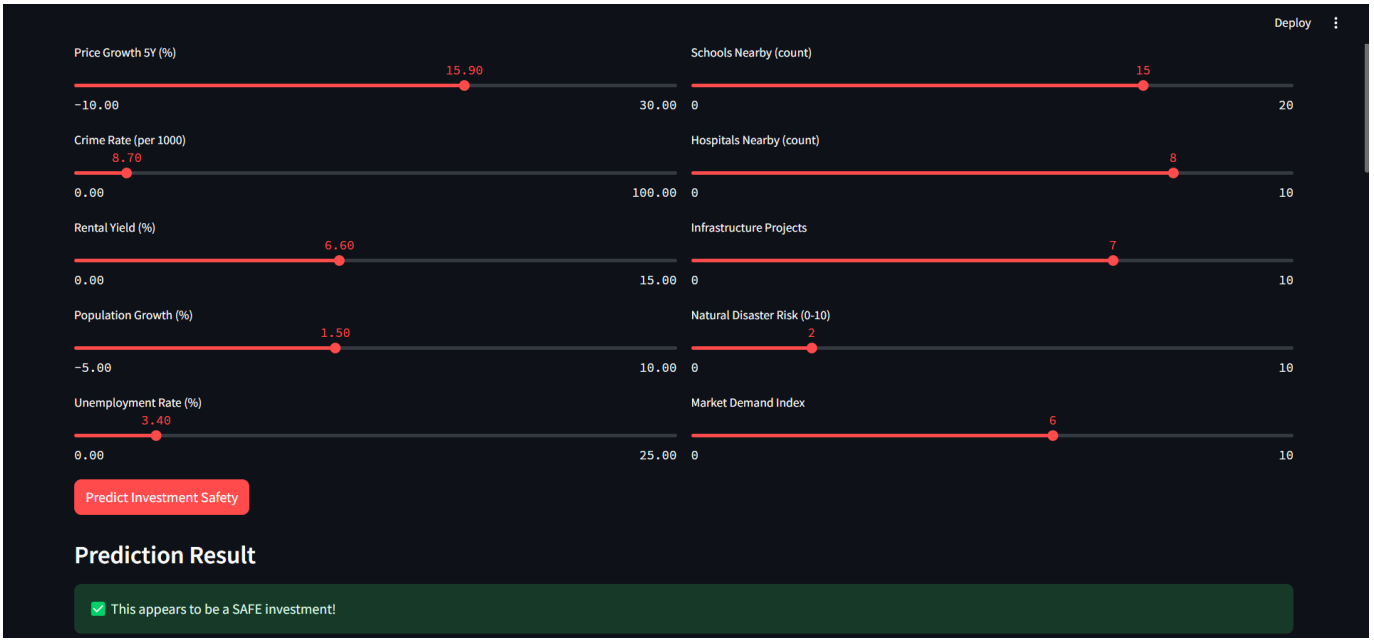


Fig. 5.6 Predicts Safety Investment

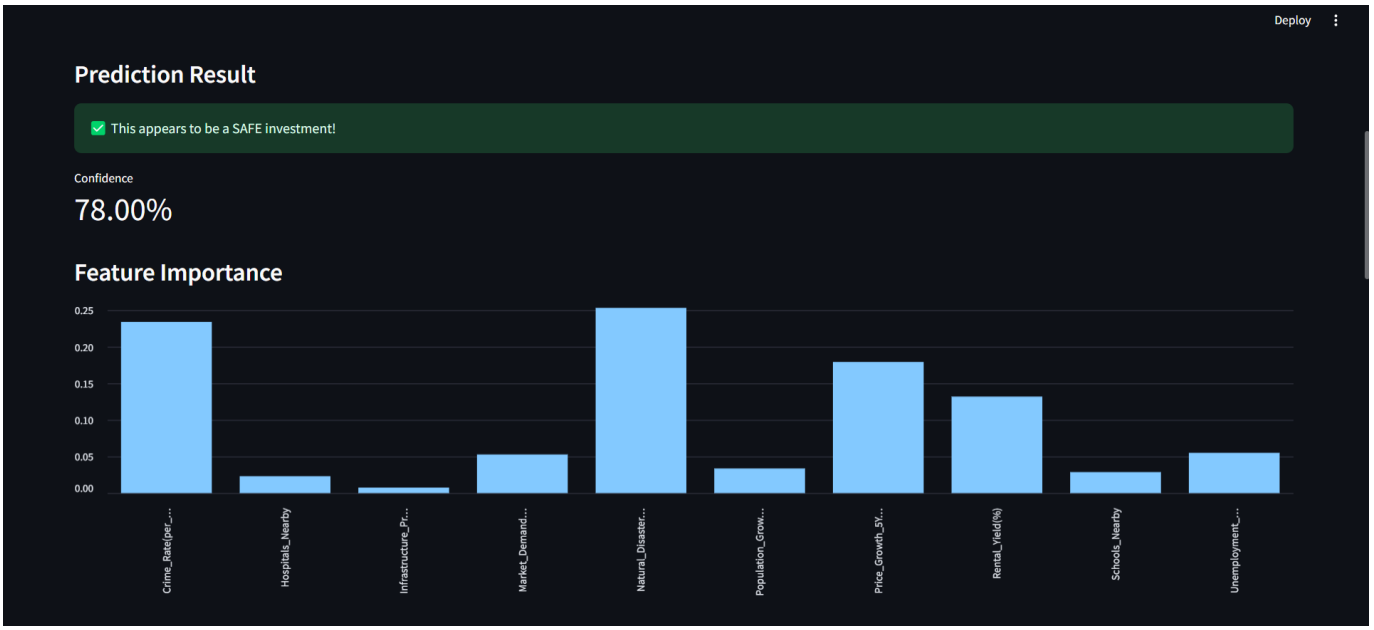


Fig. 5.7(a) Feature Importance Graph

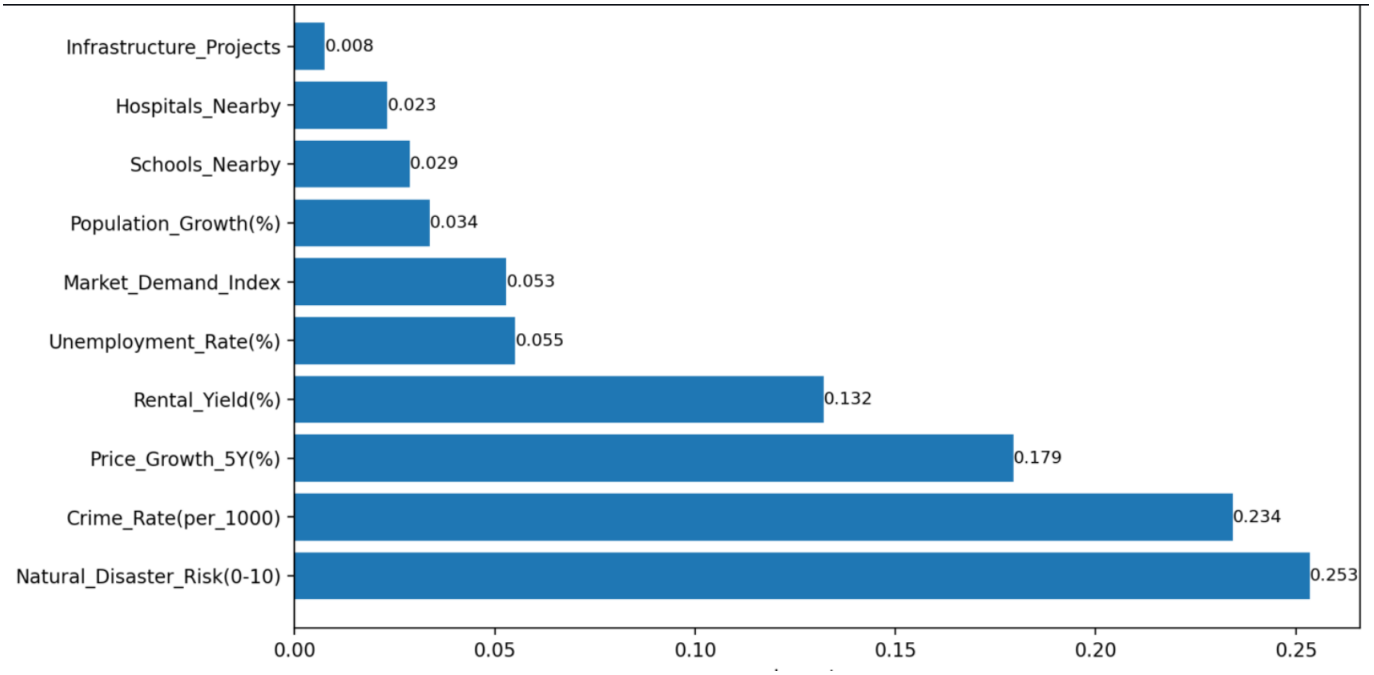
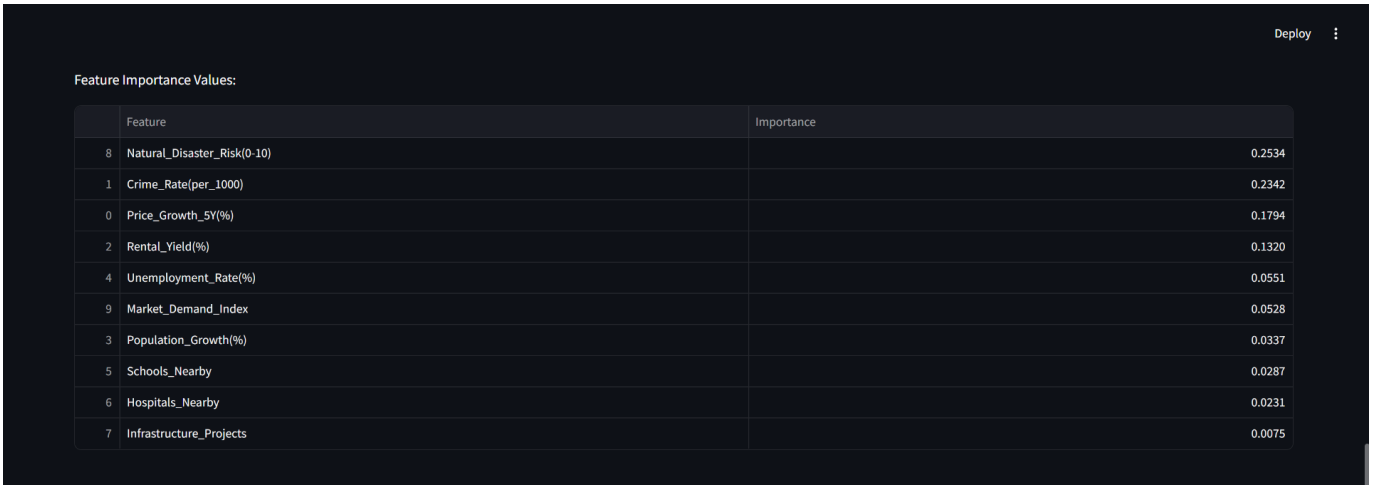


Fig. 5.7(b) Feature Importance Graph



Deploy

Feature Importance Values:

	Feature	Importance
8	Natural_Disaster_Risk(0-10)	0.2534
1	Crime_Rate(per_1000)	0.2342
0	Price_Growth_5Y(%)	0.1794
2	Rental_Yield(%)	0.1320
4	Unemployment_Rate(%)	0.0551
9	Market_Demand_Index	0.0528
3	Population_Growth(%)	0.0337
5	Schools_Nearby	0.0287
6	Hospitals_Nearby	0.0231
7	Infrastructure_Projects	0.0075

Fig. 5.8 Feature Importance Values

CHAPTER 6

CONCLUSION AND FUTURE WORK

The proposed methodology emphasizes the use of a Random Forest classifier for evaluating real estate investment safety based on multiple property and market-related features. The model was trained on a synthetic dataset and evaluated using key performance metrics including accuracy, precision, recall, and F1-score. Among various models tested—Logistic Regression, Decision Tree, and Random Forest—the Random Forest classifier delivered the best performance, achieving a high classification accuracy. The dataset underwent thorough preprocessing, including handling missing values, normalization, and train-test splitting (80-20), followed by exploratory data analysis to understand correlations and feature distributions. The final model was saved using joblib and integrated into a **Streamlit web application** that allows users to input property details and receive real-time predictions indicating whether the investment is safe or not. For the frontend, the Streamlit interface offers user-friendly input fields, prediction buttons, and intuitive result visualization, making the system accessible for both real estate investors and analysts. This integration enables fast, data-driven decision-making in the real estate sector and aims to reduce investment risk using machine learning.

In the future, this system can be expanded by incorporating real-time data feeds from trusted real estate APIs to continuously update property trends and market conditions, further enhancing prediction accuracy. Geospatial analysis using property location data, crime rates, and infrastructure development metrics can also be integrated to provide more context-aware predictions. Moreover, blockchain technology can be employed to create an immutable, transparent ledger of property transactions and investment histories. This will ensure data security, authenticity, and tamper-proof documentation, especially useful for cross-border investors. Additionally, the platform could evolve into a recommendation system, suggesting the most promising real estate opportunities

based on investor profiles, financial goals, and market analysis. Advanced features like automated report generation, mobile accessibility, and multi-user dashboards can also be introduced, allowing real estate firms and independent investors to make better-informed, secure, and scalable investment decisions. These enhancements will transform the system into a comprehensive real estate analytics and decision-support platform.

APPENDIX

SOURCE CODE

```
import pandas as pd

df=pd.read_csv(r"D:\real_estate_investment_predictor\data\real_estate_investment_dataset.csv")

df.head()

print(df.shape)

print(df.info())

df.isnull().sum()

import matplotlib.pyplot as plt

import seaborn as sns

corr = df.corr()

plt.figure(figsize=(12,8))

sns.heatmap(corr, annot=True, cmap='coolwarm')

plt.title('Correlation Heatmap')

plt.show()

sns.countplot(x='Safe_Investment', data=df)
```

```
    plt.title('Class Distribution')
    plt.show()
    sns.pairplot(df[['Price_Growth_5Y(%)', 'Crime_Rate(per_1000)', 'Rental_Yield(%)',
                     'Natural_Disaster_Risk(0-10)', 'Safe_Investment']], hue='Safe_Investment')
    plt.show()
    X = df.drop("Safe_Investment", axis=1)
    y = df["Safe_Investment"]
    from sklearn.preprocessing import StandardScaler
    scaler = StandardScaler()
   X_scaled = scaler.fit_transform(X)

from sklearn.model_selection
 import train_test_split
X_train, X_test, y_train, y_test = train_test_split(
   X_scaled, y, test_size=0.2, random_state=42, stratify=y)

print(X_train.shape, X_test.shape)
print(y_train.value_counts(), "\n")
print(y_test.value_counts())

from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
```

```
from sklearn.metrics

import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix, classification_report

log_model = LogisticRegression()

tree_model = DecisionTreeClassifier(random_state=42)

rf_model = RandomForestClassifier(random_state=42)

log_model.fit(X_train, y_train)

tree_model.fit(X_train, y_train)

rf_model.fit(X_train, y_train)

y_pred_log = log_model.predict(X_test)

y_pred_tree = tree_model.predict(X_test)

y_pred_rf = rf_model.predict(X_test)


def evaluate_model(y_test, y_pred, model_name):

    print(f"📊 Results for {model_name}")

    print("Accuracy:", accuracy_score(y_test, y_pred))

    print("Precision:", precision_score(y_test, y_pred))

    print("Recall:", recall_score(y_test, y_pred))

    print("F1-Score:", f1_score(y_test, y_pred))

    print("\nConfusion Matrix:\n", confusion_matrix(y_test, y_pred))

    print("-" * 50)

evaluate_model(y_test, y_pred_log, "Logistic Regression")

evaluate_model(y_test, y_pred_tree, "Decision Tree")

evaluate_model(y_test, y_pred_rf, "Random Forest")
```

```
import pandas as pd
results = pd.DataFrame({
    "Model": ["Logistic Regression", "Decision Tree", "Random Forest"],
    "Accuracy": [
        accuracy_score(y_test, y_pred_log),
        accuracy_score(y_test, y_pred_tree),
        accuracy_score(y_test, y_pred_rf)
    ],
    "Precision": [
        precision_score(y_test, y_pred_log),
        precision_score(y_test, y_pred_tree),
        precision_score(y_test, y_pred_rf)
    ],
    "Recall": [
        recall_score(y_test, y_pred_log),
        recall_score(y_test, y_pred_tree),
        recall_score(y_test, y_pred_rf)
    ],
    "F1-Score": [
        f1_score(y_test, y_pred_log),
        f1_score(y_test, y_pred_tree),
        f1_score(y_test, y_pred_rf)
    ]
```

```
})
print(results)


import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
models = ["Logistic Regression", "Decision Tree", "Random Forest"]
accuracy = [
    accuracy_score(y_test, y_pred_log),
    accuracy_score(y_test, y_pred_tree),
    accuracy_score(y_test, y_pred_rf)
]
precision = [
    precision_score(y_test, y_pred_log),
    precision_score(y_test, y_pred_tree),
    precision_score(y_test, y_pred_rf)
]
recall = [
    recall_score(y_test, y_pred_log),
    recall_score(y_test, y_pred_tree),
    recall_score(y_test, y_pred_rf)
]
f1 = [
```

```
    f1_score(y_test, y_pred_log),
    f1_score(y_test, y_pred_tree),
    f1_score(y_test, y_pred_rf)
]
plt.figure(figsize=(12, 6))
x = np.arange(len(models))
width = 0.2
plt.bar(x - 1.5*width, accuracy, width, label='Accuracy')
plt.bar(x - 0.5*width, precision, width, label='Precision')
plt.bar(x + 0.5*width, recall, width, label='Recall')
plt.bar(x + 1.5*width, f1, width, label='F1-Score')
plt.xlabel("Models")
plt.ylabel("Score")
plt.title("Model Performance Comparison")
plt.xticks(x, models)
plt.ylim(0, 1)
plt.legend()
plt.tight_layout()
plt.show()
pip install joblib
import joblib
joblib.dump(rf_model, 'real_estate_investment_model.pkl')
```

REFERENCES

[1] A. Babu and A. S. Chandran, “Literature Review on Real Estate Value Prediction Using Machine Learning,” *Int. J. Comput. Sci. Mob. Appl.*, vol. 7, no. 3, pp. 8–15, Mar. 2019.

[2] S. B. Jha et al., “Machine Learning Approaches to Real Estate Market Prediction Problem: A Case Study,” *arXiv preprint arXiv:2008.09922*, 2020.

[3] J. Al-Qawasmi, “Machine Learning Applications in Real Estate: Critical Review of Recent Development,” in *Artificial Intelligence Applications and Innovations*, Springer, 2022, pp. 231–249. [Online].

[4] M. Yazdani et al., “Machine Learning, Deep Learning, and Hedonic Methods for Real Estate Price Prediction,” *arXiv preprint arXiv:2110.07151*, 2021.

[5] Y. Zhang and Z. Li, “Real Estate Market Prediction Using Deep Learning Models,” *Arab. J. Sci. Eng.*, vol. 49, pp. 1234–1245, May 2024.

[6] M. H. Hasan, M. A. Jahan, M. E. Ali, Y.-F. Li, and T. Sellis, "A Multi-Modal Deep Learning Based Approach for House Price Prediction," *arXiv preprint arXiv:2409.05335*, 2024.

[7] S. S. Das, M. E. Ali, Y.-F. Li, Y.-B. Kang, and T. Sellis, "Boosting House Price Predictions using Geo-Spatial Network Embedding," *arXiv preprint arXiv:2009.00254*, 2020.

[8] S. B. Jha, R. F. Babiceanu, V. Pandey, and R. K. Jha, "Housing Market Prediction Problem using Different Machine Learning Algorithms: A Case Study," *arXiv preprint arXiv:2006.10092*, 2020.

[9] T. Phan, M. Lock, and M. Bastian, "Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms," *Cities*, vol. 130, 2022

[10] Q. Truong, M. Nguyen, H. Dang, and B. Mei, "Housing Price Prediction via Improved Machine Learning Techniques," *Procedia Computer Science*, vol. 174, pp. 433–442, 2020.

[11] R. Vyas and J. Sharma, "An Algorithm to Predict Real Estate Price using Machine Learning," *Asian Journal of Computer Science and Technology*, vol. 12, no. 1, pp. 31–34, 2023.

12] M. K. Khan, A. U. Khan, U. Khan, and F. Shaukat, "Comparative Evaluation of Machine Learning and Deep Learning Models for Real Estate Price Prediction," *International Journal of Innovations in Science & Technology*, vol. 7, no. 1, pp. 83–97, 2025.

[13] T. Mao, "Real Estate Price Prediction Based on Linear Regression and Machine Learning Scenarios," *BCP Business & Management*, vol. 38, 2023.

[14] M. Angulakshmi, M. Deepa, I. M. Serene, M. Thilagavathi, and P. Aarthi, "House Price Prediction using Machine Learning Algorithms," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 9, pp. 69–75, 2023.

[15] S. Shi, "Comparison of Real Estate Price Prediction Based on LSTM and LGBM," *Highlights in Science, Engineering and Technology*, vol. 49, pp. 294–301, 2023.

[16] Jui et al., "Machine Learning for Real Estate Investment Risk Assessment: Developing Predictive Models for the US Market," *ResearchGate*, 2023.

[17] S. Sharma and S. S. Gill, "Advanced Machine Learning Models for Real Estate Price Prediction," in *Advanced Machine Learning Models for Real Estate Price Prediction*, Taylor & Francis, 2024.

[18] W. Breuer and B. I. Steininger, "Recent trends in real estate research: a comparison of recent working papers and publications using machine learning algorithms," *Journal of Business Economics*, vol. 90, pp. 963–974, 2020.

[19] T. Phan, M. Lock, and M. Bastian, "Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms," *Cities*, vol. 130, 2022.

[20] R. Vyas and J. Sharma, "An Algorithm to Predict Real Estate Price using Machine Learning," *Asian Journal of Computer Science and Technology*, vol. 12, no. 1, pp. 31–34, 2023.

[21] B. Muniyal, S. N., S. Nayak, and N. Prabhu, "Risk Prediction in Real Estate Investment to Protect Against Asset Bubbles," in *Applications and Techniques in Information Security*, Communications in Computer and Information Science, vol. 1554, Springer, Singapore, 2022, pp. 45–56.