Artificial Intelligence Approach for Modeling House Price Prediction

Melihsah Cekic
Department of Computer Engineering
Istanbul Sabahattin Zaim University
Istanbul, Turkey
cekic.melihsah@std.izu.edu.tr

Kübra Nur Korkmaz
Department of Computer Engineering
Istanbul Sabahattin Zaim University
Istanbul, Turkey
korkmaz.kubra@std.izu.edu.tr

Habib Müküs
Department of Computer Engineering
Istanbul Sabahattin Zaim University
Istanbul, Turkey
mukus.habib@std.izu.edu.tr

Alaa Ali Hameed
Department of Computer Engineering
Istinye University
Istanbul, Turkey
0000-0002-8514-9255

Akhtar Jamil
Department of Computer Science
National University of Computer and Emerging Sciences
Islamabad, Pakistan
0000-0002-2592-1039

Faezeh Soleimani
Department of Mathematical Sciences
Ball State University
Muncie, Indiana, USA
fsoleimani@bsu.edu

2022 2nd International Conference on Computing and Machine Intelligence (ICMI) | 978-1-6654-7483-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICMI55296.2022.9873784

Abstract—Real estate has a vast market volume across the globe. This domain has been growing significantly in the past few decades. An accurate prediction can help buyers, and other decision-makers make better decisions. However, developing a model that can effectively predict house prices in complex environments is still a challenging task. This paper proposes machine learning models for the accurate prediction of real estate house prices. Furthermore, we investigated the feature importance and various data analysis methods to improve the prediction accuracy. Linear Regression, Decision Tree, XGBoost, Extra Trees, and Random Forest were used in this study. For all models, hyperparameters were first calculated using k-fold cross-validation, and then they were trained to apply to test data. The models were tested on the Boston housing dataset. The proposed method was evaluated using Root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) metrics.

Keywords— Convolutional Neural Network real estate price prediction, Convolutional Neural (CNN), machine learning, house price prediction.

I. INTRODUCTION

House price prediction has become a hot research topics in recent years. Due to the exponential increase in demand for houses, there is a need to develop a framework that can facilitate both buyers and sellers in fast decision-making. The end-user is expecting comprehensive information about the houses and due to volatile prices and demand needs, it becomes very difficult for the users to make decisions. Introducing a predictive analytics system into house pricing can help many stakeholders properly plan and make decisions. In fact, the house market volume has become very large, and to mitigate the demand needs, introducing the new method is imperative and can ease the problems.

There have been a number of approaches proposed for predicting house prices based on the literature. Among these, machine learning-based methods have been used very frequently and have shown very satisfactory performance. For instance, [1] focused on the utilization of the Google search engine for real estate price change prediction based on the search results. Machine learning models were used for the predictive potential of Google user search, including Regularized Regression (GLMNET), SVM, kNN, Naive Bayes, CART, C5.0, Bagged CART, and Random Forest. Similarly, authors in [2] employed neural networks to forecast house prices in China's major cities between June 2010–May 2019.

An end-to-end self-attention-based model for predicting house prices is proposed in [3]. The data from public facilities such as schools, parks, etc., were used and exploited the satellite maps to analyze their environments. Different models were then trained for prediction. These models include Gradient Boosting and Light Gradient Boosted machine, deep learning-based and attention-based models. Similarly, a method based on optimized gradient boosting for improving the real estate house price prediction with higher accuracy is proposed in [4].

The authors in [5] used five commonly used machine learning models for Green Building (GB) price prediction. Linear Regression, Decision Tree, Random Forest, Ridge, and Lasso are among these models. Some authors, such as in [6], have used regression techniques for house price prediction.

Machine learning techniques are applied to predict house prices from historical data of property markets in [7]. The data was preprocessed and then fed into the network for classification. Several models were exploited, which showed high prediction accuracy.

Gradient Boosting Model and XGBoost are also commonly used models for various classification and regression tasks. In [8], the authors used these two approaches for house price prediction for Karachi city. It is a highly dense city with a large population residing in it. The results analyzed with these models showed that both are suitable for house price prediction.

An integrated data mining and machine learning model for predicting real estate house prices is proposed in [9]. In addition, it utilized a quadratic, exponential smoothing time-varying algorithm for trend estimation.

Artificial Neural Networks (ANNs) obtained excellent prediction accuracy for real estate price prediction [10]. This paper focused on optimizing the parameters of the model for price prediction in Helsinki, Finland. Different parameters of the model, such as activation functions, weight initialization, number of hidden layers, learning rate, etc., were optimized to obtain high accuracy. The authors follow a similar approach in [11]. In this paper, four machine learning models are proposed to forecast real estate prices: Least Squares Support Vector Regression (LSSVR) as well as Classification and

Regression Trees (CART), General Regression Neural Networks (GRNN), and Backpropagation Neural Networks (BPNN).

A novel machine learning approach for real estate estimation using Call Detail Records (CDR) is proposed in [12]. CDR provides detailed insights into mobility characterization, which can be used for predicting the real estate price. Multi-Layered Perceptron (MLP) trained with Particle Swarm Optimization (PSO) is used for prediction. In [13], an artificial intelligence-based approach is used to forecast real estate auction prices. Similarly, several real estate features have been developed and classified using various machine learning classifiers, including Logistic Regression, Random Forest, Voting Classifier, and XGBoost [14].

This paper investigated the problem of accurately predicting house prices using various machine learning models. The steps required to prepare the data for training the model and then predicting using test data are described in detail. Feature selection was also performed based on feature importance, and various analysis techniques were implemented to get more insights into the data. Experiments with the Boston dataset were performed to evaluate the proposed methods.

The rest of the paper is arranged as follows. Section II describes both the dataset and the summary of each algorithm. In section III, the details of the experiments are presented. Finally, the paper is finished with a conclusion section.

II. MATERIALS AND METHODS

A. Data Set

The data set used in this study was downloaded from the Kaggle website [15]. In this data set, different values of the houses in Boston are processed according to the situations. The data set consists of 14 features and 7084 samples. The attributes of the data set are shown in Table I.

TABLE I. HOUSING DATASET ATTRIBUTES AND DESCRIPTIONS

Feature	Description
CRIM	Crime rate per capita by city.
ZN	Zoned for parcels over 25,000 sq.ft. residential land rate.
INDUS	Ratio of non-retail business per city.
CHAS	Charles River dummy variable (= road river 1 if limiting; otherwise 0).
NOX	Nitrogen oxide concentration (one part 10 million) part).
RM	Average number of rooms per residence.
AGE	Built before 1940, by owner the ratio of units used.
DIS	Distances to five Boston employment centers weighted average.
RAD	Index of accessibility to radial highways.
TAX	Full value property tax rate per $10,000.
PTRATIO	Student-teacher ratio by city.
BLACK	1000 (Bk - 0.63) ^ 2 where Bk, by city is the proportion of blacks.
ISTAT	Lower status of the population (percent).
MEDV	Median of owner-occupied houses value as $1000.

B. Feature Importance

Many attributes are more relevant to determining the prices of houses. Some of these are controllable, measurable factors. It is necessary to know to what extent these attributes affect house prices. In this way, estimation success rates of the regression method may increase if fewer features that can better represent the data set can be determined instead of all the features in the data set. In this study, we used the feature selection method of Correlation-based Feature Subset Evaluation (PPC). All extracted features are ranked through feature importance analysis and then the optimum feature subset is selected after ordering [16]. In the project, quantitative information about each data point is given by "RM", "LSTA", and "DIS". The target variable, "MEDV", will be the variable that helps us predict. Fig. 1 and Fig. 2 show the connection of RM and LSTA data with MEDV data. That indicates which feature is used mostly when estimating the MEDV value.

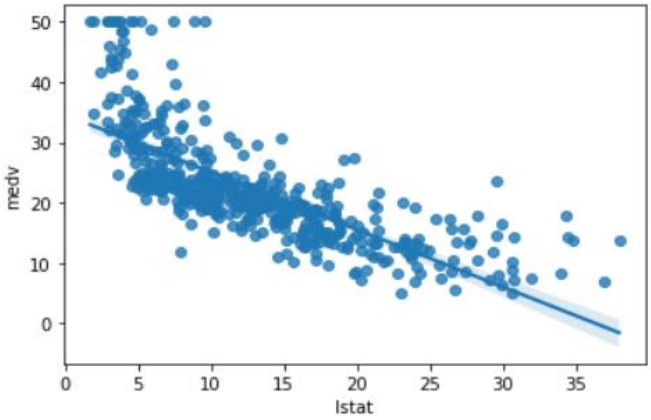

Fig. 4. Interconnection of stat and medv data

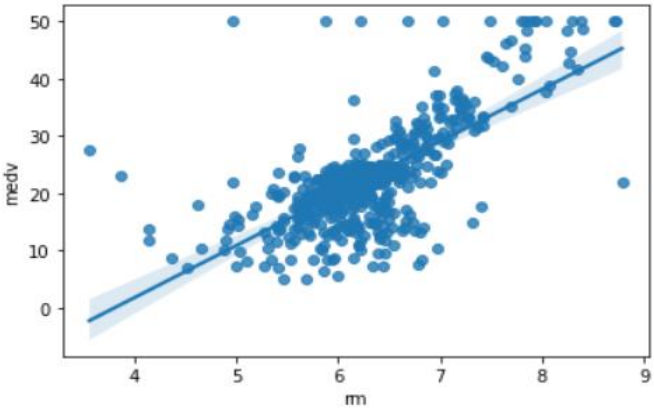

Fig. 5. Interconnection of medv and rm data

C. Data Analysis

Density Estimation in Dataset: Density estimation is to estimate the number range of data in the data set. This post reviews and compares a set of data for density estimation. Histograms are an effective means of visualizing the probability density of a sample of data. In parametric probability density estimation, a sample of data is used to estimate parameters for the density function based on parameters corresponding to a common distribution.

$$P(x) = \int_{-\infty}^{x} p(\xi)d\xi. \quad (1)$$

Correlation Analysis: It is the value that represents the direction and strength of the linear relationship between two variables. A CCA model consists of two representations of the same objects with each representation being projected in such

a way that they have the maximum correlation in reduced dimensions. Normally, CCA calculates projection vectors w_x∈[IR]^d and w_y∈[IR]^k correlation coefficient.

$$\rho = \frac{w_x^T X Y^T w_y}{\sqrt{(w_x^T X X^T w_x)(w_y^T Y Y^T w_y)}} \quad (2)$$

Linear Regression Analysis: To determine the relationship between variables, linear regression is the most common estimation model. Apart from univariate and multivariate types, linear data are also considered. Generally, the linear regression models fall into one of the following two categories: simple linear regression and multiple linear regression. Simple linear regression models deal with one dependent and one independent variables. Multiple linear regression models work with one dependent and more independent variables.

Random Forest Regression Analysis: It is also a supervised learning algorithm. It combines several individual decision trees to form a robust ensemble model. By dividing the original data into smaller subset and using subsets of attributes different models are grown. Finally, the predictions are made by taking the average of predictions obtained from each tree.

Decision Tree Regression Analysis: Based on a tree structure, regression and classification models can be created by Decision Tree models. An associated decision tree is progressively developed by dividing a dataset into smaller and smaller subsets. Finally, a tree with decision nodes and leaf nodes is created. A decision node (e.g., house) has two or more branches (e.g. crim, tax, and nox) that each represent value.

D. Regression Methods

If one of the variables discussed is dependent (y) and the other is independent (x), the relationship expressed as a function of y and x is called regression. In this function, the continuous variable y is calculated for the given x attribute values. Regression is supervised learning. Regression analysis is an analysis method that allows finding the cause-effect relationship between the variables. In this study, we estimated the status of the houses according to the features using Linear Regression, Decision Tree Regressor, and Random Forest Regressor methods.

In a linear regression model, y and x are called the dependent and independent variables. β_0 is the intersection point (the value of the y variable when the x variable takes the value zero), β_1 coefficient (slope of the line), and ε is the random variable that represents the noise that causes the actual values to deviate from the function [16].

$$y = \beta_0 + \beta_1 x + \varepsilon \quad (3)$$

E. Decision Tree Regression

In decision tree algorithms, how the split will take place is one factor that affects the tree's accuracy. The type of target variable is important to select the algorithm. Gini, and Classification Error for categorical variables are the most commonly used algorithms in decision trees. For regression reasons, we will use standard deviation instead of information gain. First, the standard deviation for the target is calculated. The formula in (4) is used in the calculation of the standard deviation with two parameters [17].

$$S = \sqrt{\frac{\sum(x-\mu)^2}{n}} \quad (4)$$

$$S(T,X) = \sum_{c \in X} P(c) S(c) \quad (5)$$

$$SDR(T,X) = S(T) - S(T,X) \quad (6)$$

F. Random Forest Regression

Since the training takes place on different data sets in the random forest model, variance, in other words, overfitting, which is one of the biggest problems of decision trees, is reduced. In addition, we reduce the chance of finding outliers in the sub-datasets we created with the bootstrap method.

$$h(x) = (1/K) \sum_{k=1}^{K} h(x; \theta_k). \quad (7)$$

III. Experimental Results

The experiments were conducted to evaluate the effectiveness of each model for house price prediction. These experiments are described in detail in this section.

The models were evaluated in terms of RMSE, MAE, and R^2. These values are obtained using the following metric [18], [19]:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} \left| Price_{predict} - Price_{true} \right| \quad (8)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(Price_{predict} - Price_{true} \right)^2} \quad (9)$$

$$R^2 = 1 - \frac{\sum \left(Price_{predict} - Price_{true} \right)^2}{\sum \left(Price_{predict} - Price_{true} \right)^2} \quad (10)$$

Where $Price_{predict}$ is the predicted price of the house and $Price_{true}$ is the actual price of the house. To evaluate the performance of the model, the error for each model is derived and then averaged for *n* number of samples.

The dataset was divided into training (70%)and test (30%) subsets. Each model was executed five times and the average values were calculated. These experiments were conducted on an intel i5 8th generation processor with 16 GB RAM. In addition, the model parameters were obtained using K-fold cross-validation and grid search. The obtained optimal values of parameters were then used in all experiments.

Table II summarizes the results obtained for each model when the dataset was used without feature selection. In terms of MAE, linear regression, random forest, decision tree, extra tree, and XGB resulted in 76.59, 98.36, 71.87, 52.83, and 68.25, respectively. Similarly, these classifiers produced RMSE as 56.09, 76.98, 59.93, 40.31, and 56.84, respectively. The coefficient of determination for same models remains 0.87, 0.70, 0.92, 0.96 and 0.93 respectively.

TABLE II. CLASSIFICATION ACCURACY OBTAINED FOR EACH MODEL FOR THE BOSTON HOUSE PRICE DATASET

Algorithms	MAE	RMSE	R^2
Linear Regression	76.59	56.09	0.87
Random Forest Regressor	98.36	76.98	0.70
Decision Tree Regressor	71.87	59.93	0.92
Extra Trees Regressor	52.83	40.31	0.96
XGB Regressor	68.25	56.84	0.93

A further analysis was performed to further improve the prediction accuracy by using feature selection. The feature selection was performed using principal component analysis. Instead of selecting all the available features, we selected half

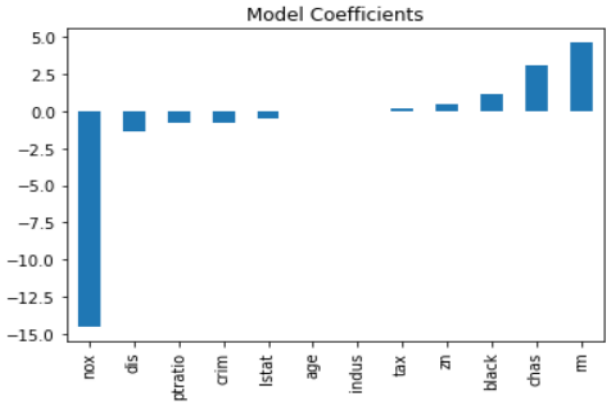

Fig. 3. Feature importance for all features

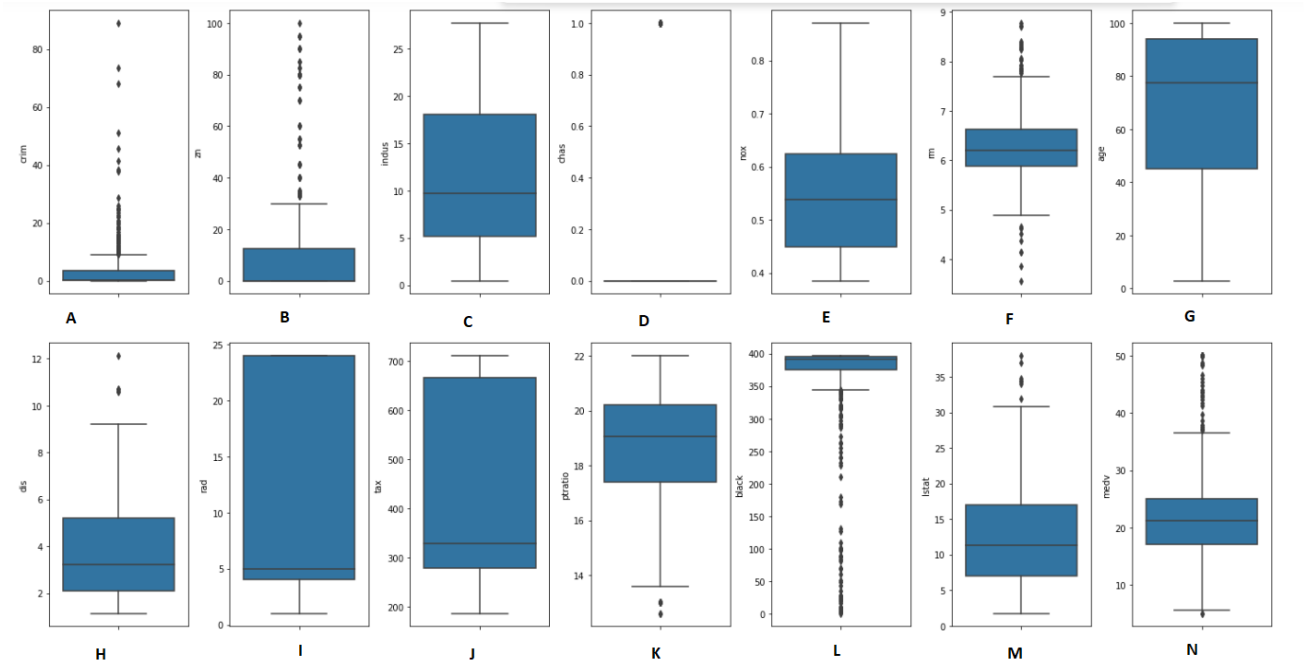

Fig. 4. Distribution of the feature values

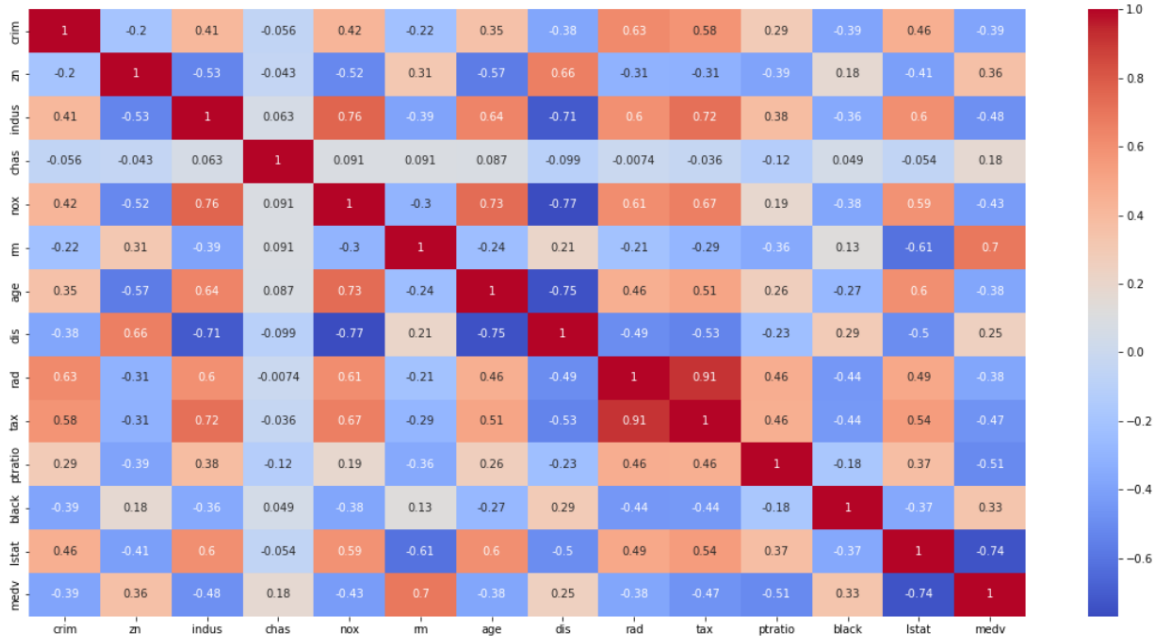

Fig. 5. Correlation matrix for all available features

(7) features and used the same classifiers for classification. The features were selected based on their higher eigenvalues. The time required to train the models with fewer features was less, and the accuracy obtained for different models was high. Fig. 3 shows the feature importance in the form of bilateral relations. The regression coefficient shows the magnitude and direction of the relationship between an indicator and the response variable. The negative bar in the graph indicates little correlation, while positive bars indicate that the features are highly correlated. Fig. 4 and 5 show the distribution of the values and feature correlation matrix, respectively.

Table III summarizes the results obtained after the application of feature selection. In terms of MAE and RMSE, a significant improvement was observed. Overall, an average of 8-12% improvement in accuracy was observed. The linear regression model produced the highest MAE while other models resulted in a similar MAE. Similarly, for the RMSE, all the models produced similar values except linear regression (35.58) and decision tree (39.65). In terms of coefficient of determination, the results were very similar for all the models.

TABLE III. THE RESULTS OBTAINED FOR ALL MODELS AFTER FEATURE SELECTION ARE APPLIED TO THE ORIGINAL DATA

Algorithms	**MAE**	**RMSE**	**R^2**
Linear Regression	23.87	35.58	0.90
Decision Tree Regressor	10.93	39.65	0.91
Random Forest Regressor	10.44	21.75	0.91
Extra Trees Regressor	10.56	19.87	0.90
XGB Regressor	10.22	18.76	0.92

IV. CONCLUSION

This paper focused on house price prediction using machine learning algorithms. Five different models were evaluated: linear regression, decision tree, XGBoost, extra trees, and random forest. The study showed that all machine learning models were effective for house price prediction. Two sets of experiments were conducted separately. In the first experiment, the available features (fourteen) were fed into the models and obtained the results. In the second experiment, a feature selection approach was followed to reduce the number of features. PCA was applied to select highly discriminative features. These features were ordered according to their eigenvalues. The features with high eigenvalues were selected as the best features. These features were then fed into the models again to predict the house prices. The evaluation of the models was performed using three different metrics, which include root mean square error, mean absolute error, and coefficient of determination. All models were very effective for house prediction as the error metrics obtained acceptable values.

The main limitation of this study is that it is developed on historical data. However, the changing condition and price trends must be included to get a more up-to-date prediction. In addition, the method cannot be extended to other regions or cities without considering the environmental factors. In the future, we would like to extend the model to other cities by including global features. Also, it is interesting to exploit both deep learning and machine learning models for house prediction with higher accuracy

REFERENCES

[1] Murat GÖK,Makine Öğrenmesi Yöntemleri İle Akademik Başarının Tahmin Edilmesi,2017

[2] International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7 Issue-6S2, April 2019

[3] A Review on Predicting Student's Performance Using Data Mining Techniques

[4] M. Lenzerini, "Data integration: a theoretical perspective", *Proc. of the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of database systems*, pp. 233-246, 2002

[5] S. Gidaris and N. Komodakis. Object detection via a multi-region & semantic segmentation-aware cnn model. In ICCV, 2015.

[6] Angela Scaringella, Regression Trees And Contingent Valuation, Roma, Italy,2016

[7] R. O. Duda, P. E. Hart, and D. G. Stork. Pattern Classification. Wiley Interscience, New York, 2001.

[8] Partially Supervised Classification of Remote Sensing Images Through SVM-Based Probability Density Estimation, March 2015 //

[9] DC Montgomery, EA Peck, and GG Vining, "Introduction to linear regression analysis", *Wiley Series in Probability and Statistics*, 2015

[10] CNN-based Cascaded Multi-task Learning of High-level Prior and Density Estimation for Crowd Counting

[11] Construction the Model on the Breast Cancer Survival Analysis Use Support Vector Machine, Logistic Regression and Decision Tree

[12] U. bin Mat, N. Buniyamin, P. M. Arsad, R. Kassim, An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention, in: Engineering Education (ICEED), 2013 IEEE 5th Conference on, IEEE, 2013, pp. 126–130.

[13] Predicting popularity of online articles using Random Forest regression, January 2017

[14] Gaf Seber and AJ Lee, "Linear regression analysis", *Wiley Series in Probability and Statistics*, 2012.

[15] Ordinal Regression Methods: Survey and Experimental Study, Pedro Antonio Gutierrez, Senior Member, IEEE, Marıa Perez-Ortiz, Javier Sanchez-Monedero, Francisco Fernandez-Navarro, Member, IEEE, and Cesar Herv as-Mart ınez, Senior Member, IEEE,2015

[16] Roger Bivand, Revisiting the Boston data set,6 April 2017

[17] J.R. Quinlan, C4.5 programs for machine learning, Morgan Kaufmann, San Mateo (1993)

[18] Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modeling, Guillermo F. Martinez, Koray K.Yilmaz, 2009.

[19] A. A. Hameed, B. Karlik, M. S. Salman, Back-propagation algorithm with variable adaptive momentum, Knowledge-Based Systems, Volume 114, 2016, Pages 79-87, 2016