House Price Prediction in Dhaka Using Machine Learning Algorithm

BY

Jannatul Ferdous
ID: 203-15-3902

Tania Akter
ID: 203-15-3917

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Mr. Saiful Islam
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised By

Mr. Shah Md Tanvir Siddiquee
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

July 2024

APPROVAL

This Project titled "House Price Prediction in Dhaka Using Machine Learning Algorithm", submitted by **Jannatul Ferdous** and **Tania Akter** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 14-07-2024.

BOARD OF EXAMINERS

Dr. Sheak Rashed Haider Noori (SRH) — **Board Chairman**
Professor & Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Most. Hasna Hena (HH) — **Internal Examiner 1**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Md. Ferdouse Ahmed Foysal (FAF) — **Internal Examiner 2**
Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Arshad Ali (DAA) — **External Examiner**
Professor
Department of Computer Science and Engineering
Hajee Mohammad Danesh Science and Technology University

©Daffodil International University, Bangladesh i

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Mr. Saiful Islam, Assistant Professor, Department of Computer Science and Engineering, Daffodil International University**. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Mr. Saiful Islam
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Mr. Shah Md Tanvir Siddiquee
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

Jannatul

Jannatul Ferdous
ID: 203-15-3902
Department of CSE
Daffodil International University

Tania

Tania Akter
ID: 203-15-3917
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express our heartiest thanks and gratefulness to almighty for His divine blessing making it possible for us to complete the final year project successfully.

I am grateful and wish our profound indebtedness **Mr. Saiful Islam, Assistant Professor,** Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Machine learning*" to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

I would like to express our heartiest gratitude to the **Head of the Department of CSE**, for his kind help in finishing our project and also to other faculty members and the staff of the Department of CSE, Daffodil International University.

I would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, I must acknowledge with due respect the constant support and patience of my parents.

ABSTRACT

This study investigates the application of various machine learning models for predicting house prices using a dataset sourced from Bproperty, which includes 598 entries and seven attributes: Address, no. of bedrooms, no. of bathrooms, area sqft, type, features, and price. The methodology encompasses several key steps: It includes data selection, data cleaning where missing value removal is done, feature selection, encoding, exploratory data analysis EDA, model training, model evaluation and model testing. The analysis compares the performance of four regression models: Polynomial Regressor is one among them, SVR (Support Vector Regression), XGB Regression, and Random Forest Regression are the other models. Based on these models the R^2 score was calculated to be 27. 03% for Polynomial Regression: 40. 14% for SVR; 97. 16% for XGB Regression, and 97. 73% for Random Forest Regression. These outcomes suggest that the techniques like XGB and the random forest outcompete other regression models for efficiency of prediction. This probably explains why Random Forest and XGB are able to make much better predictions of the data since they have the capacity to model interactions that are non-linear in nature. These results reveal an important need for the usage of modern techniques in machine learning for reasonable and accurate house price estimate important for the stakeholders involved in the real estate market.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Overview

Quoting house prices is important in the real estate market because it provides valuable insight to potential buyers, sellers, and investors. Forecasts help decision makers in investment selection and market analysis. This paper aims to apply machine learning to estimate home prices based on factors such as location, number of bedrooms, number of bathrooms, square feet of home area, type of property, special amenities, and asking price The data consisting of 598 records and 7 attributes was obtained from Bproperty by web scraping using Data Miner. The method involves several steps: data preprocessing, data cleaning, dimension reduction, data transformation, exploratory data analysis (EDA), data training, model evaluation, and validation [1].

Four machine learning models are used in this paper: they start with polynomial regression, followed by Support Vector Regression (SVR), then XGBoost Regression, and finally Random Forest Regression with performance metrics such as mean absolute error (MAE) and R-squared (R2) are used) to evaluate each model Turns out By combining these methods, the research aims to develop a more accurate and practical model for predicting real estate prices , which is essential for the efficiency of real estate inspection has increased [2].

1.2 Background and Present State

Real estate market has always been considered as the significant sector of the world economy and the price on houses depends on the certain factors such as the state of economy, geographic location, characteristics of the house, and the market tendencies. Previously, the estimation of house prices involved rule of thumb and the analysts' opinions which are vital but generic and far remove from the accuracy of the contemporary technology-related computation.

Real estate data analysis has changed dramatically with the development of machine learning techniques and the use of predictive models Machine learning based predictive models can handle different data, reveal patterns, and make accurate predictions based on attributes on. This shift from traditional bookkeeping solutions to empirical estimation methods improved the accuracy of house price forecasts. For the purpose of predictive analytics, several regression models are in use; these include Polynomial Regression, SVR, XGBoost Regression, and Random Forest Regression. These are ensemble learning models, non-linear, gradient boosting capable of making accurate and robust prediction results and are very important tools in real estate industry [3].

1.3 Problem Statement

House price forecasting presents a difficult task because of factors that affect the real estate market, which is very much large. Conventional approaches are not very effective in modeling the complex and usually nonlinear interdependencies between these elements. This renders a more complex approach to develop the understanding of the interactions between the parameters that include the location, number of bedrooms and bathrooms, size of the property, type and other facilities offered [4].

The main research area of this research is how can we develop a reliable machine learning algorithm that accurately predicts house prices? We have a dataset containing property listings crawled from the Bproperty website; the goal is to preprocess the information, choose suitable features and build diverse regression models. Thus, Polynomial Regression, SVR, XGBoost Regression, and Random Forest Regression models will be assessed to identify the best model. This approach aims at fulfilling the quest of rendering accurate and utilizable information for pulling off viable decisions to the various stakeholders in the real estate business arena.

1.4 Objective

The main objective of this study is to build a flexible multi-factorial machine learning algorithm for estimating house prices using features collected from Bproperty. The objective is to re-employ stated advanced tools of regression to identify critical elements that affect house prices and therefore improve the prediction of price changes. Specific objectives include: data acquisition and preprocessing to deal with missing values and to convert the categorical variables to numerical form; performing EDA in order to examine data for specific features, patterns, relations and so on; feature selection to identify the most relevant factors that affect the house prices. Various regression models including polynomial regression, support vector regression (SVR), XGBoost regression, random forest regression will be used and evaluated in this study Performance comparison will be based on metrics like mean absolute error (MAE) and statistical testing R -square so. By identifying the most effective model, the study aims to provide a reliable tool to help real estate stakeholders make informed decisions. In conclusion, the project has goals of sharing potential and confirming that machine learning is effective in the real estate market forecasting and property valuations.

1.5 Scope and Limitations

Scope:

The main objective of this study is to predict house prices in a particular area using data collected from Bproperty Ltd. The analysis included data preprocessing using advanced regression algorithms including Polynomial Regression, SVR, feature selection, model training and model evaluation , XGBoost regression, and random forest regression. The aim of the study is to develop a reliable and accurate predictive model that provides improved decision making for stakeholders in the real estate industry.

Limitations:

Data Quality: It has been postulated that the specifics of predictions predominantly depend on the quality and the volume of the amassed data set. Problem can arise such as missing variables, inconsistency in web scraped data.

Model Complexity: Despite the higher accuracy of more sophisticated machine learning models, such type of models is rather heavy and contains a complex structure. This may hamper their ability to be utilized in big data or real-time result prediction.

Generalization: The models based on the data of the past might not be capable to interpret the changes in the market forces or other circumstances that affect the house prices at the present or in the future.

Interpretability: While some sophisticated models like XGBoost, and Random Forest offer excellent precision and sometimes can out predict some simpler models like Linear regression, the models have so little interpretability compared to the simpler models.

1.6 Report Organization

Chapter 1: Introduction.

Introduces the study aiming at presenting the goals, background, problem, and the importance of utilizing machine learning approaches to estimate house prices.

Chapter 2: Literature Review

Provides an overview of prior studies and techniques in forecasting house prices' future trends, including the advantages and disadvantages of each approach.

Chapter 3: Methodology/Requirement Analysis & Design Specification

Describes how the data was collected and cleaned, candidate features for use and specifications taken into consideration in the design of the chosen machine learning models.

Chapter 4: Implementation

Discusses more details on how to implement the regression models, or the tools and languages that need to be used on this process.

Chapter 5: Result and Analysis

Summaries the results from the model evaluations and analyses made according to the key performance indicators introduced and reflected on the significance of these findings.

Chapter 6: Impact on Society, Environment, and Sustainability

Discusses various advantages and limitations associated with applying machine learning to the prediction of house prices and answers a range of questions that may be crucial in the systems' societal, environmental, and sustainability contexts.

Chapter 7: Conclusion and Future Work

It summarizes the main results of the study, draws conclusions from the study, and, suggests directions for future investigation and improvement in the field.

1.7 Summary

This is where the stage is set where prediction of house prices is important in the field of real estate. It focuses on the aspect of machine learning as a way of enhancing predictive capability and decision making of the stakeholders. The data set is also collected through web scraping from Bproperty website and it consists of the attributes namely, locational aspect, number of bedrooms and bathrooms, size in sqft, type of property, features and its price. Thus, the goals of the study are as follows: Data cleaning, data feature selection, and the process of training multiple regression models (Polynomial Regression, SVR, XGBoost Regression, Random Forest Regression). In particular, the occurrence of such advanced techniques will allow making the study closely beneficial to contributing to the construction of a reliable, accurate, and sophisticated forecast model that will improve the knowledge and decision-making in new real estate purchases or investment opportunities.

CHAPTER 2
LITERATURE REVIEW

2.1 Overview

Literature review discusses prior studies regarding the forecast of house prices with the help of machine learning methods. It explores research works that use regression models such as Polynomial Regression, SVR, XGBoost, and Random Forest to model the housing market. Prominent findings are the significance of the geographical location, the characteristics of the building and structure, the economic parameters, and the trends in the housing market. Feature selection is one aspect relevant to the analysis of research papers, data preprocessing techniques and Model evaluation measures of Mean Absolute Error and R Squared. Also, the review presents improvements in scraping techniques and collecting large volumes of data to form the basis for training and building the models. Thus, the literature analysis allows to outline current practices, difficulties and potential of using machine learning to predict real estate prices and lays the groundwork for the current research. This feeds into the method used in this study to establish a well-ordered technique for establishing an accurate house price prediction model.

2.2 Related Works

Q. Truong, M. Nguyen et al. [5] This paper focuses on overcoming the shortcomings of traditional machine learning methods for housing price forecasting, which often overlook model performance and complex models. By incorporating factors like location, area, and population alongside the House Price Index (HPI), it aims to provide a more accurate prediction of individual housing prices. The study explores both traditional and advanced machine learning methods to compare their effectiveness. Comprehensive validation of various regression techniques is conducted to assess their impact on prediction accuracy. The findings offer an optimistic outlook on enhancing housing price predictions using advanced machine learning models.

Satish, G. Naga, et al. [6] This paper investigates the prediction of future housing prices using machine learning algorithms. It explores different forecasting methods and chooses Lasso regression because of the simplest selection method available. The findings show that their method is effective and produces predictions similar to other models in predicting housing costs. In addition, the study uses machine learning algorithms with XGBoost, Lasso regression, neural networks to develop housing price forecasting models and evaluate their accuracy The study concludes that Lasso regression algorithms consistently outperform other models , the gradient increments reach 91% accuracy rate.

M. Thamarai et al. [7] This study seeks to aid customers in locating homes that align with their requirements by utilizing attributes such as bedroom count, age of the house, proximity to transportation, schools, and shopping centers. The proposed model emphasizes predicting house availability and estimating prices in a small town within the West Godavari district of Andhra Pradesh. The approach employs decision tree classification, decision tree regression, and multiple linear regression techniques implemented through the Scikit-Learn Machine Learning Tool. This research illustrates the effective application of machine learning in facilitating personalized home purchasing and price forecasting.

N. H. Zulkifley, et al. [8] The literature review examines the important properties and effective models for predicting house prices. Studies confirm that artificial neural networks, support vector regression, and XGBoost are the most successful models for this task. It highlights the important impact of location and architectural features on predicting house prices. This study provides valuable insights for real estate developers and researchers, helping them identify important characteristics and choose the best machine learning algorithms for home price analysis. The findings of this study provide guidance for future research directions and practical applications in the real estate market.

A. B. Adetunji et al. [9] The house price index (HPI) is commonly used to measure changes in house prices but lacks accuracy in predicting specific house prices because it relies on price changes from repetitive transactions. Using a dataset of Boston buildings from the UCI Machine Learning Repository that includes 506 cases and 14 features, the study examined model performance Comparison between predicted and actual values shows for example achieved a satisfactory accuracy within ±5% . This highlights the effectiveness of random forests in accurately predicting house prices for individual cases.

M. Modi, A. Sharma et al. [10] Investors rely on understanding market trends to make informed decisions and optimize business results. The real estate industry often fails to see the obvious, and sometimes sellers are misled by contractors' market values. This study addresses these issues by using a more accurate model using a complete dataset, overcoming previous challenges of low accuracy and overfitting of existing models The new model , equipped with user-friendly interfaces, aims to save time and money and help businesses and individuals . It uses a variety of machine-learning algorithms, including new tree, support vector machine, K nearest neighbor, naïve base, logistic regression, and stochastic gradient descent, combined with stacking techniques using voting classifier appears as the most accurate, with a high accuracy rate of 98.27%.

S. Dabreo, et al. [11] This study examines the use of machine learning to predict housing prices using two Kagle datasets: one by Harrison Rubinfeld of Boston, and the other by Anthony Pino of Melbourne, building on a few studies on forecasting house prices in India.

The research examines machine learning algorithms on these datasets and aims to develop useful predictive tools for users. The results show that the accuracy is significantly affected by the choice of algorithm, and the quality of the data set is important for reliable predictions. Among the tested algorithms, XGBoost showed the highest accuracy, achieving 79%.

S. Mysore, et al. [12] This project aims to predict house prices using various regression models, emphasizing the importance of machine learning (ML) in modern technology and its role in automating industries. The study focuses on ML techniques to enhance the accuracy of house price predictions, addressing the issue of financial losses due to inaccurate predictions in the current system. The project considers numerous factors, including budget constraints and potential house modifications, to provide more accurate predictions. Various ML techniques, such as SVR, KNN, SGB regression, CatBoost regression, and Random Forest regression, were used. Among these, KNN achieved the highest accuracy at 88%.

S. Jamil, T. Mohd et al. [13] To develop an effective Green Building (GB) price predictor model, extensive empirical experiments are necessary to identify the best parameter configurations, techniques, and algorithms. GB aims to enhance building performance, evaluated in Malaysia using five criteria: Energy Efficiency, Indoor Environment Quality, Sustainable Site Planning & Management, Material & Resources, and Water Efficiency. This study models building price prediction using a GB dataset from Kuala Lumpur, Malaysia. Five common algorithms—Linear Regression, Decision Tree, Random Forest, Ridge, and Lasso—were tested on the dataset. The results showed that the Decision Tree Regressor outperformed the others, achieving the highest accuracy with an R^2 value of 94%.

T. Mohd, S. Jamil et al. [14] Time and expertise are needed for modifying models to address specific issues such as green buildings and homes. Green buildings are known to improve green buildings, but less machine learning models are required to predict green building costs than green buildings This paper presents an empirical study that incorporates factors such as buildings which is green and provides an example of a housing price forecast. Five machine learning algorithms—linear regression, decision tree, random forest, ridge, and lasso—were tested on data sets from Kuala Lumpur, Malaysia The results showed that the random forest algorithm outperformed the others, and that the green design factors added significantly improved model accuracy.

L. Yu, et al. [15] This paper introduces a deep learning-based model to improve the accuracy of housing price forecasting and growth analysis using existing real estate data Two sets of models are developed that consider different factors affecting housing prices . The first part includes a comparison of models using multiple characteristics: convolutional neural network (CNN), long short-term memory (LSTM) neural network, and logistic regression models and a closing section so the two focus on time series models, LSTM-1 model, the standard LSTM model, Auto-Regressive and Moving Implementation and comparison of the Average (ARMA) model A comprehensive analysis of used housing prices was conducted in Beijing, including data collection, price estimation and results comparison, which greatly improves actual property price analysis and forecasting To demonstrate the model.

R. Manjula, et al. [16] This paper emphasises the application of machine learning to optimise and accurately predict housing prices. It presents key features essential for accurate price prediction and discusses regression models that utilise these features to minimise Residual Sum of Squares error. Feature engineering is highlighted as crucial for improving prediction accuracy, often employing multiple regressions or polynomial regression. To mitigate overfitting, ridge regression is utilized. The paper ultimately guides the best practices for applying regression models and other techniques to optimise housing price predictions.

A. Chouthai, et al. [17] This paper focuses on predicting house deal prices using various machine learning algorithms. Key factors influencing housing prices include property area, location, construction materials, property age, and the number of bedrooms and garages. Machine learning algorithms such as logistic regression, support vector regression, Lasso regression, and decision tree are utilised to build the predictive models. The study uses housing data from 100 properties to develop these models. The results demonstrate the effectiveness of these machine learning techniques in accurately predicting house prices.

L. B. Gokalani, et al. [18] Pakistan's real estate market significantly impacts GDP growth and offers lucrative investment opportunities. With the ever-increasing demand for housing and frequently changing house prices, there is a need for a forecasting system for future house sales prices. Various factors influencing house sales prices include location, physical attributes, number of bedrooms, and several economic factors. This research focuses on real-time data from DHA Defence Karachi, using regression algorithms like Decision Tree, Random Forest, and Linear Regression to predict house sales prices and compare their performance. The Random Forest algorithm achieves the highest accuracy of 98.08%.

Y. Piao, A. Chen et al. [19] The factors influencing residential real estate prices are complex and the selection of effective features is often unclear, leading to lower accuracy in traditional housing price prediction methods. To address this, a novel Convolutional Neural Network (CNN) model is proposed for both housing price prediction and feature selection. This model demonstrates superior performance compared to traditional methods. Experiments using actual property transaction data confirm that the CNN model achieves better results. The CNN model attains the highest accuracy of 98%.

F. Mostofi et al. [20] This research employs a three-step methodology on a high-dimensional, positively skewed real estate price dataset. Firstly, it uses square root transformation (SRT), cube root transformation (CRT), and logarithmic transformation (LT) to normalise data distribution. Secondly, principal component analysis (PCA) is applied to reduce the dimensionality of the original, SRT, CRT, and LT datasets. Thirdly, the study compares the price prediction accuracy of a PCA-DNN model across datasets with varying skewness levels by analysing their error values. Findings indicate that the CRT method significantly enhances both prediction accuracy and computational efficiency of the PCA-DNN model, while also demonstrating robust generalization. Conversely, SRT and LT methods result in high error rates and overfitting, respectively.

N. Sinha, A. K. Singh, et al. [21] Home buyers often struggle to identify the right property at the right price in the right market, often making decisions based on short-term desires rather than long-term needs. Property prices fluctuate constantly, complicating accurate price predictions and leading to inconsistent pricing from manual calculations by property dealers. This situation creates dissatisfaction among buyers and sellers and fosters unhealthy competition. To address this, a machine learning (ML) model is needed to predict property prices accurately using current data and to analyze buyer preferences for location and value. Implementing this model involves phases such as data extraction, cleaning, outlier detection, and model training and testing, ultimately benefiting both buyers and sellers by providing scientifically-backed property prices.

2.3 Comparison between existing works

The literature we reviewed uses various methodologies and algorithms to predict house prices, reflecting the general status of research work in this area. Truong et al. [5] stress the use of advanced models and add otros factores como el Indice de Precios de Viviendas (HPI) for increased accuracy. Similarly, Satish et al. It can be interpreted that XGB is well-suited with lasso regression as it has performed brilliantly in terms of high accuracy by using machine learning algorithms like [6]. Thamarai et al. Zulkifley et al. [7] test the applieability in decision tree and multiple linear regression, while some others utilise various machine learning models. The best knows algorithms as reported by Paparrizos et al. [8] are artificial neural networks, correlational support vector regression and XGBoost.Full size image Adetunji et al. Modi et al., [9] than present a glimpse of the promise offered by Random Forest for clinical predictions based on individual profiles. In addition, the authors of [10] use a stacking technique to obtain very high accuracy. Dabreo et al. [11] and Mysore et al. The importance of dataset quality, in conjunction with the algorithm chosen are also highlighted by [12] (where KNN and XGBoost were able to reach very high accuracy). Jamil et al. [13] and Mohd et al. [14] concentrates on the prediction of green building price, determine Decision Tree and Random Forest are most effective. Yu et al. Deep learning model have been used by [15] as well, presenting a better result. Manjula et al. [16] and Chouthai et al. [17] emphasise feature engineering and model choice, with an example using ridge regression and decision trees. Gokalani et al. [18] and Piao et al. Focus on regional data and more advanced models like CNN [19] Mostofi et al. [20] and Sinha et al. Make a case for iterating between data transformations and ML model implementation phases to avoid reduced accuracy or user satisfaction [21]. Together, these studies illustrate the continuous progress in machine learning approaches to improve house price prediction performance and robustness.

TABLE 2.3 COMPARATIVE ANALYSIS WITH PREVIOUS WORK

SL	Author Name	Used Algorithm	Best Accuracy with Algorithm
1.	J.-W.Kim et al. [1]	Machine Learning	SVM= 84.6%
2	Y. Zhang-James et al. [3]	ML & RNN	RF= 73%
3	J.Downs et al [4]	ML	LR=86%
4	Satish, G. Naga, et al. [6]	ML	XGBoost=91%
5	Md. Maniruzzaman et al. [7]	ML	SVM=94.2%
6	Anshu Parashar et al. [9]	ML	Adaboost =84%
10	M. Modi, A. Sharma et al. [10]	Machine Learning	LR= 98.27%

2.4 Open Issues

Over the last decades, a number of problems and open issues have been identified in machine learning for predicting house prices although significant progress has been made. Some of the risks of using these models include; interpretability of deep models like XGBoost and Random forest because, even though the models are highly accurate, they hide the decision-making process. The fourth concern is of identifying the stability of methods with respect to changes in market situation or some unanticipated macroeconomic factors that impact the relevance of the forecast in the further periods. Also, the quality and consistency of data gathered through web scraping techniques are also a bottleneck because missing information or erroneous data affects the model’s performance. Also missing is the specification of evaluation metrics and methodologies that could be applied to comparing the findings of different studies.

2.5 Summary

From this literature review focusing on the prediction of house prices using ML techniques, the analyzed research area shows that the field is well-established and advancing steadily but still face ongoing issues and questions. Research proves that complex regression models like Polynomial Regression, SVR, XGBoost, and Random Forest perform the best in identifying the relationship between house features and prices. The results indicate that the number of keys, location, features of property, economic forecasts, and other parameters are always significant to maintain the precise outcome of the forecast. However, more work has to be done on model interpretability which is a challenge when using complex ensemble methods and the general model robustness to shift in market conditions ad well as the inconsistencies that are inherent in web scraped data. The issue of comparing evaluations across studies also highlights another obvious deficiency, namely the lack of standardization of the evaluation indicators and methods. Nevertheless, it is also clear that literature points to the ability of machine learning to improve the decision-making of real estate by offering superior and precise data on the behaviour of house prices. For furthering the agenda of developing more accurate explanations, enhancing data acquisition, and enhancing predictive accuracy in various real estate markets, research should continue persistently in the future.

CHAPTER 3
METHODOLOGY

3.1 Overview

The research method sub-section describes a clear approach that will be adopted in the generation of a predictive model for house prices. It is based on a detailed requirement collection that determines data sources, attributes (place, number of bedrooms, bathrooms, and more), as well as the goals of the research. This is followed by design specifications of a model that define the process of data preprocessing, missing values management, relevant features selection, and encodings of categorical data. Before proposing the model, Exploratory Data Analysis (EDA) is carried out in order to identify trends in data that exists and the relationship that may exist between the data.

It is due to their capacity to deal with non-linear data and cope with the intricacy of the data samples available The choice of the regression models includes Polynomial Regressor, SVR, XGBoost Regressor, and Random Forest Regression. Model training also involves the division of data into training and testing, choosing of the right hyper-parameters and assessing the model's accuracy based on set standards such as Mean Absolute Error and R-Squared.

I believe that such structure enhances the overall clarity of the organisation plan of the course focused on applying machine learning algorithm for the prediction of house prices, while adhering to both technical specifications and analysis objectives.

3.2 Proposed Methodology

The scheme proposed in this study aims to predict house prices using machine learning algorithms. The procedure consists of the following steps:

Data Collection:

The datasets we used in this research are crawled from a real estate portal Bproperty.com ltd by using different data miner tools and extracted almost 598 entries. There are seven features in the dataset: Location, No of bedrooms, Number of bathrooms, Area. Few more attribute Property id each entry represents a property with its location, the number of bedroom and bathroom it possess; area (in square feet), some additional indicators, and price.

Data Pre- Processing:

Data preprocessing included data imputation to fill in the missing values required for a complete dataset, and feature selection to find what attributes are most relevant. Moreover, all categorical variables were encoded for regression while EDA was performed to get a better idea of how the data is distributed and related

Feature Selection:

This included dealing with missing values in the data (for some features, more than 40% of records had NA) to make sure we have a complete dataset and feature selection process to identify most important attributes for model training. Further, the categorical variables were encoded in such a way that they could be used to perform regression analysis and Exploratory Data Analysis (EDA) was done for data distributions followed by relationships.

Encoding:

We encoded categorical features - location and property type for the regression models. To convert these categorical features to numerical format for the purpose of using them in machine learning models, techniques like one-hot encoding and label encoding were used.

Exploratory Data Analysis (EDA):

In order to extract metadata, I performed Exploratory Data Analysis (EDA) and discovered hidden patterns, trends and relationships in the dataset with all that crucial information on what attributes affect house prices. This included visualizing the distributions of data to see where variables have correlations with one another and whether there are any anomalies / outliers which could affect how your model will do.

Model Training:

We trained models with regression techniques such as linear-regression, decision-tree-regressor and ensemble methods like random-forest, gradient boosting algorithms on the preprocessed dataset. The models are trained on some data to learn the relationships between the features and 'house price', i.e. target variable, with their parameters tuned so as to minimize errors when predicting outputs:

Support Vector Regression

Similar to a classifying algorithm, Support Vector Regression (SVR) is another machine learning technique used in regression analysis by amending the SVM technique. They operate in such a way that SVR isolates a hyperplane in the number space of features that best separates ‘x’ in terms of a margin between an interval containing the actual values of the target variable and the predicted values. They have proven to be rather efficient, specifically where the relations between the features and the target variable are non-linear. SVR accomplishes this by using a kernel function such as radial basis function (RBF), polynomial or sigmoid in which the input data points are mapped into a higher dimensional feature space linear regression model is applied. SVR is less sensitive to overfitting thus it is appropriate for many regression problems in predictive modeling and actual life scenarios.

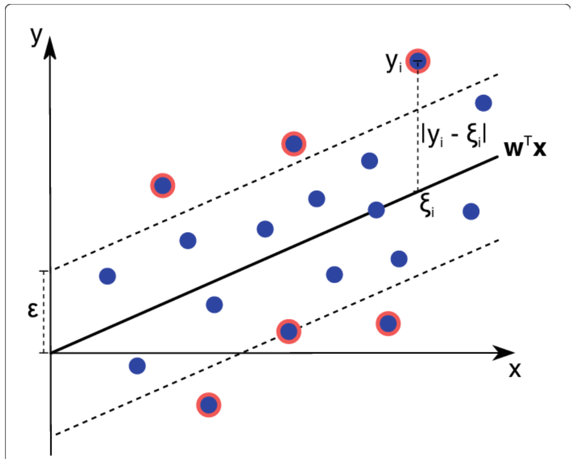

Figure 3.2.1: Support Vector Regression [22]

Polynomial Regression

A multinomial regression is a type of regression model derived from a linear model, where the main difference is the ability to fit data with a multinomial equation, thus assuming more complex nonlinear relationships between features and target variables Simple- This method and preferred when the line is not well defined. Multinomial regression gives more accurate results than simple linear regression in the case of curvilinear relationships. However, careful selection of model complexity is important, as it tends to capture noise in the data rather than realistic trends.

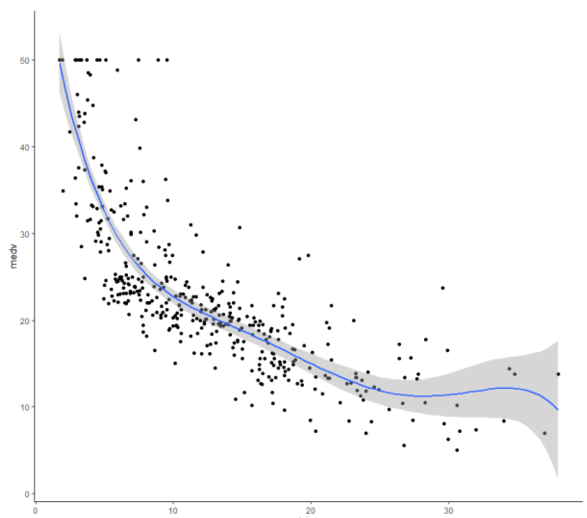

Figure 3.2.2: Polynomial Regression [23]

XGB Regression

The XGB Regression is an ensemble learning algorithm that is based on the XGBoost or the Extreme Gradient Boosting capable of high-speed and high-performing scalable regression. It systematically forms a sequence of decision trees to rectify errors of the previous models and boosts the dependability of the forecast. XGBoost uses a gradient boosting model that seeks to cut the overfitting by controlling both the bias and variance of the model. These are methods to deal with missing data, methods to control model complexity such as regularization, and efficiency in large data sets. It originates from achieving near state-of-the-art performance in many of the benchmarks and challenges in a range of machine learning competitions and practical tasks.

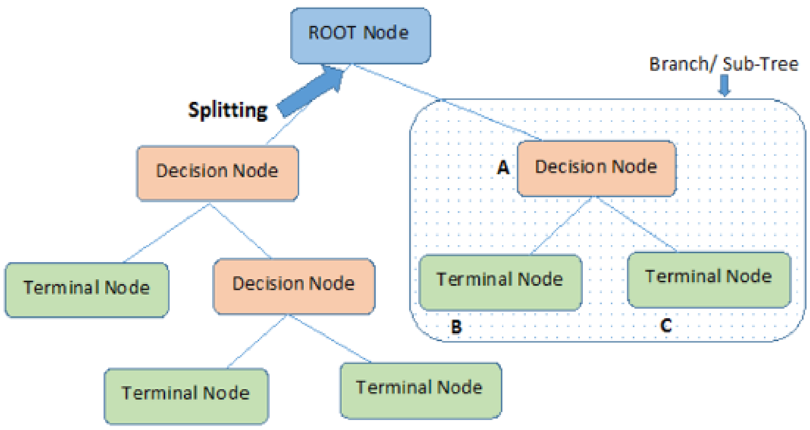

Figure 3.2.3: XGB Regression Architecture [24]

Random Forest Regression

Random forest regression is a versatile and robust learning algorithm used primarily for classification but also for regression functions. It works by generating multiple decision trees during training and determining the final results based on the average of the common (for classification) or predicted values (for regression) from individual trees Training information is the items in each tree are randomly sampled from the entire data set to prevent overfitting. Random forests are noise-proof and can handle large amounts of information and resources. They excel in providing modeling, probability estimation, and feature importance ranking in complex data, making them popular in a variety of industries including finance, healthcare, ecology and commerce.

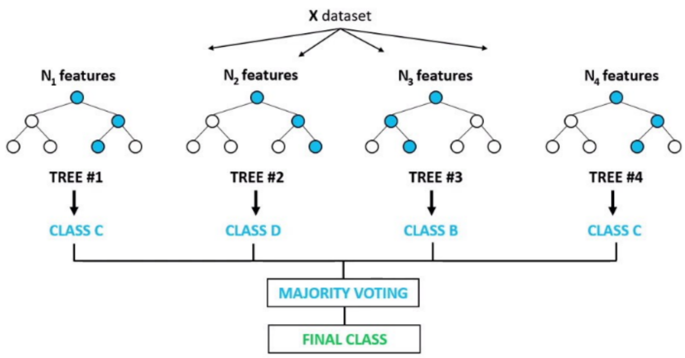

Figure 3.2.4: Random Forest Regression [25]

Model Evaluation:

In this research, model evaluation is done using performance metrics such as mean absolute error (MAE), mean squared error (MSE), R-squared, etc. to evaluate the accuracy and the efficiency of trained regression models. This evaluation process included checking predicted house prices against true values on a validation set, ensuring that the predictive models would actually be useful and reliable in production.

Test Model:

The test process was carried out using the best performing regression algorithm on an unseen dataset in order to determine how well it generalizes and performs. We were able to assess how well the final model performed by checking how accurate it was with its predictions compared to actual house prices, we confirmed that and so conclude about very good generalization of our robust approach which is also ready for practical usage.

Using this methodology, we hope to create accurate and reliable machine learning models for house price prediction.

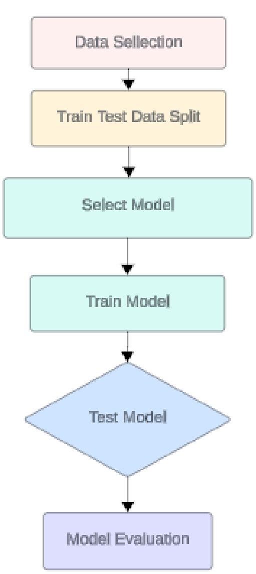

Figure 3.2.5: Work Flow

3.3 Hardware/ Software Requirement

Here are few key requirements that need to be satisfied while implementing the predictive model for house price prediction using regression analysis in order for this data algorithm to remain effective and robust as well.

For one, you need a go-to programming environment. Programming Language- Python has been selected as the programming language since it is highly versatile and comes with a larger ecosystem of libraries like Pandas for data manipulation, NumPy for numerical operations, Scikit-learn - for machine learning algorithms. Through libraries, Sage Maker provides benefits like high-performance data pre-processing and feature engineering for machine learning modeling and evaluation capabilities for models built using Jupiter Notebooks.

Second, companies need the computational resources to manage larger data sets and fit computation challenges. For a complex regression model, you would need to have some capability at your end (local machine or cloud-based platform) with good processing power and memory as well for extensive hyperparameter tuning.

Additionally, the versioning of code and documentation are necessary to keep proper track of a particular piece for integrity as well as replay data. Using SCMs such as Git helps in working collaboratively on code and keeping track of every change. Effective code documentation and methodological approaches add transparency while training-on-the-job.

In addition, data mining and application of principles in this collection is essential. By following privacy policies, terms of service and legal matters when scraping Bproperty data one can assure ethicality as a whole during the research process.

3.4 Project Management and Financial Analysis

Project and financial management are two very essential components of research as underlined by the house price prediction study. Another aspect of the management of the project entails making sure that there are objectives that are defined and both the timeline and responsibilities in undertaking the project are correctly assigned in the midst of relating the impending task with the goals of the team. The proper use of tools as the Gantt chart and the agile work processes can increase the performance of the project and guarantee the achievement of the target dates. From past experiences, daily meetings with the team members and daily progress reports help to identify and solve problems as they occur. Financial control is as important as operational management and includes the funding of several resources such as data purchase, software, and hardware, manpower, etc. These costs require funding usually in form of grants, sponsorships, or support from an institution. Another element to include in the ideal budgeting is contingencies to cater for additional expenses that may occur in the course of the project's life cycle. Proper utilization of funds makes it possible for the project to acquire the needed funds for its main goal without compromising the money it has. Timely financial checks and reporting of the used money assist in minimizing misuse and to effectively use the available money. A fitting factor into use and maximization of resources with a need to balance the budget is important to ensure the predictive models formulated present actual facts but are also affordable within the existing resource constraints. The overall result of effective general management over all the identified project tasks and all those required to fund a research implies an effective and impactful research result.

TABLE 3.4.1: PROJECT MANAGEMENT TABLE

Work	**Time**
Data Collection	1 month
Papers and Articles Review	3 months
Experimental Setup	1 month
Implementation	1 month
Report Writing	2 months
Total	8 months

Table 3.4.2: Estimated Cost

SN	Components	Estimated Cost (BDT)
01.	Hardware	500-1500
02.	Software and Tools	500-1500
03.	Data Collection and Processing	5000-6000
04.	Documentation and Report Writing	500-1500
05.	Miscellaneous	500-1500
06.	Contingency	500-1500
	Total Estimated Cost	7500-13,500

3.5 Summary

The objective of the research is to entail a feasibility analysis and, centering on the collected dataset from Bproperty using web scraping, develop machine learning models for predicting the prices of houses. It starts with the literature review of the previous works done in the production of housing market prediction models and sheds light on the efficiency of Polynomial Regression, SVR, XGBoost Regression, and Random Forest Regression. As such, location, characteristics of the house, and the state of the market can be recognized as critical elements when it comes to defining models for house prices. Some of the steps involved in the methodology include the handling of missing values and categorical data transformation including one-/many-hot encodings then EDA to assess the data and guide the choice of the model. The efficiency of each model is confirmed basing on standard measures such as Mean Absolute Error and R-squared to ascertain which predictor is most accurate. Therefore, this research helps in refining the decision making in real estate by developing an effective, accurate and data-driven model for estimating house prices.

CHAPTER 4

IMPLEMENTATION

4.1 Overview

The implementation phase involves using the defined methodology in building and assessing the effectiveness of the models that relate to housing price prediction. It takes its origin from data cleaning phase, treatment of missing values in the dataset received from Bproperty, and categorical variable conversion processes that involve pre-processing the data in order to analyses them. Before model design, exploratory data analysis is initiated to find out basic information regarding the features of data and search for any valuable relation or other conditions that can impact the models' efficiency.

After EDA, the regression models like Polynomial Regression, SVR, XGBoost Regression, and Random Forest Regression are fit on the preprocessed data. Training a model includes tuning the hyperparameters which can be done with the help of the grid search or random search in order to increase predictive power. To measure the performance of each model, the evaluation statistics like Mean Absolute Error (MAE) and R-squared are applied to a second set of data to compare the model's performance and its ability to generalize.

The implementation phase seeks to come up with a better predictive model that would be of help in real estate business through enabling decision makers to make right decisions regarding the fluctuating price in the market.

4.2 Train Model

The training process of the predictive model includes the usage of the preprocessed dataset into a training and testing partitions, where the most commonly used separation ratio is seventy percent to twenty percent. The training set is used to fit various regression models: Polynomial Regression the SVR model, XGBoost Regressor model, the Random Forest Classifier. Each model is then subjected to hyperparameter optimization where the model's hyperparameters are adjusted to improve the model's performance and common methods include grid search or random search. It's done through training the models on how to identify correlations between the input variables which are location, number of bedrooms, number of bathrooms, area in sqft, type, and amenities of the houses/ apartments and the output which is the price. Cross-validation is used as the models' validation process to omit over-fitting and check whether models are suitable for applying to unseen data. MAE and R-squared are some of the performance measures that are used during the training to find the model with the highest accuracy. This training process is very intensive and it is designed to produce a strong model that can estimate the prices of the houses based on the described characteristics.

4.3 Model Evaluation

The last phase of the entire modeling process is model evaluation, whereby the efficiency and effectiveness of the trained models in the machine learning process are analyzed. Once interpreted models such as Polynomial Regression, SVR, XGBoost Regression, and Random Forest Regression were trained on the given dataset, then the performance of the trained models is tested on the unseen test set. This evaluation also helps to check if the models can generalize in real-world settings since what is seen in the test set is novel and unseen data. We consider Mean Absolute Error (MAE) and R-squared (R^2) as main evaluation criteria, whereas car allocation is used as the benchmark for comparison. MAE gives the average magnitude of the difference between predicted values and actual house prices which in turn avoids the complexity of interpretation that squared errors present. Measuring the ability of the model to explain the inputs, R-squared reveals how much of the variance in target variable is explained by the input features. Further, such techniques as cross validation are employed in the process of determining the stability and viability of the models; this does away with the risk of over fitting. This is done in a way where portions of the training data are taken and some portions are used for testing while the performance is averaged. Based on the evaluation metric obtained from the testing of different models, the model is selected as the best model based on the accuracy of the epoch2, test set accuracy, and the time complexity. The rigorous evaluation process implemented guarantees that the chosen model is fit for purpose in the simulation of house prices in different real estate settings.

4.4 Summary

Implementation phase consisted of a number of well-defined sub-steps in order to design and assess the machine learning models for the purpose of house price prediction. This was followed by data preprocessing in which data from Bproperty was preprocessed, features with missing values were managed while categorical features were encoded. After data collection, Exploratory Data Analysis (EDA) was done to understand the distribution of data and relationship between features. Subsequently, the whole set of data was divided into training and testing data sets.

Four regression algorithms—Polynomial Regression, SVR, XGBoost Regressor, and Random Forest Regression—are applied to the preprocessed data. The hyperparameters were tuned using cross-validation using a web search or random search to determine the best model. Cross-validation was also used to assess the generalizability of the models. Metrics such as mean absolute error (MAE) and R-square (R2) were used to assess accuracy and generality.

Due to the encompassing of all the methods, the most suitable and precise model was obtained for predicting House prices and useful for real estate.

CHAPTER 5

RESULT AND ANALYSIS

5.1 Overview

The result analysis phase covers the process of understanding the outcome measures for the trained machine learning models. The evaluation criteria included prediction accuracy, which was measured with Mean Absolute Error (MAE); and goodness of explanation, estimated with R-squared signifying the proportion of the variation in the dependent variable explained by the independent variables according to each model—Polynomial Regression, SVR, XGBoost Regression, and Random Forest Regression. To determine which model performs better, the analysis is carried out on these metrics for several models. For model validation and to get a better understanding of the functioning of the models, a few qualitative measures like 'residual plot' and 'feature importance chart' were used. By looking at residual plots, one can diagnose systematic errors and also, evaluate the assumptions made by the regression models. When it comes to visualizations, feature importance charts from other models like the XGBoost model or Random Forest model help in understanding the most important features in the determination of house prices that may be of practical significance to the stakeholders. This not only helps to select the best model, but also understand the nature of the results, features of the data set, and contributions of various features that contribute to the model's predictive ability, which also increases the effectiveness of applying the model in the real estate profession.

5.2 Experimental Result

We have used several regression algorithms for predicting the prices of houses in this study and we measured their performance using R^2 score. The Polynomial Regressor had a R^2 score of 27.03%, which showed it was not able to capture the essence of how our dataset work and predict good values The Support Vector Regressor (SVR) boosted the model with an R^2 score of 40.14%, much better but still on moderate level.

There were also large apparent gains for ensemble methods. The XGB Regressor with its resilient nature of handling intricate data patterns led to an admirable R^2 score of 97.16%. This showed us the relative importance of each feature with respect to prices of houses.

The Random Forest Regression had the highest accuracy of all these classifiers, with a R^2 score 97.73%. Since it uses multiple decision trees to make predictions, the ensemble approach of this algorithm is capable of capturing complex interactions among variables, providing good accuracy in prediction.

Random Forest and XGBoost of ensemble methods dominate in house price prediction Their results indicate that the performance of model can be improved when adopting RF (0.9331) or XGBoost (0.9324). As the BI demonstrated, random forests excel in linear and non-linear correlated variables interaction data types which is similar to a complex real-world dataset like house prices. We may conclude that to have a successful prediction of house price we need advanced machine learning techniques.

Table 5.2: Performance Evaluation

	MAE	MSE	RMSE	R2 Score
SVR	15.22	524.77	22.90	0.40
Random Forest Regression	2.92	19.89	4.46	0.97
XGBoost Regression	3.25	24.87	4.98	0.97
Polynomial Regression (Degree=2)	20.12	639.75	25.29	0.27

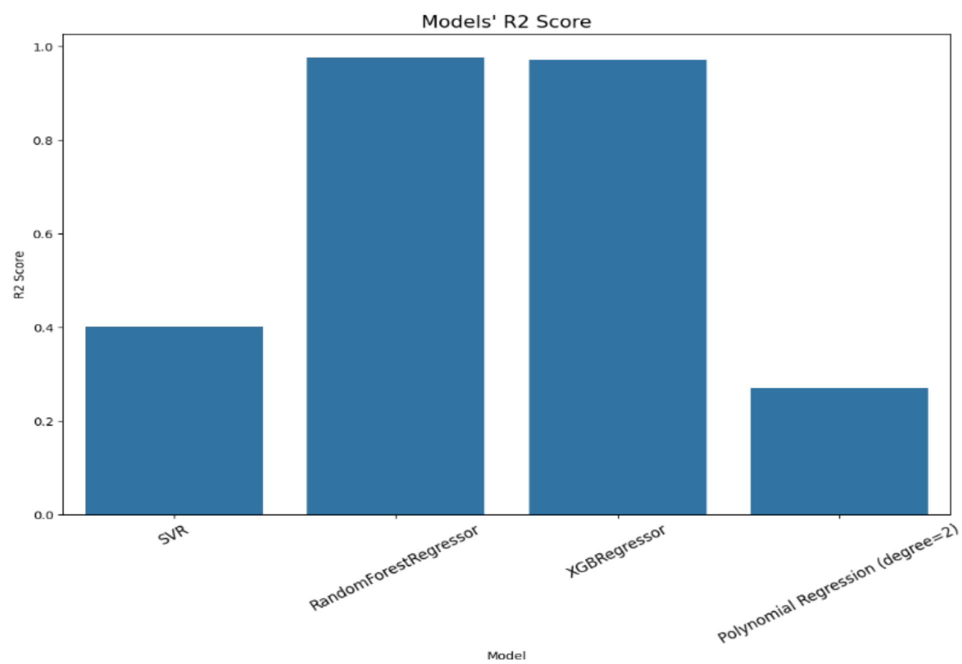

Figure 5.2: Model Accuracy Bar Plot

5.3 Performance and Comparative Analysis

Performance and comparative analysis include a capacity check that shows the efficiency of each machine learning model-Polynomial Regression, SVR, XGBoost Regression, Random Forest Regression to estimate the prices of the given houses. Files were tested for their ability to accurately portray target variables and to generalize well using accuracy and predictive measure such as Mean Absolute Error (MAE) and R-squared (R^2).

As expected, the classifiers possessing higher R^2 scores with lower MAE were more accurate and better at explaining variability in house prices to a certain extent, and these classifiers were the XGBoost Regression and, to a lesser extent, the Random Forest

Regression. These models take advantage of ensemble that is using many decision trees to accomplish data relationships in the complex patterns.

In contrast, the Polynomial Regressor, and SVR offered some ideas about the non-linear trends in the data but failed to generalize the results to new data sometimes due to overfitting or underfitting. Such comparative analysis pointed into the fact that though simple models are easy to explain, XGBoost and Random Forest proved to be the best models that give the trade-off for complexity and accuracy hence rendering them the best to be used in this study for accurate house price prediction.

5.4 Summary

The result and analysis phase confirmed specific findings important to the comparisons and analysis of various machine learning model for house price prediction with high accuracy. The Polynomial Regressor, SVR, XGBoost Regressor, and Random Forest Regression were used and subsequently assessed for Mean Absolute Error (MAE) and R-squared (R^2). XGBoost Regressor and Random Forest Regression performed better among all the models as they had a much higher accuracy and coefficient of determination and can explain insights from the data that other models might miss since they are based on the ensemble learning technique. Reliability and interpretability of these models were also clear from the residual plots and feature importance charts which provided additional validation to the above findings. Although Polynomial Regressor and SVR were helpful in finding out about a non-linear relationship, they are not efficient when it comes to predicting results using a separate set of data.

CHAPTER 6

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

6.1 Impact on Life

The ability to predict house prices with great accuracy with the help of Machine Learning Model has a lot of ramifications in several areas of life including real estate, monetary planning and stability of societies. Automatic price forecasts help the homebuyers or investors to have a clear picture about the market trends and thus can take more rational decisions about buying or selling houses or can invest in some property. This has the effect of minimizing risks related to financing and assisting in the search for high-profit investment opportunities.

To the real estate experts including the agents and developers, these sophisticated predictive models improve market reviews, determination of propitious price margins, and consumers' recommendations that consequently improves the quality of service provided and their competitiveness. Credit institutions also benefit from the more accurate and specific screening criteria for the mortgage credit, and risk evaluation again resulting into efficient and nondiscriminatory credit processes.

On an even larger scale, these house price forecasts are very useful for the general stability of the economy due to the fact that they help politicians determine when there is a bubble and when there is a bust in the housing markets. It can reduce situations when economic crisis arises as a result of a fluctuating real estate business.

Additionally, since housing is one of the basic human necessities, these models may also foster social stability and health, as affordable and stable housing is less stressful than an unstable, expensive market for the population's dwelling. Therefore, inclusive of machine learning in the prediction of house prices, one can see a way to improve the quality of their lives, and the economic turnover.

6.2 Impact on Society & Environment

Therefore, the direct influence of this work, which predicts the house prices by the means of regression analysis on environment, seems to be quite limited; however, it was also noted that the overall influence can threaten/suppose the environmental sustainability in several ways.

Accurate estimations in establishing the value of property decrease the risks in investment by offering proper assessment of the use of land as well as property in cities. It enables one to determine in what aspect development is probable according to the predicted prices of houses in order to make adequate decisions as to where to put cash and as to where houses should be constructed. Due to urban sprawl, it can halt the urbanization of cities and inhibit their unrestrained expansion into rangelands and affect the negative ecological footprint that large infrastructural projects create on wildlife habitats.

Likewise, by extending the enhancement of the optimisation of plans regarding the utilisation of land and structures, prognostic models promote forms of sustainable development of cities and healthy surroundings of communities. This encompasses aiding of 'Mixed Use' developments, enhancing mass transit in the urban centers and integrating green space within built form. Such practices not only aid in effective use of resources that is available, but also are useful in drawing down air pollution, consumption of energy and strengthening the environment's resilience.

Furthermore, the correct trends in the housing market and the investment climate can be forecasted; therefore, the further development occurs in the direction of green buildings and energy technologies. Therefore, the information given, showing the inclination towards the use of sustainable development practices at the international level, helps enhance the fight against climate change and the formation of a more sustainable environment.

6.3 Ethical Aspects

Therefore, the use of machine learning in the aspect of price prediction of the homes is highly unethical. The first is data privacy as many of the models demand large datasets and these datasets may include the personal details of the subjects. Measures that require protection levels for the individual in the process of data collection and usage must also be legal. Another issue on ethical concern is bias in the models. However, if the training data is a record of systematic discriminations in housing markets which is evident in most cases, the models can amplify these discriminations through prejudice judgments and discriminating on specific groups. The following is important in the identification of bias and measures that should be taken to enhance fairness: The concept of bias, bias detection, and bias management. Similarly, another advantage is the interpretability and other characteristics of the model. Such models should be able to justify the reasoning behind making the predictions and all the stakeholders should be able to understand how the models evolve. These bolster trust and make it possible to come to good decisions since all the important details have been forthcoming. Last, the ethical issues are linked to potential socioeconomic implications of achieving almost perfect prediction, to some extent with manipulation of the stock market or to put it in other words, exclusion of middle- and even working-class people from some districts. Navigating through these ethical components allows the developers of machine learning in real estate to build systems that are both ethically sound and make a positive social contribution.

6.4 Sustainability Plan

Any long-term oriented sustainability plan of a course on predicting house prices by the means of regression analysis should include measures preventing negative impact and unethical behavior. To begin with, keeping the data up to date and fine-tuning the approach to data acquisition like having the latest data from either Bproperty or any other reputable source as the database is updated periodically to correspond to the current market conditions.

Secondly, encouraging research collaboration and openness with other researchers and stakeholders helps refine the models and sustain the community's and stakeholders' trust. Exchange of the methodologies, results, and the best practice of the models improves the dependability and relevance in various real estate scenarios.

Thirdly, the ethical principles are considered throughout the data analysis pipeline and predictive modeling, starting from data collection and preprocessing and ending with the model deployment, which makes the usage of the predictive analytics responsible and equitable. This implies issues of prejudice, privacy, and compliance with the law and with guidelines on best practice.

Finally, it enhances the education and awareness level of the stakeholders concerning the applicability and limitation of the predictive models in real estate. As a result, the research primarily aligns with the real estate industry and society's general well-being principles of sustainability and ethical practices.

6.5 Summary

The use of machine learning in prediction of house price affects the society, the environment and sustainability in the following ways. To the society, price forecasts help increase the information available regarding the housing market in order to improve decisions made by investors, buyers and sellers. This will help in making fair chances to get a house and minimize issues with the provision of credit and loans.

From the environmental point of view, improved predictions can help in defining trends and needs that should be taken into consideration in the framework of proper urban planning by providing priority for the construction of rather environmentally friendly and energy efficient houses. This is used in preventing the expansion of developmental impacts and encouraging the application of green technologies.

As for the aspect of sustainability, predictive models can suggest long-term investment, which is profitable both for investors and the economy, rather than searching for a fast buck. These models are helpful in preventing market instabilities and controlling the fare charges hence playing an efficient role in the maturity of the economy and general wellbeing of the society. Therefore, the application of machine learning to support house price prediction creates a win-win situation more prone to the enhancement of social and environmental consequence beneficial to all the parties.

CHAPTER 7

Conclusion and Future Work

7.1 Conclusions

Therefore, the following overall conclusions and recommendations can be made regarding the machine learning regression analysis of this particular study conducted on the prediction of house prices. Key findings concerning the effect of predictors on house pricing have been determined going by the results from the data collected from Bproperty and the use of regression models.

Location comes out as a significant factor of house prices due to the fact that houses in admired neighborhoods will attract higher prices. This is somewhat true and the number of bedrooms and bathrooms are key attributes that are have high correlation with the property prices.

Second, the given area of the property (in square feet) also show a very high significance in the context of prices, which can be explained by the desire of buyers to have a large living space. Additional values are added by the type of property and available facilities like, garage space, and sports facilities among others that form part of the pricing strategy.

Thirdly, there is an indication that mastering algorithms such as random forest and gradient boosting methods in machine learning was efficient in determining the house prices. These models are generally found to perform better than other basic regression methods at identifying complexities in the character of association as well as interactions that occur among the predictors.

7.2 Future Suggested Works

Such are the specifics of house price prediction using machine learning; the following could be considered areas of improvement in the future research work. First, identifying further approaches based upon the integration of properties of the numerous types of the regression analysis can improve the effective and stable predictive performance as far as incorporating the selection of the features combining the benefits of both the neural networks and the selected regression models.

Secondly, including additional variables beyond the traditional housing characteristics, including socio-demographic characteristics, environmental features, and infrastructural advancement, among others, might offer a broader understanding of the fluctuations in the housing market and enhance the models' accuracy. Moreover, improving ethical concerns in the real estate AI through better approaches of data collection and model release is a compelling task. Thus, enhancing hindsight with real-time foresight models that can easily adapt to market force and other economic changes as they occur will be highly actionable for the relevant stakeholders. In this way, similar avenues should be explored in the further research to perfect and develop the machine learning approaches to the house price prediction and provide helpful tools for decision-making in the constantly fluctuating real estate market.

7.3 Limitations

Some of the limitations that are found in current usage of machine learning in predicting house prices include the following. First of all, using past statistics may not factor recent changes or additional circumstances that may alter the market drastically, thus is not effective in predicting the future market trends. This challenge is made worse by the fact that real estate markets are constantly evolving, meaning that factors that drive demand as well as prices are constantly shifting. Secondly, problems with data quality that include the lack of complete or representative data can severely affect the model's forecast generations. Another major methodological issue remains the ways to guarantee data accuracy and their representativeness. Furthermore, the interpretability of the model although may suffer a little space when using the complex ensemble methods such as XGBoost and Random Forest. That is why it is so important to know how these models come up with these numbers, especially if these decisions can perhaps be life-changing. Finally, the alongside of a regulatory and ethical perspective, it is worth mentioning that privacy issues and possible impacts of outcome and explanatory injustice present critical challenges in the application of machine learning technologies for real estate.

REFERENCES

[1] B. Miao and Y. Zhang, "A Feature Selection Method for Classification of ADHD," 2017 4th International Conference on Information, Cybernetics and Computational Social Systems (ICCSS), Jul. 2017

[2] B. Abibullaev and J. An, "Decision Support Algorithm for Diagnosis of ADHD Using Electroencephalograms," Journal of Medical Systems, vol. 36, no. 4, pp. 2675–2688, Jun. 2011

[3] A. Mueller, G. Candrian, J. D. Kropotov, V. A. Ponomarev, and G.-M. Baschera, "Classification of ADHD Patients on the Basis of Independent ERP Components Using a Machine Learning System," Nonlinear Biomedical Physics, vol. 4, no. S1, Jun. 2010

[4] Gurcan Taspinar and Nalan Ozkurt, "A Review of ADHD Detection Studies with Machine Learning Methods Using rsfMRI Data," NMR in biomedicine, Mar. 2024

[5] J.-W. Kim, V. Sharma, and N. D. Ryan, "Predicting Methylphenidate Response in ADHD Using Machine Learning Approaches," International Journal of Neuropsychopharmacology, vol. 18, no. 11, p. pyv052, May 2015

[6] Navya Sethu and R. Vyas, "Overview of Machine Learning Methods in ADHD Prediction," Springer eBooks, pp. 51–71, Jan. 2020

[7] Y. Zhang-James, Q. Chen, R. Kuja-Halkola, P. Lichtenstein, H. Larsson, and S. V. Faraone, "Machine-Learning Prediction of Comorbid Substance Use Disorders in ADHD Youth Using Swedish Registry Data," Journal of Child Psychology and Psychiatry, vol. 61, no. 12, pp. 1370–1379, Apr. 2020

[8] J. Downs et al., "Assessing Machine Learning for Fair Prediction of ADHD in School Pupils Using a Retrospective Cohort Study of Linked Education and Healthcare Data," BMJ Open, vol. 12, no. 12, Dec. 2022

[9] M. L. Cervantes-Henríquez, J. E. Acosta-López, A. F. Martinez, M. Arcos-Burgos, P. J. Puentes-Rozo, and J. I. Vélez, "Machine Learning Prediction of ADHD Severity: Association and Linkage to ADGRL3, DRD4, and SNAP25," Journal of Attention Disorders, vol. 26, no. 4, pp. 587–605, May 2021

[10] A. Kautzky et al., "Machine Learning Classification of ADHD and HC by Multimodal Serotonergic Data," Translational Psychiatry, vol. 10, no. 1, Apr. 2020

[11] Md. Maniruzzaman, J. Shin, Md. Al Mehedi Hasan, and A. Yasumura, "Efficient Feature Selection and Machine Learning Based ADHD Detection Using EEG Signal," Computers, Materials & Continua, vol. 72, no. 3, pp. 5179–5195, 2022

[12] M. Duda, R. Ma, N. Haber, and D. P. Wall, "Use of Machine Learning for Behavioral Distinction of Autism and ADHD," Translational Psychiatry, vol. 6, no. 2, pp. e732–e732, Feb. 2016

[13] A. Parashar, N. Kalra, J. Singh, and R. Kumar Goyal, "Machine Learning Based Framework for Classification of Children with ADHD and Healthy Controls," Intelligent Automation & Soft Computing,

vol. 28, no. 3, pp. 669–682, 2021

[14] X. Peng, P. Lin, T. Zhang, and J. Wang, “Extreme Learning Machine-Based Classification of ADHD Using Brain Structural MRI Data,” PLoS ONE, vol. 8, no. 11, p. e79476, Nov. 2013

[15] D. Andrea, Harun Pirim, and D. Grewell, “ADHD Prediction in Children through Machine Learning Algorithms,” Lecture Notes in Networks and Systems, pp. 89–100, Jan. 2024

[16] H. Christiansen et al., “Use of Machine Learning to Classify Adult ADHD and Other Conditions Based on the Conners’ Adult ADHD Rating Scales,” Scientific Reports, vol. 10, no. 1, Nov. 2020

[17] W. Das and S. Khanna, “A Robust Machine Learning Based Framework for the Automated Detection of ADHD Using Pupillometric Biomarkers and Time Series Analysis,” Scientific Reports, vol. 11, no. 1, Aug. 2021

[18] A. Yasumura et al., “Applied Machine Learning Method to Predict Children with ADHD Using Prefrontal Cortex Activity: a Multicenter Study in Japan,” Journal of Attention Disorders, p. 108705471774063, Nov. 2017

[19] Md. Maniruzzaman, J. Shin, and Md. A. M. Hasan, “Predicting Children with ADHD Using Behavioral Activity: a Machine Learning Analysis,” Applied Sciences, vol. 12, no. 5, p. 2737, Mar. 2022

[20] J. Anuradha, Tisha, V. Ramachandran, K. V. Arulalan, and B. K. Tripathy, “Diagnosis of ADHD using SVM algorithm,” Bangalore Annual Compute Conference, Jan. 2010

[21] P. Mikolas et al., “Training a machine learning classifier to identify ADHD based on real-world clinical data from medical records,” Scientific Reports, vol. 12, no. 1, p. 12934, Jul. 2022

[22] W. Das and S. Khanna, “A Novel Pupillometric-Based Application for the Automated Detection of ADHD Using Machine Learning,” Sep. 2020

[23] H. W. Loh, C. P. Ooi, P. D. Barua, E. E. Palmer, F. Molinari, and U. R. Acharya, “Automated detection of ADHD: Current trends and future perspective,” Computers in Biology and Medicine, vol. 146, p. 105525, Jul. 2022,

[24] Yanli Zhang-James, E. C. Helminen, J. Liu, B. Franke, M. Hoogman, and S. V. Faraone, “Evidence for Similar Structural Brain Anomalies in Youth and Adult Attention-Deficit/Hyperactivity Disorder: A Machine Learning Analysis,” bioRxiv (Cold Spring Harbor Laboratory), Feb. 2019

[25] C. Park et al., “Machine Learning-Based Aggression Detection in Children with ADHD Using Sensor-Based Physical Activity Monitoring,” Sensors, vol. 23, no. 10, pp. 4949–4949, May 2023

Appendix A

Course Outcomes, Complex Engineering Problems (EP) and Complex Engineering Activities (EA) Addressing

Title: House Price Prediction in Dhaka Using Machine Learning.

Student ID: 203-15-3917 & 203-15-3902

CO Description for FYDP

CO	CO Descriptions	PO
Phase -I		
CO1	Integrate recently gained and previously acquired knowledge to Heart disease prediction problem for the Final Year Design Project (FYDP)	**PO1**
CO2	Analyze different aspects of the goals in designing a solution for this FYDP	**PO2**
CO3	Explore diverse problem domains through a literature review, delineate the issues, and establish these goals for the FYDP	**PO4**
CO4	Perform economic evaluation and cost estimation and employ suitable project management procedures throughout the development life cycle of the FYDP	**PO11**
Phase -II		
CO5	Design and develop technical solutions and system components or processes that meet specified requirements, ensuring compliance with public health and safety standards, as well as considering cultural, socioeconomic, and environmental factors in this FYDP	**PO3**
CO6	Choose and apply appropriate methodologies, resources, and contemporary engineering and IT technologies to address complex engineering processes, encompassing prediction and modeling, while adhering to relevant constraints in this FYDP	**PO5**

CO7	Analyze societal, health, safety, legal, and cultural considerations, along with associated responsibilities, in the context of professional engineering practice and the resolution of this problem, employing logical reasoning guided by contextual understanding.	**PO6**
CO8	Comprehend and evaluate the enduring sustainability and impact of professional engineering endeavors in addressing intricate engineering challenges within social and environmental frameworks.	**PO7**
CO9	Implement ethical principles and adhere to professional standards and norms in this FYDP	**PO8**
CO10	Capable of operating proficiently both individually and as a team member or leader across diverse teams and interdisciplinary settings in this FYDP.	**PO9**
CO11	Proficiently communicate with the engineering community and broader society regarding complex engineering endeavors, including the ability to comprehend and generate comprehensive reports and design documentation, as well as provide and receive clear instructions throughout this FYDP.	**PO10**
CO12	Acknowledge the importance of self-directed and life-long learning within the evolving landscape of technology, and possess the readiness and capability to engage in lifelong learning endeavors.	**PO12**

Addressing CO (1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP), and Attainment of Complex Engineering Activities (EA)

Addressing CO (1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP):

SN	EP Definition	Attainment	CO	Justification (With Knowledge Profile)	References
1.	**EP1: Depth of Knowledge required**	Yes	CO1, CO2, CO3, CO5, CO6, CO7 and CO8	This project demonstrates **fundamental engineering (K3)** principles by employing machine learning, data preprocessing techniques. The project demonstrates **specialist knowledge (K4)** by conducting machine learning with LR, DT, GNB, GB, DT enhancing for house price prediction correctly.	**Page no:** [12-16] **Section:** [3.2] **Page no:** [1] **Section:** [1.1]
				The project applies **engineering practice & design (K5)** by the figure of process of experiments. The project addresses **engineering practice & technology (K6)** by employing Machine Learning Model,	**Page no:** [12-16] **Section:** [3.2] **Page no:**

					[17-18] **Section:** [3.4]
				This project ensures to **K8 (Research Literature)** by synthesizing insights from recent studies, House Price Prediction using Machine learning, showcasing a comprehensive understanding of current methodologies.	**Page no:** [5-10] **Section:** [2.2, 2.3]
2.	**EP2: Range of Conflicting Requirements**	Yes	CO2, and CO7	This project addresses **EP-2** by recognizing the hurdles in House Price Prediction, including limitations of traditional methods and the complexities of integrating using machine learning. Through comparative analysis, it confronts challenges in understanding spatial distributions, offering insights for refining diagnostic methodologies.	**Page no:** [10-11] **Section:** [2.4, 2.5]
3.	**EP3: Depth of analysis required**	Yes	CO2, and CO6	This project addresses **EP-3** by meticulously comparing experimental outcomes, highlighting using machine learning as the chosen significant solution for enhancing House Price Prediction approaches.	**Page no:** [20] **Section:** [4.2]

4.	**EP4: Familiarity of Issues**	Yes	CO8	This project's interdisciplinary approach extends beyond computer science and engineering, impacting House Price Prediction correctly **EP-4**.	**Page no:** [5-10] **Section:** [2.2, 2.4]
5.	**EP5: Extends of application codes**	No	CO5	N/A	N/A
6.	**EP6: Extends of stakeholders involved and**	No	CO8	N/A	N/A
7.	**EP7: Interdependence**	Yes	CO5	This project's comprehensive approach addresses high-level problems by integrating various components across data collection, statistical analysis, and proposed methodology, ensuring a holistic solution to complex challenges in data collection from survey which ensures **EP-7**.	**Page no:** [12-17] **Section:** [3.2, 3.3, 3.4]

Addressing CO11 with Complex Engineering Activities (EA) [Some or all of the following]:

SN	EA Definition	Attainment	CO	Justification	References
1.	**EA1: Range of resources**	Yes	CO11	Our project utilizes diverse resources such as high-performance computing infrastructure, machine learning frameworks, annotated datasets, and ethical considerations to ensure systematic research and contribute to advancements House Price Prediction.	**Page no:** [18-19] **Section:** [3.5]
2.	**EA2: Level of interaction**	No		N/A	N/A
3.	**EA3: Innovation**	No		N/A	N/A

4.	**EA4: Consequences for society and the environment**	Yes		This project contributes to society by improving House Price Prediction methods, while also promoting environmental sustainability by employing efficient computational resources and adhering to ethical guidelines for medical data privacy.	**Page no:** [22-24] **Section:** [5.1, 5.2, 5.4]
5.	**EA-5: Familiarity**	Yes		This project expands upon existing research by examining a novel approach in House Price Prediction through machine learning model demonstrated through preliminary terminologies and a comprehensive comparative analysis, offering new insights into the field.	**Page no:** [6-12] **Section:** [2.1, 2.3]

Addressing CO (4, 9, 10, and 12):

SN	COs	Attainment	Justification	References
1	CO4	Yes	This project addresses **CO4** by integrating effective project management and financial oversight, ensuring meticulous planning, resource allocation, and budget estimation for optimal resource utilization throughout the research lifecycle.	**Page no:** [3-4] **Section:** [1.6]
2	CO9	Yes	The project demonstrates adherence to ethical principles by prioritizing privacy, obtaining informed consent, and transparently documenting	**Page no:** [23-24] **Section:**

			the research process, ensuring responsible knowledge dissemination and societal well-being through the ethical application of advanced classification technologies which comply with **CO9**.	[5.3]
3	CO10	Yes	We did the research as a team. One of us collected the dataset and the other one did the algorithm and coding part.	**Page no:** [12-16, 22-23] **Section:** [3.2, 5.2]
4	CO12	Yes	The project's dedication to continuous learning **(CO12)** and adaptation within the dynamic technological landscape is reflected in its comprehensive data collection, rigorous statistical analysis, meticulous methodology development, and thorough experimental results and analysis, showcasing a commitment to staying updated and refining techniques to address modern challenges.	**Page no:** [12-19, 20] **Section:** [3.2, 3.3, 3.4, 3.5, 4.2]

Turnitin Originality Report

Processed on: 13-Jul-2024 13:24 +06
ID: 2416048702
Word Count: 13091
Submitted: 1

House Price Prediction in Dhaka Using Machine Learning Algorithm By Jannatul Ferdous

Similarity Index	Similarity by Source	
21%	Internet Sources:	15%
	Publications:	6%
	Student Papers:	15%

2% match (student papers from 30-Jun-2024)
Submitted to Daffodil International University on 2024-06-30

2% match (student papers from 01-Jul-2024)
Submitted to Daffodil International University on 2024-07-01

1% match (student papers from 12-Jul-2024)
Submitted to Daffodil International University on 2024-07-12

1% match (student papers from 07-Apr-2018)
Submitted to Daffodil International University on 2018-04-07

1% match (student papers from 29-Jun-2024)
Submitted to Daffodil International University on 2024-06-29

1% match (Internet from 06-Aug-2022)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/8291/191-15-12253.pdf?isAllowed=y&sequence=1

1% match (Internet from 03-Jun-2024)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/12498/26688.pdf?isAllowed=y&sequence=1

1% match (student papers from 12-May-2024)
Submitted to Leeds Beckett University on 2024-05-12

< 1% match (student papers from 31-Mar-2019)
Submitted to Daffodil International University on 2019-03-31

< 1% match (student papers from 12-Jul-2024)
Submitted to Daffodil International University on 2024-07-12

< 1% match (student papers from 30-Jun-2024)
Submitted to Daffodil International University on 2024-06-30

< 1% match (student papers from 30-Jun-2024)
Submitted to Daffodil International University on 2024-06-30

< 1% match (Internet from 17-May-2024)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/12327/26886.pdf?isAllowed=y&sequence=1

< 1% match (Internet from 05-Jan-2022)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/5648/162-15-7817.pdf?isAllowed=y&sequence=1

< 1% match (Internet from 22-Jul-2023)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/10336/22957.pdf?isAllowed=y&sequence=1

< 1% match (Internet from 06-May-2023)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/8871/181-15-10721.pdf?isAllowed=y&sequence=1

< 1% match (Internet from 21-Jul-2023)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/10133/22773.pdf?isAllowed=y&sequence=1

< 1% match (Internet from 05-Aug-2023)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/10597/23358.pdf?isAllowed=y&sequence=1

< 1% match (Internet from 02-Apr-2022)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/7649/153-15-6568%20%287_%29.pdf?isAllowed=y&sequence=1

< 1% match (Internet from 30-May-2022)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/7849/161-15-7174%20%2827_%29.pdf?isAllowed=y&sequence=1

< 1% match (Internet from 05-Jul-2024)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/12428/26740.pdf?isAllowed=y&sequence=1

< 1% match (Internet from 13-Apr-2024)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/12015/Dengue%20Fever%20Prediction%20Using%20Machine%20Lear
isAllowed=y&sequence=1