THE REPUBLIC OF TURKEY

BAHÇEŞEHİR UNIVERSITY

FAST, ACCURATE AND APROXIMATE VALUATION OF REAL ESTATE

Master Thesis

İSMAİL SERDAR TAŞ

İSTANBUL, 2012

THE REPUBLIC OF TURKEY

BAHÇEŞEHİR UNIVERSITY

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

COMPUTER ENGINEERING

FAST, ACCURATE AND APROXIMATE VALUATION OF REAL ESTATE PROPERTY

Master Thesis

İSMAİL SERDAR TAŞ

Supervisor: ASSOC. PROF. DR. SELIM NECDET MIMAROĞLU

İSTANBUL, 2012

BAHÇEŞEHİR UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
COMPUTER ENGINEERING

Title of the Master’s Thesis	: Fast, accurate and approximate valuation of real estate property
Name/Last Name of the Student	: İsmail Serdar Taş
Date of the Defense of Thesis	: 25 Dec 2012

The thesis has been approved by the Graduate School of Natural and Applied Sciences.

Assoc. Prof. Dr., Tunç BOZBURA
Graduate School Director

I certify that this thesis meets all the requirements as a thesis for the degree of Master of Science.

Asst. Prof. Dr., Tarkan AYDIN
Program Coordinator

This is to certify that we have read this thesis and we find it fully adequate in scope, quality and content, as a thesis for the degree of Master of Science.

Examining Comittee Members	Signature
Thesis Supervisor Assoc. Prof. Dr. Selim Necdet MİMAROĞLU	--------------------------------
Member Asst. Prof. Dr. Egemen ÖZDEN	--------------------------------
Member Asst. Prof. Dr. Olcay KURŞUN	--------------------------------

ABSTRACT

FAST, ACCURATE AND APPROXIMATE VALUATION OF REAL ESTATE PROPERTY

İsmail Serdar Taş

Computer Engineering
Thesis Supervisor: Assoc. Prof. Dr. Selim Mimaroğlu

Jan 2013, 47 Pages

Target of real restate valuation is to find the market value of the real estate property which is a hard task. In stock markets buyers list their price while seller does the same and market price for the stock is the price of latest transaction. But, in real estate properties transactions do not take place frequently and every real estate property is different than each other. So, there is no simple way of determining the market value of the real estate properties.

On the other hand financial institutions loaning money in return of a mortgage on the real estate properties need to know the market value of the real estate property. So, in Turkey, for the financial institutions the valuation of real estate properties is being done by real estate experts. Valuation of the real estate experts depend on a simplified multiple regression analysis and expert's personal judgement. Assuming that personal judgements are always correct, still this is not the best way to make a bulk valuation.

In this thesis, by combining data mining methods of clustering and classification, we propose a novel, fast, accurate and approximate real estate property valuation method that is robust and intuitive.

Keywords: Price Valuation, Pricing, Real Estate Valuation, Valuation, Real Estate Price Determination

ÖZET

GAYRİMENKULLERİN HIZLI, DOĞRU VE YAKLAŞIK OLARAK DEĞERLENDİRİLMESİ

İsmail Serdar Taş

Bilgisayar Mühendisliği
Tez Danışmanı: Doç. Dr. Selim Mimaroğlu

Ocak 2013, 47 Sayfa

Gayrimenkul değerlemenin hedefi, zor bir görev olan, gayrimenkulün piyasa fiyatını belirlemektir. Hisse senedi piyasasında satıcılar satmak istediği hisse adedi ve satış fiyatını listeler ve alıcılar da aynısını yaparlar ve piyasa fiyatı ise hisse senedinde gerçekleşen son işlem fiyatıdır. Fakat, gayrimenkullerde, alım satım işlemleri sık gerçekleşmez ve her gayrimenkul birbirlerinden farklıdır. Bu yüzden, gayrimenkul piyasa değeri belirlemenin basit bir yolu yoktur.

Diğer yandan ise gayrimenkul ipoteği karşılığında kredi veren finansal kurumların, ipotek ettikleri gayrimenkulün piyasa değerini bilmeleri gerekir. Bu yüzden, Türkiye'de, finansal kurumlar için, gayrimenkul ekspertizini, gayrimenkul eksperleri yaparlar. Gayrimenkul eksperlerinin değerlemeleri, basitleştirilmiş bir çoklu regresyon analizi ve kişisel kanaatlerinine dayanır. Kişisel kanaatlerin hep doğru olduklarını varsaysak bile, bu yol hala gayrimenkulleri toplu olarak değerlendirmede kullanılabilecek en iyi yol değildir.

Bu tez çalışmasında, veri madenciliği yöntemleri olan, kümeleme ve sınıflandırmayı birleştirerek, yeni, hızlı, güçlü ve objektif, doğru ve yaklaşık sonuçlar veren bir gayrimenkul değerlendirme yöntemi öneriyoruz.

Anahtar Kelimeler: Fiyat Belirleme, Fiyatlandırma, Emlak Ekspertizi, Değerleme, Emlak Fiyat Hesaplama

TABLE OF CONTENTS

TABLES viii

FIGURES ix

ABBREVIATIONS x

SYMBOLS xi

1. INTRODUCTION 2

  1.1. REAL ESTATE VALUATION 2

  1.2. AUTOMATED VALUATION MODELS 3

    1.2.1. Hedonic Pricing 4

    1.2.2. Valuation Using Decision Trees 4

    1.2.4. FAVREP 5

  1.3. THESIS OVERVIEW 5

2. RELATED WORK 7

  2.1. HEDONIC PRICING 7

    2.1.1. Important Issues in Hedonic Price Method 8

  2.2. RELATED RESEARCH 9

3. MATERIAL AND METHODS 12

  3.1. INTERQUARTILE RANGE 12

  3.2. WEIGHTED MANHATTAN DISTANCE 12

  3.3. DBSCAN 13

  3.4. C4.5 14

    3.4.1 Pseudocode 15

4. DATA AND METHOD 17

  4.1. DATA 17

    4.1.1. Attributes 17

      4.1.1.1. City 17

4.1.1.2. District..17

4.1.1.3. Neighborhood 17

4.1.1.4. Floor count..17

4.1.1.5. Floor 17

4.1.1.6. Built year...18

4.1.1.7. Bathroom count...18

4.1.1.8. Living room count...18

4.1.1.9. Bedroom count 18

4.1.1.10. Has car park 18

4.1.1.11. Square meters..18

4.1.1.12. Has sea view..18

4.1.1.13. In building complex 19

4.2. METHOD 19

4.2.1. Algorithm...19

4.2.2. Algorithm Explanation..20

4.2.3. Noise Filtering..21

4.2.4. Distance Metric..21

4.2.4.1. Normalization of attribute values......................................22

4.2.4.2. Normalization example..22

4.2.5. Attribute Weights...23

4.2.6. Clustering...23

4.2.7. DBSCAN Input Parameters...25

4.2.7.1. MinPts 25

4.2.7.2. Epsilon..26

4.2.8. Finding the Price Range 26

4.2.9. Decision Tree Learning...27

4.2.10. Assigning Price 27

5. DEMONSTRATION ON TOY DATASET 29

5.1. DATA 29

5.2. APPLYING ALGORITHM STEPS 29

6. EXPERIMENTAL RESULTS 35

6.1. TEST METHOD AND DATA 35

6.2. ERROR PERCENTAGE 36

6.3. METHODS COMPARED TO EACH OTHER 36

6.3.1. FAVREP – Approximate and Accurate Valuation of Real Estate Property (Valuation Using Density Based Clustering) 36

6.3.2. Real Estate Property Valuation Using Hierarchical Clustering & Decision Tree (Using AGNES in FAVREP) 37

6.3.3. Real Estate Property Valuation Using Decision Tree 37

6.3.4. Real Estate Property Valuation Using Hedonic Pricing Method 37

6.4. TEST RESULTS & COMPARISON 37

6.4.1. FAVREP vs. FAVREP with AGNES 39

6.4.2. FAVREP vs Real Estate Property Valuation Using Decision Tree 41

6.4.3. FAVREP vs Hedonic Pricing Method 43

7. DISCUSSION AND CONCLUSION 46

REFERENCES 48

APPENDICES 52

Appendix 1: Toy data set 53

Appendix 2: Distance matrix created on demonstration run 55

TABLES

Table 4.1: Example record set for normalization.. 22

Table 4.2: Normalized values of table 4.1.. 23

Table 4.3: Attribute values used in distance metric... 23

Table 5.1: Normalized values in training set... 31

Table 5.2: Records in same cluster with test set.. 32

Table 5.3: Price ranges.. 32

Table 5.4: Records in the same cluster after price discretization........................ 33

Table 5.5: Records in price range.. 34

Table 6.1: Amount of records in each error level example................................ 38

Table 6.2: Percentage of records in each error level example............................ 38

Table 6.3: Percentage of records to total records.. 38

Table 6.4: Amount of records in each error level, FAVREP vs. AGNES……... 40

Table 6.5: Percentage of records in each error level, FAVREP vs. AGNES...... 41

Table 6.6: Percentage of records to total records, FAVREP vs. AGNES........... 41

Table 6.7: Amount of records in each error level, FAVREP vs. DT................... 42

Table 6.8: Percentage of records in each error level, FAVREP vs. DT.............. 43

Table 6.9: Percentage of records to total records, FAVREP vs. DT.................. 43

Table 6.10: Amount of records in each error level, FAVREP vs. HPM.............. 44

Table 6.11: Percentage of records in each error level, FAVREP vs. HPM........... 45

Table 6.12: Percentage of records to total records, FAVREP vs. AGNES........... 45

FIGURES

Figure 4.1: An example decision tree.. 27
Figure 5.1: Decision tree with discretized training set.. 33
Figure 5.2: Final decision tree.. 34
Figure 6.1: Visual representation of error percentages in test results of FAVREP... 39
Figure 6.2: Visual representation of error percentages in test results of AGNES... 40
Figure 6.3: Visual representation of error percentages in test results of Decision Tree.. 42
Figure 6.4: Visual representation of error percentages in test results of HPM.... 44

ABBREVIATIONS

AGNES	: Agglomerative Nesting
C4.5	: Learning Decision Tree Algorithm of Ross Quinlan
DBDT	: Real Estate Property Valuation Using Density Based Clustering & Decision Tree
DBSCAN	: Density Based Spatial Clustering of Applications with Noise
DIANA	: Divisive Analysis
DT	: Real Estate Property Valuation Using Decision Tree
GLS	: Generalized Least Squares
GIS	: Geographic Information Systems
HCDT	: Real Eestate Property Valuation Using Hierarchical Clustering & Decision Tree
HPM	: Hedonic Pricing Method
ID3	: Iterative Dichotomiser 3
IQR	: Interquartile Range

SYMBOLS

$ddr(p, q)$	: A data object p is directly density reachable from other data object q
C	: Cluster
p	: Data object p
q	: Data object q
D	: Data Set
$D(p, q)$	: Distance of data object p to data object q
Eps	: Epsilon
$N_{Eps}(p)$	: Epsilon neighborhood of a data object or data point p
$Err(p)$	: Error function for data object p
$EvalPrice_p$	: Evaluated price of data object p
$f(x)$	: Function
Max	: Function returning biggest value in set provided
Min	: Function returning smallest value in set provided
$Info(D)$	: Information gain on D
dP	: Marginal Willingness to Pay for Higher Housing Prices
$Minpts$	: Minimum amount of data points in epsilon neighborhood of a data object for the data object to be a core point.
$Norm(A)$	: Normalized value of attribute A
d_n	: Nth Demand
Q_n	: Nth quartile
$OrgPrice_p$	: Original price of data object p
T	: Test
t	: Threshold
Att_1W	: Weight of attribute 1

1. INTRODUCTION

This chapter provides information on real estate property valuation, hedonic pricing method and our novel study of real estate property valuation by using data mining methods of clustering and classification by decision trees.

1.1. REAL ESTATE VALUATION

Process of valuating real estate property can be called real estate appraisal, property valuation or real estate valuation. And, the aim in appraisal of real estate property is to find the market value of the property. As compared to stocks and bonds, real estate transactions occur less frequently and each real estate property contains unique properties. So, a central auction as in the stock markets does not exist for real estate properties. As a result, a method is required to advice on the value of the real estate property.

In almost any country, in mortgage loans, to secure the amount loaned to the client, mortgage offices / banks usually do not loan 100 percent of the value of the property. The idea behind this is to lower the risk they take by assuring the loan amount by something much worthy and when necessary, being able to sell the property for the amount they lend in a short time. While this looks like a simple process which doesn’t require complex calculations as the loan percentage is usually limited by governments, there is an unknown variable in equation which is market value of the real estate property.

During the mortgage loan process, to learn the value of the real estate property, qualified appraisers (real estate experts) are asked to advice on the value of the real estate. Real estate experts usually examine the property, considering price of similar properties, they find a base value. Then by adding or removing the value of different features from / to the base value they decide on the final price. Then, they provide a written report on the value of the real estate property to the mortgage office / bank and to the client. Using the provided value, banks / mortgage offices calculate the maximum amount they shall loan in return of mortgage.

But, what if the appraiser fails to make an accurate estimation in the value of the property? Estimated value can be higher or lower than the market value. While lower valuation will not be a risk, higher valuation will cause a risk for the banks / mortgage offices. High valuation will cause an increase in the amount of mortgage loans but all these loans will be risky loans as the value of the real estate property will be lower than the loan amount. As a result of these, it can simply be said that accuracy is extremely important in valuation of real estate property. Even if valuation of experts is mandatory by the regulations, using an automated valuation model that can accurately evaluate the value of the real estate property is a must to make a comparison.

In this thesis we introduce a novel automated valuation method to valuate real estate properties. To valuate real estate property, we use data mining methods of clustering and decision trees by using the knowledge gathered from real estate properties with known market prices.

1.2. AUTOMATED VALUATION MODELS

Accuracy of real estate property valuation made by real estate experts usually depends on expert’s judgment. Considering an expert can only be expert on real estate properties in a certain area but not for all the real estate properties in the world and personal judgments are not always trustable, this way of the valuation of real estate property is highly vulnerable to mistakes. In the other hand there is no way of making mass appraisal in a short time using this method of making appraisals by real estate experts. As a result, developing of automated valuation models is needed. All the known automation methods for real estate property valuation use multiple regression analysis and geographic information systems. One of the well known methods used for real estate property valuation is hedonic pricing.

1.2.1. Hedonic Pricing

Hedonic pricing method relies on information provided by households when they make their location decisions. Main principle is that, when the demand for land and housing increases, the price of housing increases.

Hedonic pricing model is basically decomposing price of an item in to separate components. So it can be said that basic premise of the hedonic pricing method is that the price of a real estate property is related to its characteristics. While this premise is right, assuming that any modification made on a real estate property has an immediate effect on price of the property is not quite true.

In the other hand, application of hedonic price method requires collecting all the data related to price of real estates. Data should contain environmental characteristics, property characteristics, neighborhood characteristics, accessibility characteristics and of course selling prices and locations of residential properties.

The main problem in hedonic price method is that it assumes that all the individuals have enough resources to select the real estate property having combination of their preferred features, which is not the case in reality.

1.2.2. Valuation Using Decision Trees

Building a decision tree model that works for all the real estate is not possible. Even using the complete dataset which belong a big city will cause random results in most of the cases as the location of the real estate property is not only a multiplier in price but it also changes the importance of attributes of the real estate properties. Different attributes have different weight in total price for different locations. For example, in Amsterdam, The Netherlands, building age does not have a big weight in price while same attribute has a very big importance for Istanbul, Turkey. While in Amsterdam a building of 20 years old can be considered a new building and the price for an apartment in that building will not have a big difference with an apartment in a 2 years old building with same attributes, in Istanbul 20 years old building will be considered as an old building and price difference for apartments in a 2 years old building and 20 years old building with same attributes will be around 40 percent. This can be explained as Istanbul is on the earthquake zone and buildings of some age are more vulnerable to earthquakes. As for another example, while in Boston, USA, bathroom count has its share in total price; while in Istanbul, Turkey, bathroom count have a very little effect in total price of the real estate property. What is expected from a good decision tree algorithm is to detect these distinctions but as the count of attributes and amount of data alongside with class label count gets bigger, chances of randomness in tree generation gets bigger.

1.2.4. FAVREP

We use data mining methods of classification by decision trees and unsupervised classification, clustering, to valuate the real estate property. As described in section 1.2.2., price valuation using decision trees leads to random results when training set is not carefully chosen. Novel method in this thesis targets finding the appropriate training set to be used while building the decision tree model to be used for valuation.

When a real estate property is wanted to be valuated, first, we take all the records in same neighborhood into a record set. We use interquartile range is to eliminate outliers and extreme values. Including the record to be valuated, we make clustering on the dataset to determine most similar real estates with the real estate property to be valuated. Then we build a decision tree model using the records in the same cluster with the real estate property asked to be valuated. While building this decision tree model, we use equal width discretization on price attribute of the records. Using this model, we determine minimum and maximum prices that real estate to be valuated can have. Then, from the records used building the decision tree model, using only the records within determined price range, we build a new decision tree model to valuate the real estate property in accurate and approximate way.

1.3. THESIS OVERVIEW

This study proposes a novel and efficient data mining method for valuation of real estate property in an accurate and approximate way.

Chapter 1 provides preliminaries and non-exhaustive survey of real estate valuation.

Chapter 2 provides information on related work and known automation methods for real estate property valuation.

Chapter 3 named as “Material and Methods” provides information on known methods used in novel study.

Chapter 4 named as “Data and Method” provides data definition and information on novel method.

Chapter 5 named as “Demonstration on Toy Data Set” provides visual information on a run at toy data set.

Chapter 6 named as “Experimental Results” provides test results for the novel study and comparison of the results with other methods.

Concluding remarks, discussions and future work is presented in Chapter 7.

2. RELATED WORK

This section contains information on related work.

All known automation methods for real estate property valuation use multiple regression analysis and geographic information systems. One of the most known methods is hedonic pricing method.

2.1. HEDONIC PRICING

Origin of the word “hedonic” comes from Greek, meaning, “pleasure”. People take pleasure by living in nice places. And the hedonic pricing method relies on information provided by households when they make their location decisions. As the demand for land and housing increases, the price of housing increases.

The price is affected by the structural characteristics (s_1, s_2, s_3, ..., s_n) of the real estate property, locational characteristics (l_1, l_2, l_3, ..., l_x) of the real estate property and environmental characteristics (e_1, e_2, e_3, ..., e_y). After collecting and compiling this data, a function that relates property value to property characteristics is required to be statistically estimated. Then this function will be used to value real estate properties.

$$Price = f(s_1, s_2, s_3, ..., s_n; e_1, e_2, e_3, ..., e_y; l_1, l_2, l_3, ..., l_x) \tag{2.1}$$

Hedonic pricing model is decomposing price of an item to separate components determining the price. So it can be said that basic premise of the hedonic pricing method is that the price of a real estate property is related to its characteristics. Application of HPM requires collecting all the data related to price of real estates. Some of the required information are selling prices and locations of residential properties, property characteristics that affect selling prices, such as size and number of rooms, neighborhood characteristics that affect selling prices, crime rates and quality of schools, accessibility characteristics that affect prices, such as distances to work and shopping centers, and availability of public transportation.

Consider two houses of both same characteristics but in different locations A and B. Location A is very close to a dump yard and air quality is very low because of the smell coming from the dump yard while location B has no such problem. Price for the house in location B will be higher than the price of the house in location A. Demand 1 (d_1) is the demand for the house in location A and demand 2 (d_2) is the demand for the house in location B. The price differential, dP is the marginal willingness to pay (for higher housing prices) for the difference in environmental quality. dP will be lower for the house in location A compared to the house in location B.

Another example of environmental effect in real estate prices can be the lower prices of real estate properties in areas which are very close to the airports. Due to the affect of the noise, pleasure for household gets lower and willingness to pay also gets lower.

2.1.1. Important Issues in Hedonic Price Method

Hedonic price method assumes all individuals have the opportunity and income to select the combination of their preferred features. But, house market is affected by other influences like interest rates, taxes etc.

Hedonic price method requires a high degree of statistical knowledge to implement and interpret. Even with a good expertise on statistics, the quality of the measures has key importance. For example, as for the build quality of a house, proxy measures should be used and this might result in inaccurate coefficient being generated.

Hedonic price method ideally requires a variety of different real estate properties to be available for sale so people can choose the particular house of they desire. But, in reality, all the desired houses might not be available.

This may be the case, large houses with big gardens are only found in green areas with low pollution and small ones without a garden exist in only city center with high pollution. In such cases, it will be impossible to separate out pollution and garden size accurately.

Hedonic price method assumes that market prices for real estate properties adjust immediately to changes in the attributes. But, in reality, it's not like this, especially, in areas where house sales and purchases occur rarely.

If households are not aware of the prices and characteristics of all the properties in the market then it is likely that the prices and the implicit prices they pay for properties with different characteristics will vary from sale to sale.

2.2. RELATED RESEARCH

Researchers often specify hedonic price functions or hedonic models. Most of the researches are not suggesting new pricing methods but using hedonic pricing method to analyze factors affecting the price structure of residential property.

Adair, A., Mcgreal, S., Smyth, A., Cooper, J. & Ryley, T. 2000, examines factors affecting the prices of real estate properties in the Belfast Urban Area. Shortly, this analysis examines effects of impact of accessibility, relative influence of real estate property characteristics and the impact of socio economic factors. Conclusion of the analysis draws attention on the complexity of relationships within an urban area.

Fletcher, M., Gallimore, P. & Mangan, J. 2000, recommends using a wider range of diagnostic statistics in the specification search for a good model, in particular, but not exclusively, those concerned with predictive stability. Data used in the research is sales data of 1600 real estate properties during 1999 – 2000 in the Midlands of United Kingdom. Illustration of this approach in the paper is done by examining out of sample and in sample diagnostic tests of specifications of a hedonic house price model.

Janssen, C. B. & Soderberg, J. Z. 2001, analyses relationship of household income and apartment price with a robust method, comparing the performance of least median of squares and least squares.

To evaluate the effect of market fundamentals on housing price dynamics, Meese, R., & N., Wallace 2003, compare Kalman filter strategy and the traditional two-step procedures. Traditional method used is first estimating a house price index. Then, method is using the estimated index in subsequent structural modeling. And, the Kalman filter strategy used in comparison allows for the simultaneous estimation of the parameters of a dynamic hedonic price model, the price index and the parameters of a structural model for housing prices.

The heteroscedasticity issue in hedonic real estate valuation models is re-examined by Stevenson 2004. Boston data with high dwelling age average is used in this research. Result is, the iterative GLS (Generalized Least Squares) correction specified in terms of age eliminates all heteroscedasticity at both aggregate and disaggregate levels. Previous findings in same subject, heteroscedasticity with respect to the age of dwelling, are supported by this result.

Bin, O. 2004, incorporates geographic information systems (GIS) into account for locational attributes of real estate properties in paper he compares price prediction performance of conventional parametric models and estimating a hedonic price function using a semi-parametric regression.

Bao, H. X. H. & Wan, A. T. K. 2004, uses limited Hong Kong real estate sales data to illustrate usage of the technique of smoothing splines to estimate hedonic housing price models.

Kim, K. & Park, J. 2005, searches for a correlation with real estate property prices and real estate property price change rates by a cluster analysis. The results show that there’s no correlation between the spatial pattern of real estate property price change rates and the real estate property prices.

Filho, C. M. & Bin, O. 2005, introduced an additive nonparametric regression model as a hedonic price function for real estate properties. A local polynomial estimator is used in combination with a back fitting procedure to make the estimation. To show an evidence of the superiority of their nonparametric model, they compared their results to alternative parametric models.

Fan, G., Ong, Z. S. E. & Koh, H. C. 2006, used the decision tree classification on Singapore real estate property sales data to analyze the relationship between the resale prices of Singapore public houses and housing characteristics and it also identifies significant characteristics in predicting resale prices.

Kestens, Y., Theriault, M. & Rosier, F.D. 2006, uses geographically weighted regressions to measure the heterogeneity of implicit prices regarding household income, age, educational attainment, type and the previous tenure status of the buyers.

3. MATERIAL AND METHODS

This chapter contains information on known methods and material used in novel study.

3.1. INTERQUARTILE RANGE

Interquartile range is a measure of statistical dispersion which equals to the difference of upper and lower quartiles. A quartile is a value which represents one fourth of the values in the set. When used for finding outliers given set is divided in to four equal groups using three quartiles. First quartile (Q_1) also named as lower quartile splits the lowest 25 percent of the data. Second quartile (Q_2 or median) splits the data in two and the third quartile (Q_3) also named as highest quartile splits the highest 25 percent of the data.

Interquartile range (IQR) is the difference between third and first quartile. To be able to find out outliers, two fences are calculated using interquartile range.

$$Lower\ Fence = Q_1 - 1.5\ (IQR) \tag{3.1}$$

$$Upper\ Fence = Q_3 + 1.5\ (IQR) \tag{3.2}$$

Values below lower fence and above upper fence are considered as outliers. Demarcation line for outliers is calculated using formula 1.5 (IQR) by John Turkey at 1977 and is being used ever since. To detect extreme values, demarcation line is being calculated as 3 (IQR) and values below lower fence and above upper fence are considered as extreme values.

In this study demarcation line is calculated by using 3 (IQR) for detecting outliers and 6 (IQR) for detecting extreme values.

3.2. WEIGHTED MANHATTAN DISTANCE

To measure the distance between a data item *A* and *B*, absolute value of the difference between each attribute is multiplied by attribute's weight and summation of this operation for all attributes is divided by summation of attribute weights.

Formula for distance metric is as follows:

$$D(A,B) = \frac{Att_1W|Atr_1A-Atr_1B| + Att_2W|Att_2A-Att_2B| + \cdots + Att_nW|Att_nA-Att_nB|}{Att_1W+Att_2W+\cdots+Att_nW} \tag{3.3}$$

Result set of this distance metric is a number between 0 and 1. [0 , 1]

3.3. DBSCAN

DBSCAN is a well-known density-based clustering algorithm published in proceedings of 2nd conference on knowledge discovery and data mining. It is an effective algorithm for discovering clusters of arbitrary shape.

DBSCAN requires epsilon (*eps*) and *minpts* as input. Each data object is labeled by the algorithm as core point, border point or noise.

Epsilon neighborhood (*epsneighborhood*) of a data object or data point *p*, is denoted by $N_{Eps}(p)$, is defined by:

$$N_{Eps}(p) = \{\, q \in D \mid distance(p,q) \leq Eps \,\}$$

A core point is a data object that has more than or equal to minpts data objects in its epsilon neighborhood.

$$p = core\ point\ \rightarrow \left|\, N_{Eps}(p) \,\right| \geq MinPts$$

A border point is a data object which is in the epsilon neighborhood of a core point but having less than *MinPts* data objects in its epsilon neighborhood.

$$p = border\ point\ \rightarrow \{\, p,q \in D \mid \left|\, N_{Eps}(p) \,\right| < MinPts\ \wedge p \in N_{Eps}(q)\}$$

Finally, a noise point is a data object having less than *MinPts* data objects in its epsilon neighborhood and which is not in the epsilon neighborhood of a core point.

$$p = noise\ point\ \rightarrow \{\ p, q\ \in D\ ||\ N_{Eps}(p)\ |\ < MinPts\ \wedge p\ \notin\ N_{Eps}(q)\ \}$$

A data object p is directly density reachable from other data object q if q is a core point and p is in the epsilon neighborhood of q.

$$ddr(p, q)\ \rightarrow \{\ p, q\ \in D\ |\ \left|N_{Eps}(q)\right| \geq MinPts\ \wedge p\ \in\ N_{Eps}(q)\}$$

If there is a chain of data objects $p_1, p_2, \ldots, p_n$, $p_1 = q,\ p_n = p$ such that p_{i+1} is directly density - reachable from p_i than p_1 is density reachable from p_n. This is not a symmetric relation but it is transitive. Although in general this is not a symmetric relation, obviously, it is symmetric for core points. In the original paper this is described as “Two border points of the same cluster *C* are possibly not density reachable from each other because the core point condition might not hold for both of them. However, there must be a core point in *C* from which both border points of *C* are density - reachable.” (Ester, M., Kriegel, H.P., Sander, J. & Xu, X. 1996, pg.3)

Two data objects p and q which are density reachable from a data object o are called density connected. This is a symmetric relation.

DBSCAN defines a cluster as a set of density connected data objects while defining noise as set of data objects in data set which don’t belong to any of its clusters.

$$noise = \{\ p\ \in D\ |\ \forall\ i{:}\ p\ \notin\ C_i\ \}$$

3.4. C4.5

C4.5 is a decision tree generation algorithm which is an extension of ID3 algorithm. Both the algorithms are developed by Ross Quinlan. The algorithm uses the concept of information entropy to build a decision tree from a set of training data. Main reason of why C4.5 is used in this study is the ability of the algorithm to handle both discrete and continuous attributes.

To construct a decision tree from a training set, C4.5 uses a method known as divide and conquer, which was pioneered by Hunt (Hunt, Marin, & Stone, 1966). A short description of the method is described below. See Quinlan (1993) for a more complete treatment.

With given training set T= $\{t_1,\ t_2,\ t_3, \dots t_n\}$ for each node of the tree, C4.5 calculates entrophy gain for each attribute to find the attribute that most effectively splits data. Recursively, attribute with the highest entrophy gain is chosen to make the decision until each node leads to a label.

3.4.1 Pseudocode

```
Check base cases
maxGain = 0
For each attribute $a_i$ in T
{
    Calculate the entropy gain from splitting on $a_i$
    If Gain ($a_i$) > maxGain then
    {
        $a_{best}$ $a_i$
        maxGain = Gain ($a_i$)
    }
}
Create a node n that splits on $a_{best}$
Recurse on the sub lists generated by splitting on $a_{best}$ and add those nodes as children of n
```

For each node a stop criterion is being checked. One reason for stopping is that *T* contains only cases of one class.

C4.5 can handle discrete and continous attributes. So, it has two default tests to determine the best attribute to split. The default tests considered by C4.5 are:

A = ? with one outcome for each value of discrete attribute *A*.

$A \leq t$ with two outcomes, true and false for a continuous attribute A. t is the threshold that maximizes the splitting criterion. To find t, the cases in T are sorted on values of attribute A. Every adjacent pair suggests a potential threshold t = (v_i+ v_{i+1})/2 and a corresponding partition of T. Then, from these potential thresholds, the one that has the best information gain is selected.

Information gain ratio is an information based measure that takes different probabilities of test outcomes into account. Let C be the number of classes and $p(T, n)$ the proportion of cases in T that belong to the nth class. The uncertainty about the class to which a case in T belongs can be expressed as

$$Info(T) = -\sum_{n-1}^{C} p(T,n) \times \log_2(p(T,n)) \tag{3.4}$$

and the corresponding information gained by a test S with k outcomes as

$$Gain(T,S) = Info(T) - \sum_{i-1}^{k} \frac{|T_i|}{|T|} \times Info(T_i) \tag{3.5}$$

The information gained by a test is highly related to the number of outcomes and it is maximal when there is one case in each subset T_i. When the gain ratio of every possible test is determined, the split with maximum gain ratio is selected.

In some situations, every possible test splits T into subsets that have the same class distribution. All tests then have zero gain and C4.5 uses this as an additional stopping criterion.

4. DATA AND METHOD

This chapter contains information on novel method and data used in the method.

4.1. DATA

Data used in experiments in this study has been gathered from listing prices of real estates from several real estate listing web sites. It contains 923 records for real estate properties located in several neighborhoods of Istanbul, Turkey.

4.1.1. Attributes

Following attributes were available in all sources and are used in experiments made for this study during the valuation process.

4.1.1.1. City

All the records in the data set are in Istanbul.

4.1.1.2. District

Records in data set belong to 6 different districts of Istanbul.

4.1.1.3. Neighborhood

Neighborhood is used to determine the records to be used while valuating a real estate property. Records in data set belong to 74 different neighborhoods.

4.1.1.4. Floor count

Floor count is the count of the floors in the building that apartment is in.

4.1.1.5. Floor

Floor is the floor that apartment located in the building.

4.1.1.6. Built year

As mentioned before, building age is important for valuation of real estate properties. So, year that building was built is used as an attribute.

4.1.1.7. Bathroom count

Count of bathrooms in real estate property

4.1.1.8. Living room count

Count of living rooms may vary in relatively bigger real estate properties. So, this attribute is also used.

4.1.1.9. Bedroom count

Count of bedrooms in the real estate property.

4.1.1.10. Has car park

Bit field telling either real estate property has private parking space or not. For valuating real estate properties in an area with most estates having car parking spaces, instead of bit values, count of parking spaces per real estate property can be used as an attribute.

4.1.1.11. Square meters

This field contains area of the real estate property in square meters.

4.1.1.12. Has sea view

Bit field showing either real estate property has sea view or not. As sea view was the only information we could find about the view, forest view, park view etc. could not be used.

4.1.1.13. In building complex

Building complexes offer many services to residents and real estate properties in building complexes have different prices. So bit field “in building complex” is also one of the used attributes.

4.2. METHOD

Considering databases used in real estate valuation are very large, trying to build a decision tree model using all the records in database will require a lot of time building the model and the model built will not be much reliable as the method described in this thesis. Novel method in this thesis first finds the records to be used in valuation then builds a decision tree using these records to valuate the real estate property. As location is one of the most important attribute determining the market value of the real estate property, only records within the same neighborhood with the real estate property are used to valuate the price. So, first step of the algorithm is picking the records in the same neighborhood with the real estate property to be valuated. Then, noise filtering, clustering, finding the price range and building a decision tree model for final valuation is done in given order.

4.2.1. Algorithm

Input: D = $\{r_1, r_2, r_3, \dots, r_n\}$ (set of real estate properties with known market values)
R (real estate property to be valuated)
S (amount of discrete values)
Eps (epsilon for DBSCAN)
Minpts (minpts for DBSCAN)
f (interquartile range multiplier to detect outliers)
Attribute weights

Output: p (valuated price of R)

```
1.  Foreach $r_i \in D$ do
2.      If $r_i.neighborhood \neq R.neighborhood$ then
3.          Remove $r_i$ from $D$
4.  Remove outliers and extreme values from $D$ using Interquartile Range
5.  Add $R$ to $D$
6.  Create distance matrix $DM$ using weighted manhattan distance
7.  Do DBSCAN clustering on $D$ using $DM$
8.  If $R\ is\ noise$ then
9.  {
10.     Find most similar non noise data object $r_s$ to $R$ using $DM$
11.     Set $R.cluserid = r_1.clusterid$
12. }
13. Remove $R$ from $D$
14. Foreach $r_i \in D$ do
15.     If $r_i.clusterid \neq R.clusterid$ then
16.         Remove $r_i$ from $D$
17. $D_{discretized} = D$
18. Apply equal width discretization on price attribute on $D_{discretized}$
19. Build a decision tree model $DT$ using $D_{discretized}$ with C4.5
20. Find min and max prices $R$ can get by classifying $R$ using $DT$
21. Foreach $r_i \in D$ do
22.     If $r_i.price < minprice$ or $r_i.price > maxprice$ then
23.         Remove $r_i$ from $D$
24. Build a decision tree model $DT$ using $D$ with C4.5
25. Find price $p$ by classifying $R$ using $DT$
26. Return $p$
```

4.2.2. Algorithm Explanation

Code in line 1, 2 and 3 removes records from dataset which do not have the same neighborhood with real estate property to be valuated. Line 4 is noise filtering step which is explained at part 4.2 of this chapter. At line 5, real estate to be valuated is added to dataset and at line 6 distance matrix is being created as described in part 4.3 of this chapter. In lines [7, 12], clustering algorithm is being applied. Further explanation about clustering and used method in the algorithm can be found under part 4.4 of this chapter. In line 13 record for real estate property to be valuated is being removed from the dataset. In following lines, records in a different cluster than the real estate property to be valuated are being removed from the dataset. In line 18, discretization is applied on price attribute of the dataset as it is described in section 4.5., before a decision tree is built to find the price range for the real estate property to be valuated. After determining the price range, records having a price which is out of price range is deleted from the data set and a new decision tree model is being built using final dataset to determine the price using actual prices of records in the dataset.

4.2.3. Noise Filtering

Not all the real estate transactions are representing the market value of the real estate. Some transactions are done with higher or lower than the market value. Reason of most of these transactions is urgent need of money. These records should be eliminated from the record set. In this thesis, interquartile range is used to filter outliers and extreme values. Demarcation line is calculated by using 3 (IQR) for detecting outliers and 6 (IQR) for detecting extreme values.

4.2.4. Distance Metric

On clustering, a distance metric is required to measure the differences of two data points. Also for this thesis, a distance metric is required to measure difference of real estate properties from each other. Same attributes might have different weight in total price for different locations. So, requirement for measuring the distance is to have a metric that supports using different weights for different attributes. In this study, weighted manhattan distance which allows attribute weights as input is used.

$$D(A,B) = \frac{Att_1W|Atr_1A - Atr_1B| + Att_2W|Att_2A - Att_2B| + \cdots + Att_nW|Att_nA - Att_nB|}{Att_1W + Att_2W + \cdots + Att_nW} \qquad (4.1)$$

Result set of this distance metric is a number between 0 and 1. [0 , 1]

4.2.4.1. Normalization of attribute values

For the formula to return meaningful results normalization is required for each attribute field. Calculating normalized values for each attribute is done by dividing the difference of attribute value to the maximum value to, difference of maximum value and minimum value.

$$Norm(Att_n A) = \frac{Att_n A - Min(Att_n)}{Max(Att_n) - Min(Att_n)} \tag{4.2}$$

This formula will return a value between 0 and 1. [0 , 1]

4.2.4.2. Normalization example

Table 4.1: Example record set for normalization

Record Id	Attribute 1
1	100
2	90
3	80
4	95
5	120

Using the given example record set in Table 4.1, normalization of the attribute 1 for the record with id 2 will be as follows:

$$Max\ (Att_1) = 120 \tag{4.3}$$

$$Min\ (Att_1) = 80 \tag{4.4}$$

$$Att_1 = 90 \tag{4.5}$$

$$Norm(Att_1) = \frac{Att_1 - Min(Att_1)}{Max(Att_1) - Min(Att_1)} = \frac{90 - 80}{120 - 80} = 0.25 \tag{4.6}$$

When all the records are normalized, values will be as seen in Table 4.2.

Table 4.2: Normalized values of table 4.1.

Record Id	Attribute 1
1	0.5
2	0.25
3	0
4	0.375
5	1

4.2.5. Attribute Weights

Weights of the attributes should be determined and provided by user as they are subject to change with regards to the area as described in introduction. Weights for the database in this study are as seen in table 4.3.

Table 4.3: Attribute values used in distance metric.

Attribute	Weight
Square Meters	0.50
Building Age	0.25
Living Room Count	0.10
Bathroom Count	0.01
Bedroom Count	0.10
Car Park	0.02
Sea View	0.10
City State	0.05
Mid Floor	0.02

4.2.6. Clustering

Clustering, which is also known as unsupervised classification, is grouping similar data objects. So, expected result of clustering of a dataset is to have clusters containing similar data objects to each other and dissimilar to the objects in other clusters. Similarity of data objects is valuated by a distance metric that uses attribute values that define the object to calculate distances.

Requirement in selecting the records to be used in building the decision tree model to valuate real estate is to find similar records in database and only using them in valuation. To improve the accuracy of the valuation by decision tree, records to be used in building the model are first selected by clustering.

As for the most cases, correct clustering of a data set is not obvious. Assuming that data objects in the data set can be visualized as points in space and clusters can be identified by eye, still the identification of clusters will vary from person to person. While this can be easily described as the effect of different similarity thresholds, determining the threshold in a dataset with multi dimensions is a problem. Most of the clustering algorithms require amount of clusters to be pre-defined which is useful for some cases but not for the use of this thesis. There are thousands of clustering methods and most of them can be examined in three headlines. Partitioning methods, hierarchical methods and density based methods.

Partitioning clustering methods divides the dataset in to non-overlapping clusters such that each data object is assigned to one cluster. Usually, number of clusters is specified by the user as a parameter. *k-means* algorithm can be given as an example to this method. In k-means algorithm, initially, k data objects are randomly selected as cluster centers and each data object is assigned to the closest clusters by calculating their distances to cluster centers. Then, cluster centers are calculated using data points in clusters. After finding new cluster centers, data points are assigned to clusters again using the distance to cluster centers. This step is repeated until cluster centers do not change.

Hierarchical clustering methods can be examined in two groups: agglomerative and divisive. Hierarchical clustering methods construct nested clusters which can be represented by a dendogram. To get a meaningful clustering, dendogram must be cut in a certain level.

Divisive hierarchical clustering algorithms start with one cluster containing all the data objects in data set and split it until all the data objects are in one cluster containing only themselves.

Agglomerative hierarchical clustering algorithms do the opposite. Initially each data object is assigned to a cluster that only contains itself. Then merges clusters at each step by calculating the similarity. Most known hierarchical clustering algorithms are AGNES and DIANA.

In density based clustering methods, cluster is defined as high density region separated from low density regions. And clusters consist of regions with a density above a given threshold.

After running multiple tests and comparing results, popular density based clustering algorithm DBSCAN was found the most suitable algorithm for the test dataset used in this thesis.

Other algorithms tested are: k-medoids, k-means, DIANA and AGNES.

DBSCAN is building clusters by grouping data points having a distance to each other below a threshold. Data points which are not close to any other data point within a cluster are considered as noise. In case, the record to be valuated is a noise, DBSCAN will find this out so it will be easy to understand that there is not sufficient amount of similar records in the database to correctly valuate that record. Still, algorithm should keep on running and make the valuation. So, record to be valuated is considered in the cluster within the closest cluster which is simply cluster of closest data point.

4.2.7. DBSCAN Input Parameters

DBSCAN requires two input parameters *MinPts* and *Epsilon*.

4.2.7.1. MinPts

Parameter *MinPts* is in a way, the minimum count of records allowed in a cluster. As a decision tree model is being built using the records in the cluster, minimum of 2 records are required in clusters. So *MinPts* parameter is set to 2. When working with larger databases a bigger value can be used but it should always be remembered that if there are sets of similar objects with member counts less than *MinPts* than all the members of this cluster will be considered as noise. So, *MinPts* should be carefully chosen.

4.2.7.2. Epsilon

Epsilon can be described as the biggest distance value allowed for two data items to be in an eps neighborhood. As this value gets bigger, amount of items in clusters get bigger while count of clusters gets smaller. In the other hand, when epsilon is very small, amount of discovered noises will get higher. So, when determining *Epsilon* value data should be examined and an *Epsilon* value which results having meaningful clusters should be chosen.

After building a couple of distance matrixes and examining the values in them, for the database used in this study, 0.08 is used as the value for the *Epsilon* parameter in DBSCAN.

4.2.8. Finding the Price Range

The step after finding similar records with the real estate to be valuated is to find in which price range is the real estate.

Dataset used in this step contains only the records within the same cluster with real estate property to be valuated. Equal width discretization is used in price attribute of the dataset. In discretization, difference of maximum and minimum price in dataset is divided by the input parameter, which is number of price ranges. Then the result is used to determine the included range of prices. For determining the first range, value is added to the minimum price. For the next ranges, value is added to maximum amount of each price range. For example, let’s assume that we have a dataset of 10 data points with prices [1 , 10]. When we want to have two equal parts for this data set,

$$(Max(D) - Min(D))/2 = 4.5 \tag{4.7}$$

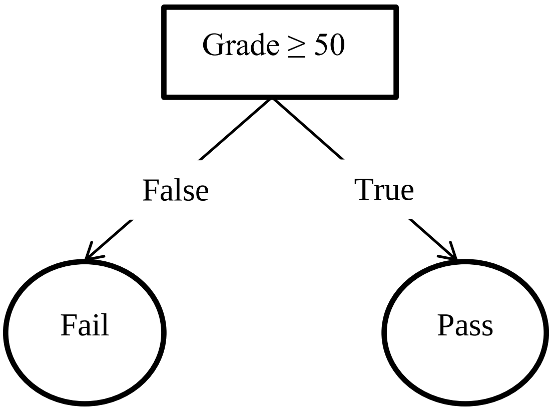
and first part will include values [1 , 5.5] while second part will include (5.5 , 10]. In this study, discretization is used to group the prices in eight equal ranges.

After making the discretization, dataset with the discretized prices is used to build a decision tree model using J. R. Quinlan's C4.5 algorithm. Then, using the model on real estate to be valuated, minimum and maximum prices real estate can have are determined.

4.2.9. Decision Tree Learning

Decision tree learning is a commonly used method in data mining. Goal in decision tree learning is to create a decision tree model that predicts the value of a target label.

Figure 4.1: An example decision tree



Decision tree learning algorithm used in this thesis is Ross Quinlan's C4.5 (1996) algorithm. Alongside with the accuracy of the trees built by C4.5, ability of handling continuous values was main reasons why it is selected.

4.2.10. Assigning Price

We also use decision trees for price assignation. Minimum and maximum values that the real estate property can get are calculated as it is described in section 4.5. Record set which we use as training set for price assignation is records which are in same cluster with real estate property to be valuated and having price between this price ranges. C4.5

is used for building the decision tree model. Then using this model, evaluated price for real estate property is determined.

5. DEMONSTRATION ON TOY DATASET

5.1. DATA

Toy data set shown in appendix 2 in appendices section, having 50 records gathered from real estate listing web sites is used in example run. Record to be valuated using the toy data set has following attributes.

RecordId: 0
Neighborhood: Anadolu Hisarı
Square Meters: 140
Built Year: 1998
Floor Count: 5
Floor: 3
Living Room Count: 1
Bedroom Count: 3
Bathroom Count: 1
Carpark: True
Sea View: False
Building Complex: True
Market Value: 360000

5.2. APPLYING ALGORITHM STEPS

First step of the algorithm is to determine the records within the same neighborhood with the real estate property to be evluated. Records with Id'sstarting from 1 up to including 24 are in same neighborhood with the real estate property to be valuated. So training set is {1, …, 24}.

Second step is applying interquartile range to remove outliers and extreme values from the training set. To apply interquartile range, training set is first ordered by market value. Then Q_1 and $Q_3$are calculated.

$$Q_1 = 350000 \tag{5.1}$$

$$Q_3 = 485000 \tag{5.2}$$

$$IQR = Q_3 - Q_1 = 135000 \tag{5.3}$$

$$Lower\ Fence = Q_1 - 3\ (IQR) = -55000 \tag{5.4}$$

$$Upper\ Fence = Q_3 + 3\ (IQR) = 890000 \tag{5.5}$$

So the records 1 and 21 having market value bigger than upper fence are outliers and will not be used in the training set.

Next step is to generate distance matrix using the training set and the test set. To build the distance matrix, first, normalization on the training set is required. Formula used in normalization is as follows.

$$Norm(Att_nA) = \frac{Att_nA - \ Min(Att_n)}{Max(Att_n) - \ Min(Att_n)} \tag{5.6}$$

Maximum room count in the training set is 5 while the minimum is 2. As the room count of record in the test set (id 0) is 3, applying the formula above will show following result.

$$Norm(RoomCount\ 0) = \frac{3 - \ 2}{5 - \ 2} = 0.33 \tag{5.7}$$

Normalized values in training set including the test (record with id 0) set will look as follows.

Table 5.1: Normalized values in training set

Id	Room Count	Bathroom Count	Building Age	Mid Floor	Car park	Sea View	Living Room Count	Square Meters	Building Complex
0	0.33	0	0.42	1	1	0	1	0.27	1
2	0.33	0	0.58	1	1	0	1	0.25	1
3	0.67	1	0.42	0	1	0	1	0.61	1
4	0.67	1	0.42	0	1	0	1	0.61	1
5	0.33	1	0.58	0	1	0	1	0.25	1
6	0.33	1	0.45	0	1	0	1	0.18	1
7	0.33	1	0.23	1	1	0	1	0.18	1
8	0.33	1	0.23	1	1	0	1	0.18	1
9	0.00	0	0.23	1	1	0	1	0.14	1
10	0.33	1	0.00	1	0	0	1	0.25	0
11	0.33	0	0.00	1	0	0	1	0.09	0
12	0.33	0	0.23	1	0	1	1	0.05	0
13	0.33	0	0.23	0	0	1	1	0.05	0
14	0.00	0	0.87	1	0	0	1	0.09	0
15	0.33	1	0.00	1	0	0	1	0.32	0
16	0.67	1	0.55	0	1	1	1	1.00	1
17	0.33	1	0.23	1	0	1	1	0.30	1
18	0.67	1	0.23	1	1	1	1	0.57	0
19	0.33	0	0.00	1	1	0	1	0.25	1
20	0.33	0	0.23	1	1	0	1	0.25	1
22	0.00	0	1.00	0	0	1	1	0.18	0
23	0.00	0	1.00	0	0	1	1	0.00	0
24	1.00	1	1.00	0	0	0	1	1.00	0

Next step is to build the distance matrix using weighted manhattan distance metric. Formula for the distance metric is as follows.

$$D(A,B) = \frac{Att_1W|Atr_1A-Atr_1B|+ Att_2W|Att_2A-Att_2B|+\cdots+ Att_nW|Att_nA-Att_nB|}{Att_1W+Att_2W+\cdots+Att_nW} \qquad (5.8)$$

The distance of record with id 0 and record with id 2 is calculated using the formula as 0.043478

$$D(0,2) = \frac{0.25|0.4194-0.5806| + 0.5|0.2727-0.25|}{1.15} = 0.04491 \qquad (5.9)$$

Distance matrix generated from training set is on appendix 3 in appendices.

Next step is to generate clusters using the distance matrix. Epsilon value is set to 0.08 for DBSCAN so record in test set (record id 0) is in cluster containing records with ids {0, 2, 5, 6, 7, 8, 9, 10, 11, 15, 19, 20}.

Table 5.2: Records in same cluster with test set

Id	R. Count	Bath room Count	B. Age	Floor	Floor Count	Car Park	Sea View	L. Room Count	m²	Building Complex	Market Value
0	3	1	13	3	5	1	0	1	140	1	?
2	3	1	18	2	4	1	0	1	135	1	360000
5	3	2	18	4	4	1	0	1	135	1	360000
6	3	2	14	5	5	1	0	1	120	1	350000
7	3	2	7	1	6	1	0	1	120	1	235000
8	3	2	7	1	6	1	0	1	120	1	235000
9	2	1	7	1	7	1	0	1	110	1	225000
10	3	2	0	1	5	0	0	1	135	0	425000
11	3	1	0	3	5	0	0	1	100	0	400000
15	3	2	0	4	5	0	0	1	150	0	500000
19	3	1	0	4	5	1	0	1	135	1	365000
20	3	1	7	4	5	1	0	1	135	1	385000

Next step is applying discretization on market value of the records within the same cluster with real estate to be valuated. Difference of maximum and minimum market values in the cluster is 275000. To create 8 equal width price ranges, each range should be 34375.

Table 5.3: Price ranges

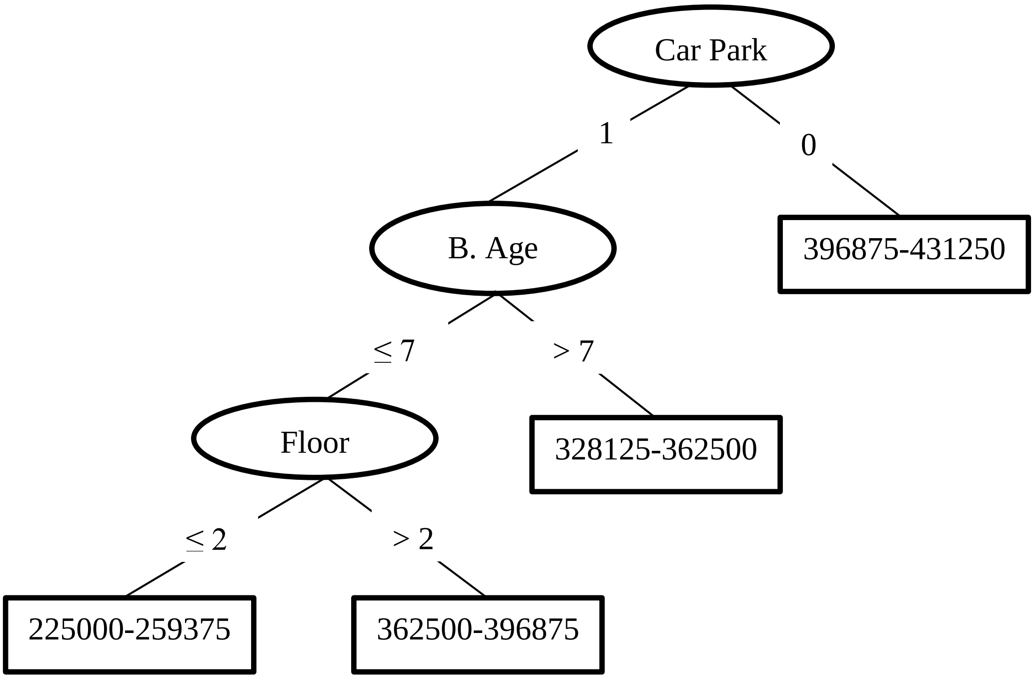
	Min	Max
1	225000	259375
2	259375	293750
3	293750	328125
4	328125	362500
5	362500	396875
6	396875	431250
7	431250	465625
8	465625	500000

Table 5.4: Records in the same cluster after price discretization

Id	Room C.	Bath room C.	B. Age	Floor	Floor C.	Car Park	Sea View	L. Room C.	Square Meters	B. Complex	Price Range
2	3	1	18	2	4	1	0	1	135	1	328125-362500
5	3	2	18	4	4	1	0	1	135	1	328125-362500
6	3	2	14	5	5	1	0	1	120	1	328125-362500
7	3	2	7	1	6	1	0	1	120	1	225000-259375
8	3	2	7	1	6	1	0	1	120	1	225000-259375
9	2	1	7	1	7	1	0	1	110	1	225000-259375
10	3	2	0	1	5	0	0	1	135	0	396875-431250
11	3	1	0	3	5	0	0	1	100	0	396875-431250
15	3	2	0	4	5	0	0	1	150	0	465625-500000
19	3	1	0	4	5	1	0	1	135	1	362500-396875
20	3	1	7	4	5	1	0	1	135	1	362500-396875

Next step is to build a decision tree using discretized records. C4.5. algorithm is used to generate decision trees.

Figure 5.1: Decision tree with discretized training set



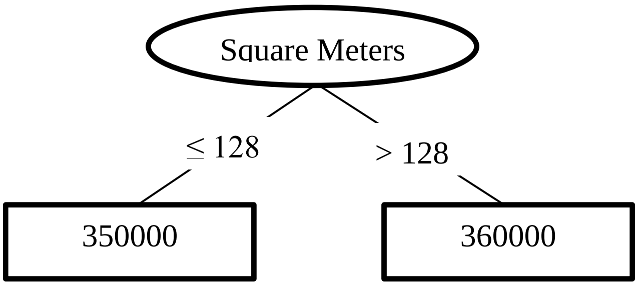
Using the decision tree created with records in same cluster with real estate to be valuated after price discretization, price range for the test set is calculated as 328125 – 362500. To make the final valuation, only records in this price range are used as training set to generate another decision tree.

Table 5.5: Records in price range

Id	Room Count	Bath room Count	B. Age	Floor	Floor Count	Car Park	Sea View	Living Room Count	Square Meters	Building Complex	Market Value
2	3	1	18	2	4	1	0	1	135	TRUE	360000
5	3	2	18	4	4	1	0	1	135	TRUE	360000
6	3	2	14	5	5	1	0	1	120	TRUE	350000

When final decision tree is generated using C4.5 with the records in Table 5.5., tree has only one leaf which is 360000. So the assigned price for the real estate in test set is 360000.

Figure 5.2: Final decision tree



6. EXPERIMENTAL RESULTS

In this chapter, results of test runs with FAVREP and comparison of the results with results from other methods can be found.

6.1. TEST METHOD AND DATA

Data used in the comparison tests are Istanbul real estate sales listing data collected from various web sites on year 2011 for several different neighborhoods. A software has been developed to gather the listing data and gathered data is randomly selected by this software. Although the dataset has around a thousand records, only 645 of them were chosen to be used in the valuation tests. Reason of this is, to be able to make a fair comparison between different methods. A neighborhood with only two real estate property records located in will result almost perfect with hedonic pricing method as the hedonic pricing function will be created using these two records in the neighborhood and results of the function for the neighborhood will be perfectly matching to these two records. On the other hand, when valuating using a decision tree, one of these two records will be removed from the training set and the other record will be the only record left to build a decision tree. Building a decision tree with only one record in training set is not meaningful and record in the test set will always have the label value of the record in the training set. To eliminate these meaningless valuations and have objective results, a limit of minimum allowed count of records in neighborhood is set to include the data of the neighborhood to tests. Limit is set to eight so all the neighborhoods having less than eight records within are kept out of the test. For all the neighborhoods having more than or equal to 8 records in the dataset, a different hedonic function is generated for each neighborhood without removing any records. As for other methods used in testing, valuation is made by removing the current record from dataset and using rest of the database for valuation. So, all of the eligible records in the dataset are used in testing with all the methods.

6.2. ERROR PERCENTAGE

Error percentage in valuation result is calculated by dividing the absolute value of the difference between original price and valuated price to original price.

$$Err(a) = \frac{|EvalPrice_a - OrgPrice_a|}{OrgPrice_a} \tag{6.1}$$

For example, a real estate property originally priced 100,000 TL has been valuated to be 105,000 TL. The error percentage for this valuation is:

$$Err(a) = \frac{|105{,}000 - 100000|}{100000} = 5\% \tag{6.2}$$

To be able to show the results in the table, 11 different levels of errors are created to group the results in. From 0 percent to 50 percent, each error level has been increased to represent 5 percent of the error. First level of error contains records which have been valuated with an error from including 0 percent to 5 percent. Second level of error contains records which have been valuated with an error from 5 percent to 10 percent inclusive. The less the error is, the better the valuation is.

6.3. METHODS COMPARED TO EACH OTHER

6.3.1. FAVREP – Approximate and Accurate Valuation of Real Estate Property (Valuation Using Density Based Clustering)

In novel method, real estate property valuation using density based clustering and decision tree, for the decision tree algorithm, C4.5 is used. For the density based clustering algorithm, DBSCAN is used. As DBSCAN does also detect noises, with regards to the data in the database, it is possible to have the real estate property to be valuated marked as noise. For those records marked as noise, yet to be valuated, they have been assumed as they belong to cluster of nearest data object. This way, to be able to compare results to other methods, regardless of the distance of the record to the dataset, all the records has been valuated.

6.3.2. Real Estate Property Valuation Using Hierarchical Clustering & Decision Tree (Using AGNES in FAVREP)

Within this method, for the decision tree algorithm, C4.5 is used. For the hierarchical clustering algorithm, Agglomerative Nesting (AGNES) is used. As most of the other hierarchical clustering algorithms, AGNES algorithm requires a way to stop. In the tests, cluster count is used to provide a stop point for the algorithm. It is asked to run until amount of generated clusters reach to eight clusters. As the algorithm can not detect noises, some of the clusters had only one record. In other words, noises were considered in their own clusters. When real estate property to be valuated is considered in its own cluster, it has been assumed that it belongs to the cluster of nearest data object. This way all the records have been valuated.

6.3.3. Real Estate Property Valuation Using Decision Tree

Another method tested is to get all the records within the same neighborhood with the real estate property to be valuated and building a decision tree model using C4.5 algorithm. At first step a decision tree model is being built using all the records within same neighborhood with the real estate property to be valuated (excluding the record to be valuated). And in the second step, generated decision tree model is used for valuating the record. By following same steps once for each record, results have been generated.

6.3.4. Real Estate Property Valuation Using Hedonic Pricing Method

Another method tested is hedonic pricing method. A general hedonic pricing function is created to be used in whole database. To get more accurate results, for each neighborhood, effect of the land price is calculated separately and used in the hedonic pricing function.

6.4. TEST RESULTS & COMPARISON

In all the tables showing test results, method Real Estate Property Valuation Using Density Based Clustering & Decision Tree described in part 5.3.1. of this chapter which is also the novel method in this study is named as “FAVREP”. Method, Real Estate Property Valuation Using Hierarchical Clustering & Decision Tree, also can be named as using AGNES as clusterng algorithm in FAVREP, described in part 5.3.2. of this chapter is named as “AGNES” and method Real Estate Property Valuation Using Decision Tree described in part 5.3.3. of this chapter is named as “DT” while no abbreviation is used for real estate valuation using hedonic pricing method and “Hedonic” is used as the name.

Three tables and one figure are generated for each comparison. The figures contain visual representation of amount of records in error percentages. First table contains comparison of amount of records in each error level.

Table 6.1: Amount of records in each error level example

	[0 - 5]
X	*n*
Y	*p*

In the example on Table 6.1., method X had *n* records in the test result with an error percentage smaller than or equal to five, while method Y had *p*.

Second table contains comparison of percentage of records in each error level.

Table 6.2: Percentage of records in each error level example

	[0 - 5]
X	0.1234
Y	0.0987

In the example on table 6.2., 12.34 percent of the records in dataset had an error percentage smaller than or equal to five in the test results of method X, while 9.87 percent was the percentage for method Y.

Third table contains comparison of percentage of records to total records.

Table 6.3: Percentage of records to total records

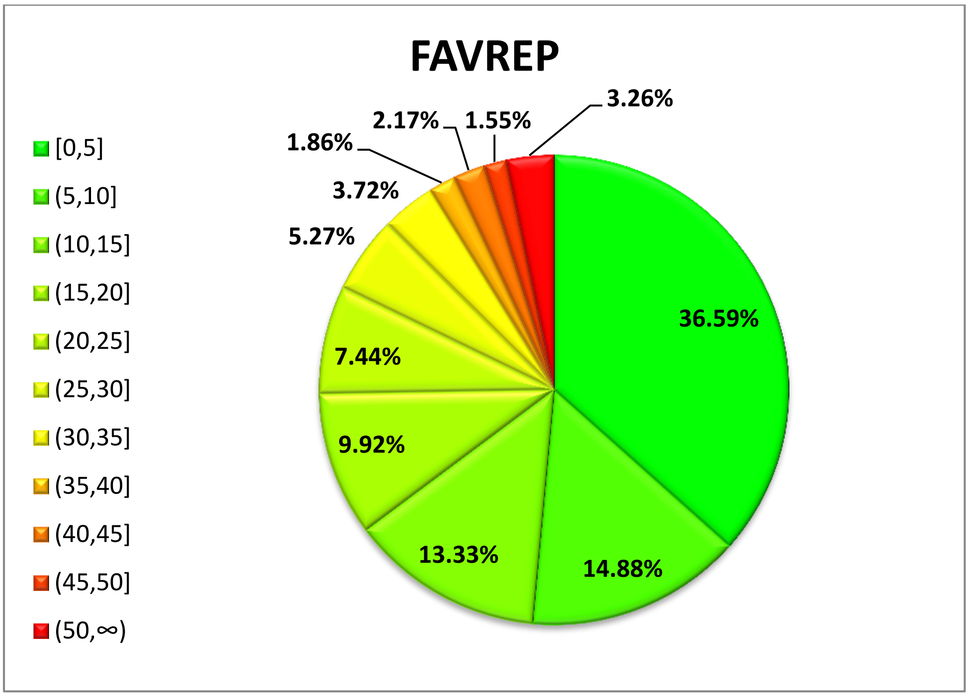
	≤5%	**≤10%**
X	0.1234	0.2345
Y	0.0987	0.1516

In the example on table 6.3., 12.34 percent of the records in dataset had an error percentage smaller than or equal to five and 23.45 percent of the records had an error percentage smaller than or equal to 10 percent in the test results of method X, while 9.87 and 15.16 were the percentages for method Y.

6.4.1. FAVREP vs. FAVREP with AGNES

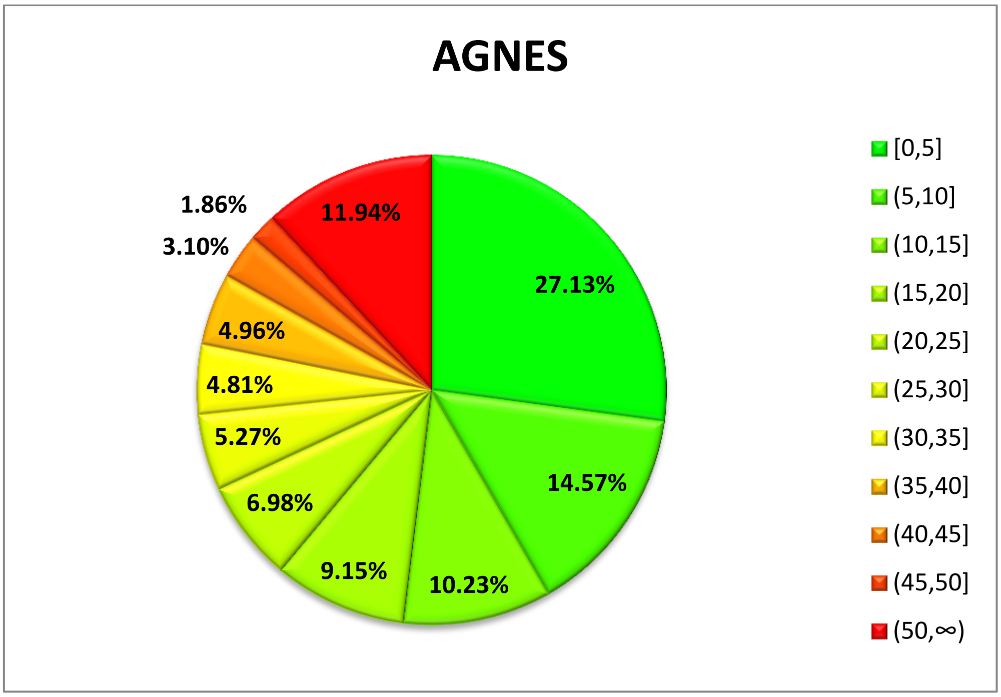
When test results from FAVREP and hierarchical clustering used version of FAVREP are compared, it's seen that using a hierarchical clustering algorithm for building clusters genereates results with bigger error percentages.

Figure 6.1: Visual representation of error percentages in test results of FAVREP



As seen in Figure 6.1, 82.2 percent of the records used in test were valuated with a difference of maximum 25 percent to the original listing price. In the other hand, 3.26 percent of the records were valuated with a difference bigger than 50 percent to the original listing price.

Figure 6.2: Visual representation of error percentages in test results of AGNES



As seen in Figure 6.2, 68.1 percent of the records used in test were valuated with a difference of maximum 25 percent to the original listing price. In the other hand, 11.94 percent of the records were valuated with a difference bigger than 50 percent to the original listing price.

Table 6.4: Amount of records in each error level, FAVREP vs. AGNES

	[0 - 5]	(5 - 10]	(10 - 15]	(15 - 20]	(20 - 25]	(25 - 30]
AGNES	175	94	66	59	45	34
FAVREP	236	96	86	64	48	34

	(30 - 35]	(35 - 40]	(40 - 45]	(45 - 50]	(50 - ∞)
AGNES	31	32	20	12	77
FAVREP	24	12	14	10	21

In Table 6.4, 11 buckets, representing the difference of the valuation result to original listing price, starting from [0,5] ending with (50, ∞) are used to show the count of records in each bucket. As seen in Table 6.4, FAVREP valuated 236 records with no difference or a difference which is smaller than or equal to 5 percent of the original listing price.

Table 6.5: Percentage of records in each error level, FAVREP vs. AGNES

	[0 - 5]	(5 - 10]	(10 - 15]	(15 - 20]	(20 - 25]	(25 - 30]	(30 - 35]	(35 - 40]	(40 - 45]	(45 - 50]	(50 - ∞)
AGNES	0.2713	0.1457	0.1023	0.0915	0.0698	0.0527	0.0481	0.0496	0.0310	0.0186	0.1194
FAVREP	0.3659	0.1488	0.1333	0.0992	0.0744	0.0527	0.0372	0.0186	0.0217	0.0155	0.0326

In Table 6.5, 11 buckets, representing the difference of the valuation result to original listing price, starting from [0,5] ending with (50, ∞) are used to show the percentage of records to dataset in each bucket. As seen in Table 6.5, FAVREP valuated 36.59 percent of all records with no difference or a difference which is smaller than or equal to 5 percent of the original listing price.

Table 6.6: Percentage of records to total records, FAVREP vs. AGNES

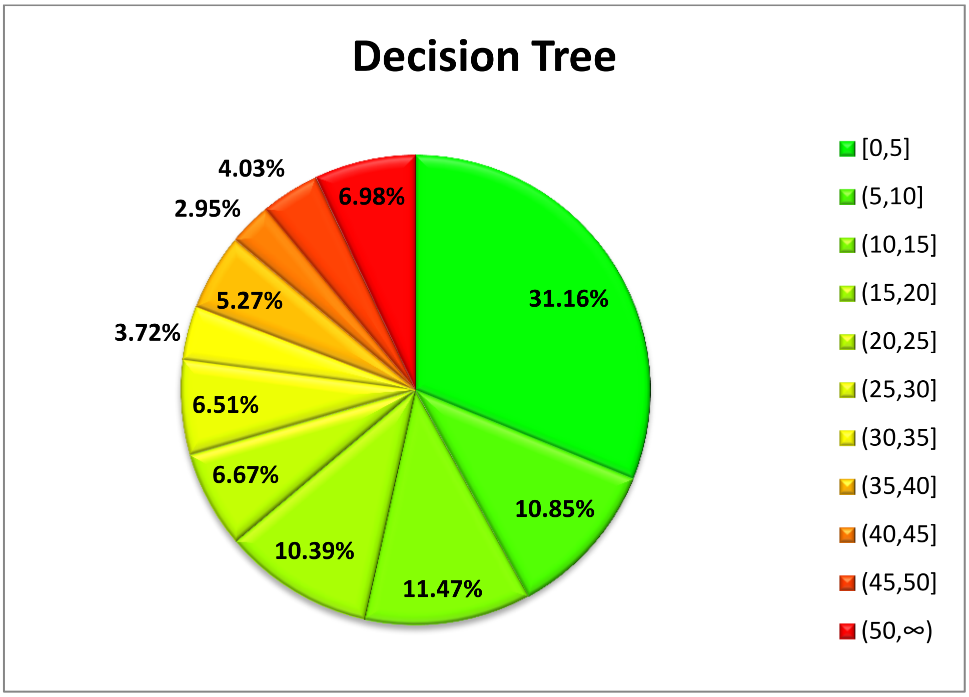
	≤5%	≤10%	≤15%	≤20%	≤25%	≤30%	≤35%	≤40%	≤45%	≤50%	>50%
AGNES	0.271	0.417	0.519	0.611	0.681	0.733	0.781	0.831	0.862	0.881	0.119
FAVREP	0.366	0.515	0.648	0.747	0.822	0.874	0.912	0.93	0.952	0.967	0.033

In Table 6.6, percentage of count of records to amount of records in dataset is shown in 11 different values. As seen in Table 6.6, FAVREP valuated 82.2 percent of all records with no difference or a difference which is smaller than or equal to 25 percent of the original listing price while valuating 3.3 percent of the records with a difference bigger than 50 percent of the original listing price.

6.4.2. FAVREP vs Real Estate Property Valuation Using Decision Tree

When test results from FAVREP and a version of FAVREP which does not have the clustering step are compared, it’s seen that not using the clustering step genereates very close yet not better results with FAVREP.

Figure 6.3: Visual representation of error percentages in test results of DT



As seen in Figure 6.3, 70.54 percent of the records used in test were valuated with a difference of maximum 25 percent to the original listing price. In the other hand, 6.98 percent of the records were valuated with a difference bigger than 50 percent to the original listing price.

Table 6.7: Amount of records in each error level, FAVREP vs. DT

	[0 - 5]	(5 - 10]	(10 - 15]	(15 - 20]	(20 - 25]	(25 - 30]
FAVREP	236	96	86	64	48	34
DT	201	70	74	67	43	42
	(30 - 35]	**(35 - 40]**	**(40 - 45]**	**(45 - 50]**	**(50 - ∞)**	
FAVREP	24	12	14	10	21	
DT	24	34	19	26	45	

In Table 6.7, 11 buckets, representing the difference of the valuation result to original listing price, starting from [0,5] ending with (50, ∞) are used to show the count of records in each bucket. As seen in Table 6.7, DT valuated 201 records with no difference or a difference which is smaller than or equal to 5 percent of the original listing price.

Table 6.8: Percentage of records in each error level, FAVREP vs. DT

	[0 - 5]	(5 - 10]	(10 - 15]	(15 - 20]	(20 - 25]	(25 - 30]
FAVREP	0.3659	0.1488	0.1333	0.0992	0.0744	0.0527
DT	0.3116	0.1085	0.1147	0.1039	0.0667	0.0651
	(30 - 35]	**(35 - 40]**	**(40 - 45]**	**(45 - 50]**	**(50 - ∞)**	
FAVREP	0.0372	0.0186	0.0217	0.0155	0.0326	
DT	0.0372	0.0527	0.0295	0.0403	0.0698	

In Table 6.8, 11 buckets, representing the difference of the valuation result to original listing price, starting from [0,5] ending with (50, ∞) are used to show the percentage of records to dataset in each bucket. As seen in Table 6.8, DT valuated 31.16 percent of all records with no difference or a difference which is smaller than or equal to 5 percent of the original listing price.

Table 6.9: Percentage of records to total records, FAVREP vs. DT

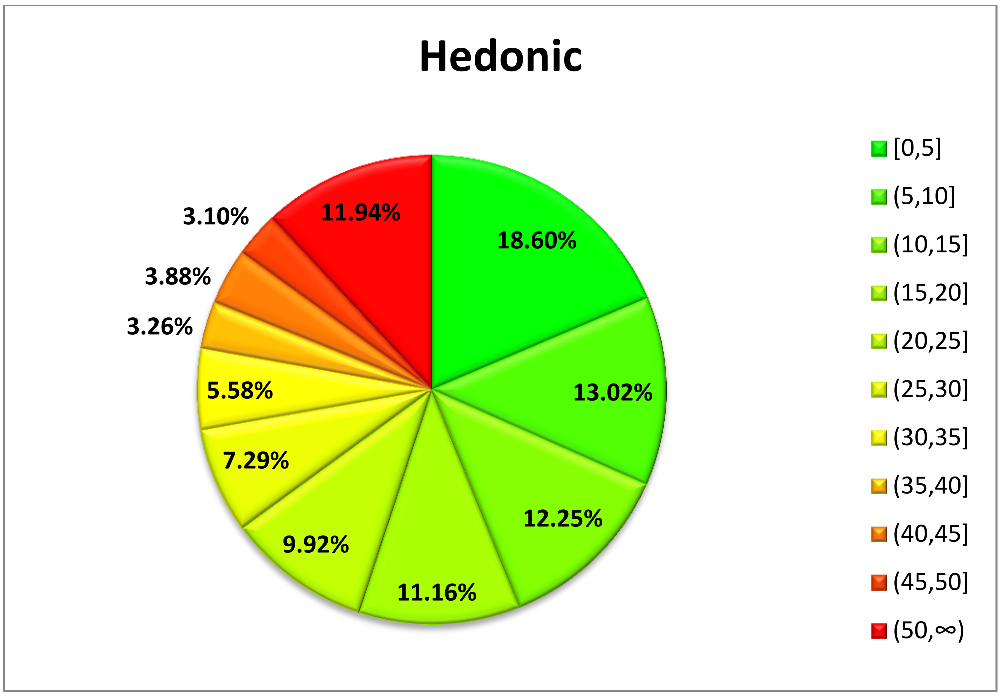
	≤5%	≤10%	≤15%	≤20%	≤25%	≤30%
FAVREP	0.366	0.515	0.648	0.747	0.822	0.874
DT	0.312	0.420	0.535	0.639	0.705	0.771
	≤35%	**≤40%**	**≤45%**	**≤50%**	**>50%**	
FAVREP	0.912	0.930	0.952	0.967	0.033	
DT	0.808	0.860	0.890	0.930	0.070	

In Table 6.9, percentage of count of records to amount of records in dataset is shown in 11 different values. As seen in Table 6.9, DT valuated 70.5 percent of all records with no difference or a difference which is smaller than or equal to 25 percent of the original listing price while valuating 7 percent of the records with a difference bigger than 50 percent of the original listing price.

6.4.3. FAVREP vs Hedonic Pricing Method

When test results from FAVREP and hedonic pricing method are compared, it’s seen that using hedonic pricing method genereates results with bigger error percentages.

Figure 6.4: Visual representation of error percentages in test results of HPM



As seen in Figure 6.4, 64.95 percent of the records used in test were valuated with a difference of maximum 25 percent to the original listing price. In the other hand, 11.94 percent of the records were valuated with a difference bigger than 50 percent to the original listing price.

Table 6.10: Amount of records in each error level, FAVREP vs. HPM

	[0 - 5]	(5 - 10]	(10 - 15]	(15 - 20]	(20 - 25]	(25 - 30]
FAVREP	236	96	86	64	48	34
Hedonic	120	84	79	72	64	47
	(30 - 35]	**(35 - 40]**	**(40 - 45]**	**(45 - 50]**	**(50 - ∞)**	
FAVREP	24	12	14	10	21	
Hedonic	36	21	25	20	77	

In Table 6.10, 11 buckets, representing the difference of the valuation result to original listing price, starting from [0,5] ending with (50, ∞) are used to show the count of records in each bucket. As seen in Table 6.10, HPM valuated 120 records with no difference or a difference which is smaller than or equal to 5 percent of the original listing price.

Table 6.11: Percentage of records in each error level, FAVREP vs. HPM

	[0 - 5]	(5 - 10]	(10 - 15]	(15 - 20]	(20 - 25]	(25 - 30]	(30 - 35]	(35 - 40]	(40 - 45]	(45 - 50]	(50 - ∞)
FAVREP	0.3659	0.1488	0.1333	0.0992	0.0744	0.0527	0.0372	0.0186	0.0217	0.0155	0.0326
Hedonic	0.1860	0.1302	0.1225	0.1116	0.0992	0.0729	0.0558	0.0326	0.0388	0.0310	0.1194

In Table 6.11, 11 buckets, representing the difference of the valuation result to original listing price, starting from [0,5] ending with (50, ∞) are used to show the percentage of records to dataset in each bucket. As seen in Table 6.11, HPM valuated 18.6 percent of all records with no difference or a difference which is smaller than or equal to 5 percent of the original listing price.

Table 6.12: Percentage of records to total records, FAVREP vs. AGNES

	≤5%	≤10%	≤15%	≤20%	≤25%	≤30%	≤35%	≤40%	≤45%	≤50%	>50%
FAVREP	0.366	0.515	0.648	0.747	0.822	0.874	0.912	0.930	0.952	0.967	0.033
Hedonic	0.186	0.316	0.439	0.550	0.650	0.722	0.778	0.811	0.850	0.881	0.119

In Table 6.12, percentage of count of records to amount of records in dataset is shown in 11 different values. As seen in Table 6.12, HPM valuated 65 percent of all records with no difference or a difference which is smaller than or equal to 25 percent of the original listing price while valuating 11.9 percent of the records with a difference bigger than 50 percent of the original listing price.

7. DISCUSSION AND CONCLUSION

In this thesis, we introduced a novel method for valuating real estate property in an accurate and approximate way. Novel method requires a training set with known market values to determine the price for the real estate property.

In the experimental results it is seen that, with limited knowledge on real estate properties, classification using decision trees generate better results when compared to hedonic pricing method. Yet the results from the decision tree classification can show improvement when records used in training set are carefully chosen. Our method determines the records to be used in building the decision tree model to be used for real estate property valuation, using data mining methods of clustering and classification with decision trees. It first filters records by their location. Then, uses clustering to find similar records with the property to be valuated, before finding the price range the property to be valuated is in. Using similar records within the determined price range and at the same location as training set improves the performance of the decision tree dramatically.

Normally, quality of the training set impacts the final valuation. Having many noises in the training set most probably will cause bad results for many methods including hedonic pricing method. Our method suggests noise detection in training set before valuation to eliminate outliers and extreme values and also as the clustering algorithm used can detect noises while determining clusters, it can be said that method introduced in this study is protected from noises in training set.

Method described in this study is novel and complete. Experimental results demonstrate that novel method described in this study works well on real estate property data, is not affected by database size, is not affected by noise and outliers, does not require statistical knowledge and can easily be implemented.

Our method requires generating clusters and decision trees on each execution which is costly as it requires generating a distance matrix to use when clustering which has complexcity of $O(n^2)$. In the other hand, when the method is used in bulk valuation of many real estates, count of running clustering algorithms and count of generation of decision trees decreases dramatically as it is not necessary to generate clusters or generate decision tree models for found clusters for every record. But, for single real estate property valuation clustering algorithm should run once and two different decision tree models should be generated. Although with the technology of the current day, speed of the method will not be a big issue, still the cost of the running algorithms are the most important shortcoming of our method.

REFERENCES

Books

Quinlan, J. R. 1979, *Discovering rules from large collections of examples: A case study*, in D. Michie, ed., `Expert Systems in the Micro Electronic Age', Edinburgh: Edinburgh University Press.

Quinlan, J. R. 1986, `Induction of decision trees', Machine Learning 1, *Readings in Machine Learning*, pp. 81-106, Shavlik and Dietterich (Eds.). San Mateo: Morgan Kaufmann.

Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann.

Periodicals

Adair, A., Mcgreal, S., Smyth, A., Cooper, J. & Ryley, T. 2000. House price and accessibility: The testing of relationships within the Belfast urban area. *Housing Studies*. **Vol. 15** (No. 5), pp. 699-716.

Bao, H. X. H. & Wan, A. T. K. 2004. On the use of spline smoothing in estimating hedonic housing price models: empirical evidence using Hong Kong data. *Real Estate Economics.* **Vol 32** (issue 3), pp. 487-507.

Bin, O. 2004. A prediction comparison of housing sales prices by parametric versus semi-parametric regressions. *Journal of Housing Economics*. (13), pp. 68-84.

Case, B.J., Clapp, R., Dubin & Rodriguez, M. 2004. Modeling spatial and temporal house price patterns: a comparison of four models. *Journal of Real Estate Finance and Economics.* **Vol. 29** (issue 2), pp. 167-191.

Fan, G., Ong, Z. S. E. & Koh, H. C. 2006. Determinants of house price: A decision tree approach. *Urban Studies.* **Vol. 43** (No. 12), pp. 2301-2315.

Fletcher, M., Gallimore, P. & Mangan, J. 2000. Heteroscedasticity in hedonic house price models. *Journal of Property Research*. **Vol 17** (2), pp. 93-108.

Filho, C. M. & Bin, O. 2005. Estimation of hedonic price functions via additive nonparametric regression. *Emprical Economics*. (30), pp. 93-114.

Goodman, A.C. 1998. Andrew Court and the invention of hedonic price analysis. *Journal of Urban Economics*, **Vol. 44**, pp. 291-298.

Janssen, C. B. & Soderberg, J. Z. 2001. Robust estimation of hedonic models of prive and income for investment property. *Journal of Property Investment & Finance*. **Vol 19**, No. 4, pp. 342-360

Kauko T., 2003, “On current neural network applications involving spatial modeling of property prices”, *Journal of Housing and the Built Environment* (18), pp. 159-181.

Kestens, Y., Theriault, M. & Rosier, F.D. 2006. Heterogeneity in hedonic modeling of house prices: looking at buyers’ household profiles. *J.Geograph Syst*. (8), pp. 61-96.

Kim, K. & Park, J. 2005. Segmentation of the housing market and its determinants: Seoul and its neighboring new towns in Korea. *Australian Geographer*. **Vol. 36** (No.2), pp. 221-232.

Meese, R., & N., Wallace 2003. House price dynamics and market fundamentals: The Parisian housing market. *Urban Studies*, **Vol. 40**, Nos.5-6, pp. 1027-1045

Quinlan, J. R., & Rivest, R. L. 1989. Inferring decision trees using the minimum description length principle. *Information and Computation*, (80), pp. 227-248.

Quinlan, J. R. 1996 Improved Use of Continuous Attributes in C4.5. *Journal of Artificial Intelligence Research* (4) pp. 77-90

Ridker, Ronald G., & John A Henning 1967 “ The Determinants of Residential Property Values with Special Reference to Air Pollution”. *The Review of Economics and Statistics* **49** (2), pp. 246-57.

Stevenson, S. 2004. New empirical evidence on heteroscedasticity in hedonic housing models. *Journal of Housing Economics*. (13), pp. 136-153.

Other Sources

Dougherty, J., Kohavi, R., & Sahami, M. 1995. Supervised and unsupervised discretization of continuous features. *Proceedings Twelfth International Conference on Machine Learning*, San Francisco: Morgan Kaufmann, pp. 194-202.

Ester, M., Kriegel, H.P., Sander, J. & Xu, X. 1996 “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 1996

APPENDICES

Appendix 1: Toy data set

Id	Neighborhood	Floor Count	Floor	Built Year	Bathroom Count	L. Room Count	Bedroom Count	Has Car Park	Square Meters	Sea View	Building Complex	Market Value
1	Anadolu Hisarı	5	3	2008	3	1	5	1	230	1	1	1325000
2	Anadolu Hisarı	4	2	1993	1	1	3	1	135	0	1	360000
3	Anadolu Hisarı	5	5	1998	2	1	4	1	215	0	1	485000
4	Anadolu Hisarı	5	5	1998	2	1	4	1	215	0	1	485000
5	Anadolu Hisarı	4	4	1993	2	1	3	1	135	0	1	360000
6	Anadolu Hisarı	5	5	1997	2	1	3	1	120	0	1	350000
7	Anadolu Hisarı	6	1	2004	2	1	3	1	120	0	1	235000
8	Anadolu Hisarı	6	1	2004	2	1	3	1	120	0	1	235000
9	Anadolu Hisarı	7	1	2004	1	1	2	1	110	0	1	225000
10	Anadolu Hisarı	5	1	2011	2	1	3	0	135	0	0	425000
11	Anadolu Hisarı	5	3	2011	1	1	3	0	100	0	0	400000
12	Anadolu Hisarı	3	1	2004	1	1	3	0	90	1	0	380000
13	Anadolu Hisarı	3	0	2004	1	1	3	0	90	1	0	350000
14	Anadolu Hisarı	4	2	1984	1	1	2	0	100	0	0	185000
15	Anadolu Hisarı	5	4	2011	2	1	3	0	150	0	0	500000
16	Anadolu Hisarı	4	0	1994	2	1	4	1	300	1	1	850000
17	Anadolu Hisarı	4	1	2004	2	1	3	0	145	1	1	470000
18	Anadolu Hisarı	8	6	2004	2	1	4	1	205	1	0	480000
19	Anadolu Hisarı	5	4	2011	1	1	3	1	135	0	1	365000
20	Anadolu Hisarı	5	4	2004	1	1	3	1	135	0	1	385000
21	Anadolu Hisarı	4	0	1998	3	1	6	1	310	1	1	1050000
22	Anadolu Hisarı	3	3	1980	1	1	2	0	120	1	0	400000
23	Anadolu Hisarı	5	5	1980	1	1	2	0	80	1	0	215000
24	Anadolu Hisarı	3	0	1980	2	1	5	0	300	0	0	525000
25	Yeşilpınar	4	2	2006	1	1	3	0	100	0	0	110000
26	Yeşilpınar	4	2	2000	1	1	3	0	140	0	0	110000

Appendix 1: Toy data set (continued)

Id	Neighborhood	Floor Count	Floor	Built Year	Bathroom Count	L. Room Count	Bedroom Count	Car Park	Square Meters	Sea View	Building Complex	Market Value
27	Yeşilpınar	4	2	2006	1	1	3	0	100	0	0	110000
28	Yeşilpınar	5	0	2006	1	1	2	1	95	0	1	125000
29	Yeşilpınar	4	2	2010	1	1	2	0	90	0	0	132000
30	Yeşilpınar	4	3	2010	1	1	2	0	90	0	0	128000
31	Yeşilpınar	4	1	2008	1	1	2	0	85	0	0	135000
32	Yeşilpınar	5	5	2000	1	1	2	0	115	0	0	128000
33	Yeşilpınar	3	1	2011	1	1	3	0	120	0	0	175000
34	Yeşilpınar	5	3	2000	1	1	2	1	95	0	1	145000
35	Yeşilpınar	3	3	2007	1	1	3	0	140	0	0	165000
36	Yeşilpınar	3	0	2009	1	1	2	0	90	0	0	92000
37	Yeşilpınar	4	0	2010	1	1	1	0	65	0	0	85000
38	Göztepe	6	4	2000	1	1	2	1	145	0	0	450000
39	Göztepe	8	3	1985	2	1	2	1	150	0	1	525000
40	Göztepe	8	3	1985	2	1	2	0	150	0	0	525000
41	Göztepe	6	4	2000	1	1	2	1	145	0	0	450000
42	Göztepe	10	3	1991	1	1	2	1	135	0	0	435000
43	Göztepe	7	3	2011	1	1	2	1	90	0	0	470000
44	Göztepe	8	3	1995	1	1	2	1	160	0	0	595000
45	Göztepe	10	4	1995	1	1	2	1	155	0	0	525000
46	Göztepe	6	1	2011	2	1	2	1	125	0	1	615000
47	Göztepe	12	2	1996	2	1	2	1	140	0	1	650000
48	Göztepe	15	3	1990	1	1	2	1	160	0	1	655000
49	Göztepe	6	1	1990	2	1	2	1	150	0	1	455000
50	Göztepe	6	2	1991	2	1	2	1	150	0	1	450000

Appendix 2: Distance matrix created on demonstration run

Id	0	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	22	23	24
0	0.000	0.045	0.203	0.203	0.071	0.073	0.090	0.090	0.130	0.171	0.231	0.289	0.306	0.267	0.181	0.486	0.165	0.339	0.101	0.052	0.360	0.439	0.588
2	0.045	0.000	0.248	0.248	0.026	0.084	0.116	0.116	0.156	0.196	0.256	0.314	0.331	0.222	0.226	0.475	0.210	0.384	0.126	0.077	0.315	0.394	0.562
3	0.203	0.248	0.000	0.000	0.222	0.224	0.276	0.276	0.334	0.357	0.434	0.492	0.475	0.470	0.327	0.283	0.331	0.210	0.304	0.255	0.529	0.608	0.384
4	0.203	0.248	0.000	0.000	0.222	0.224	0.276	0.276	0.334	0.357	0.434	0.492	0.475	0.470	0.327	0.283	0.331	0.210	0.304	0.255	0.529	0.608	0.384
5	0.071	0.026	0.222	0.222	0.000	0.058	0.124	0.124	0.182	0.205	0.282	0.340	0.323	0.248	0.234	0.449	0.219	0.392	0.152	0.103	0.306	0.385	0.536
6	0.073	0.084	0.224	0.224	0.058	0.000	0.067	0.067	0.124	0.206	0.225	0.282	0.265	0.247	0.236	0.493	0.220	0.394	0.154	0.105	0.305	0.384	0.594
7	0.090	0.116	0.276	0.276	0.124	0.067	0.000	0.000	0.057	0.140	0.158	0.216	0.233	0.278	0.169	0.559	0.154	0.327	0.088	0.038	0.371	0.450	0.660
8	0.090	0.116	0.276	0.276	0.124	0.067	0.000	0.000	0.057	0.140	0.158	0.216	0.233	0.278	0.169	0.559	0.154	0.327	0.088	0.038	0.371	0.450	0.660
9	0.130	0.156	0.334	0.334	0.182	0.124	0.057	0.057	0.000	0.197	0.159	0.216	0.234	0.221	0.227	0.617	0.211	0.385	0.128	0.078	0.353	0.393	0.718
10	0.171	0.196	0.357	0.357	0.205	0.206	0.140	0.140	0.197	0.000	0.078	0.234	0.251	0.296	0.030	0.640	0.199	0.321	0.070	0.119	0.389	0.468	0.619
11	0.231	0.256	0.434	0.434	0.282	0.225	0.158	0.158	0.159	0.078	0.000	0.156	0.173	0.218	0.108	0.717	0.277	0.399	0.130	0.179	0.390	0.390	0.697
12	0.289	0.314	0.492	0.492	0.340	0.282	0.216	0.216	0.216	0.234	0.156	0.000	0.017	0.276	0.263	0.601	0.161	0.282	0.286	0.237	0.274	0.235	0.754
13	0.306	0.331	0.475	0.475	0.323	0.265	0.233	0.233	0.234	0.251	0.173	0.017	0.000	0.293	0.281	0.584	0.178	0.300	0.303	0.254	0.257	0.217	0.737
14	0.267	0.222	0.470	0.470	0.248	0.247	0.278	0.278	0.221	0.296	0.218	0.276	0.293	0.000	0.326	0.697	0.397	0.519	0.348	0.299	0.172	0.172	0.536
15	0.181	0.226	0.327	0.327	0.234	0.236	0.169	0.169	0.227	0.030	0.108	0.263	0.281	0.326	0.000	0.610	0.190	0.291	0.099	0.148	0.419	0.498	0.589
16	0.486	0.475	0.283	0.283	0.449	0.493	0.559	0.559	0.617	0.640	0.717	0.601	0.584	0.697	0.610	0.000	0.440	0.319	0.587	0.538	0.582	0.661	0.275
17	0.165	0.210	0.331	0.331	0.219	0.220	0.154	0.154	0.211	0.199	0.277	0.161	0.178	0.397	0.190	0.440	0.000	0.208	0.182	0.133	0.316	0.395	0.681
18	0.339	0.384	0.210	0.210	0.392	0.394	0.327	0.327	0.385	0.321	0.399	0.282	0.300	0.519	0.291	0.319	0.208	0.000	0.356	0.306	0.438	0.517	0.507
19	0.101	0.126	0.304	0.304	0.152	0.154	0.088	0.088	0.128	0.070	0.130	0.286	0.303	0.348	0.099	0.587	0.182	0.356	0.000	0.049	0.441	0.520	0.688
20	0.052	0.077	0.255	0.255	0.103	0.105	0.038	0.038	0.078	0.119	0.179	0.237	0.254	0.299	0.148	0.538	0.133	0.306	0.049	0.000	0.392	0.471	0.639
22	0.360	0.315	0.529	0.529	0.306	0.305	0.371	0.371	0.353	0.389	0.390	0.274	0.257	0.172	0.419	0.582	0.316	0.438	0.441	0.392	0.000	0.079	0.538
23	0.439	0.394	0.608	0.608	0.385	0.384	0.450	0.450	0.393	0.468	0.390	0.235	0.217	0.172	0.498	0.661	0.395	0.517	0.520	0.471	0.079	0.000	0.617
24	0.588	0.562	0.384	0.384	0.536	0.594	0.660	0.660	0.718	0.619	0.697	0.754	0.737	0.536	0.589	0.275	0.681	0.507	0.688	0.639	0.538	0.617	0.000