



PREDICTING THE RISK OF SEIZING STATE LANDS USING DATA
MINING TECHNIQUES

Hussein Ali Ahmed AHMED

OCTOBER, 2017

PREDICTING THE RISK OF SEIZING STATE LANDS USING DATA MINING
TECHNIQUES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED
SCIENCES OF
ÇANKAYA UNIVERSITY

BY
Hussein Ali Ahmed AHMED

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF
MASTER OF SCIENCE
IN
INFORMATION TECHNOLOGY

OCTOBER, 2017

Title of the Thesis: **PREDICTING THE RISK OF SEIZING STATE LANDS**
USING DATA MINING TECHNIQUES

Submitted by: **Hussein Ali Ahmed AHMED**


Approval of the Graduate School of Natural and Applied Sciences, Çankaya University.



Prof. Dr. Can ÇOĞUN

Director


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



Prof. Dr. Erdoğan Dođdu

Head of Department

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



Prof. Dr. Erdoğan Dođdu

Supervisor

Examination Date: 17/10/2017

Examining Committee Members:

Prof. Dr. Erdoğan Dođdu


(Çankaya Univ.)


Assoc. Prof. Dr. Osman Abul


(TOBB Univ.)

Assoc. Prof. Dr. Reza Hassanpour

(Çankaya Univ.)







R. Hassanpour

STATEMENT OF NON-PLAGIARISM PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name : **Hussein Ali Ahmed AHMED**

Signature

:



Date

: **17/10/2017**

ABSTRACT

PREDICTING THE RISK OF SEIZING STATE LANDS USING DATA MINING TECHNIQUES

AHMED, Hussein

M.Sc., Information Technologies Department

Supervisor: Prof. Dr. Erdoğan Dođdu

October 2017, 68 pages

Adverse possession is a common problem that exists in many countries all around the world but in different rates. These rates depend on many factors, some are related to the conditions in the countries and others are related to the individuals who are adversely possessing a land. The adverse possession may be divided into two types, intentional and unintentional adverse possession. The intentional adverse possession poses difficult challenges to municipalities responsible for the regions where lands are intentionally seized to be adversely possessed in the future. The conditions that Iraq has been going through in last few years contributed to a huge increase in state's land seizing in order to adversely possess these land in the future, for both residential and commercial purposes. Thus, it is important to find ways to limit this phenomenon to reduce its effects over the cities where they occur. The existing seizures rates are way beyond the capabilities of the municipalities who are responsible for finding an appropriate solution. Thus, it is important to reduce the number of squatters that need to be processed by the municipality to a reasonable number that matches the limited resources of these municipalities using data mining techniques.

In this study, six classifiers are examined to predict the risk of a squatter going further to seize another land by extracting the knowledge from the existing dataset that

contains information of the squatters and the seizures they made. The classifiers tested in this study are Bayesian, decision tree and lazy classifiers by evaluating their performances using the accuracy and F-measure of the classification results. Then, the number of features, used by the classifiers to predict a class for each tuple, is reduced by ranking these features and removing the least ranked features one by one while evaluating the classifiers' performances using the remaining features. Two features ranking techniques are used, which are information gain and one rule methods. The highest number of features that are removed without affecting the performance of the classifiers is six features based on the ranking of the one rule method. This ensures faster prediction, for squatter's risk level of seizing another land.

The classifier that is selected to implement a method, which assists the municipality of Baquba in data management and squatters risk assessment, is the k -NN classifier, depending on the results acquired from the conducted experiment. The implemented method is then used to classify the squatters in the database and predicted that 395 (76.10%) of the squatters are high-risk of going further toward seizing another land, while the remaining 124 (23.90%) are predicted to be of low risk.

ÖZ

Veri Madenciliği Teknikleri Kullanarak Kamu Alanlarının İşgali Riskinin Tahmin Edilmesi

AHMED, Hussein

Yüksek Lisans, Bilgi Teknolojileri Anabilim Dalı

Tez Yöneticisi: Prof. Dr. Erdoğan Doğdu

Ekim 2017, 68 sayfa

Fiili işgal, tüm dünyada birçok ülkede olan ortak bir sorundur, ancak farklı oranlarda bulunmaktadır. Bu oranlar birçok faktöre bağlıdır; bazıları ülkelerdeki koşullarla ilişkili ve bazıları ise arazisine olumsuz bir şekilde sahip olanlarla ilgilidir. Fiili işgal, kasıtlı ve kasıtsız mülkiyet olmak üzere iki kısma ayrılabilir. Fiili işgal, sorumlu belediyelere gelecekte olumsuz bir şekilde, topraklarda kasıtlı olarak ele geçirilen bölgeler için, zorlu koşullar ortaya koyar. Irak'ın son birkaç yıldır içinde bulunduğu şartlar; hem yerleşim hem de ticari yönden bu topraklara sahip olması geleceğe olumsuz bir şekilde neden oldu. Dolayısıyla bu durumda, şehirler üzerindeki meydana gelen etkilerini azaltmak için bunu sınırlandırmanın yollarını bulmak önemlidir. Mevcut işgal oranları, uygun bir çözüm bulma sorumluluğunu üstlenen belediyelerin kapasitesinin ötesine geçmektedir. Bu nedenle, belediye tarafından işlenmesi gereken gecekonduların sayısı, veri madenciliği tekniklerini kullanarak bu belediyelerin sınırlı kaynaklarıyla eşleşen makul bir sayıya düşürülmesi önemlidir.

Bu çalışmada, gecekonduda bulunanlar hakkında bilgi içeren mevcut veri kümesindeki bilgileri kullanarak, başka bir araziye ele geçirme riskini bulmak için, altı sınıflandırıcı incelenmiştir. Bu çalışmada incelenen sınıflandırıcılar Bayesian,

karar ağacı, ve tembel sınıflandırıcılardır. Sınıflandırma sonuçlarının doğruluğu ve F-ölçüsünü kullanarak performansları değerlendirilmiştir. Ardından sınıflandırıcıların, her bir grup için bir sınıf tahmin etmesinde kullandığı özelliklerin sayısı, bu özellikler derecelendirilerek ve en düşük sıralı özellikler birer birer kaldırılıp kalan özellikler kullanılarak sınıflandırıcıların performansları değerlendirilmiştir. Bilgi kazanımı ve bir-kural (one-rule) yöntemleri kullanılarak özellikler derecelendirilmiştir. Sınıflandırıcıların performansını etkilemeden kaldırılan en fazla özellik sayısı altıdır, bir kural yönteminin sıralamasına dayanarak elde edilmiştir. Bu, başka bir yeri işgal etme riskinin hızlı tahminini sağlar.

Baquba belediyesindeki veri yönetiminde ve işgal risk değerlendirmesinde yardımcı olan bir yöntem olarak uygulamak üzere seçilen sınıflandırıcı, yapılan deneyden elde edilen sonuçlara bağlı olarak k-NN sınıflandırıcısıdır. Uygulanan yöntem, daha sonra işgalde bulunan kişilerin sınıflandırılması için kullanıldı ve kalan 395'i (%76.9) diğer arazileri ele geçirme yönünden yüksek risk taşıdığı ve kalan 124'ünün ise (%23.9) düşük riskli olduğu tahmin edilmektedir.

ACKNOWLEDGMENTS

I would like to thank my supervisor, Prof. Dr. Erdoğan Dođdu, for all the guidance and assistance he provided during all the stages of the thesis. I also would like to show my gratitude to the one who taught me my very first word, my mother. I would also appreciate all the wisdom provided by the person who put my foot on the first step on the way, my father. Many thanks also go to the one who supported me along the way, my wife. Finally, I would like to thank my son and daughter who have always lightened my life with their beautiful smiles.

TABLE OF CONTENTS:

Statement of Non-Plagiarism Page	iii
Abstract	iv
ÖZET	vi
Acknowledgments.....	viii
Table of Contents:.....	ix
List of Tables:	xi
List of Figures	xiii
List of Abbreviations:	xiv
Chapter 1	1
Introduction	1
1.1. Adverse Possession	2
1.2. Data Mining.....	4
1.3. Methods evaluation	4
1.4. Problem definition.....	5
1.5. Aim of the thesis.....	5
Chapter 2.....	7
Related Work	7
Chapter 3.....	13
Squatters Risk Assessment.....	13
3.1. Data Collection.....	13
3.2. Data Preprocessing	15

3.2.1.	Data Labeling	15
3.2.2.	Identification Number Removal	16
3.2.3.	Numerical Data Conversion	16
3.3.	Data Classification.....	17
3.3.1.	Bayesian Classifiers.....	17
3.3.2.	Lazy Classifiers	18
3.3.3.	Decision trees	20
3.4.	Performance Evaluation	21
3.4.1.	Accuracy	21
3.4.2.	F-measure	22
3.5.	Features Selection.....	23
3.5.1.	Information Gain Method	23
3.5.2.	One-Rule Method	24
Chapter 4	25
Experimental Results	25
4.1.	Classifiers' Performance	25
4.1.1.	Split Data Evaluation.....	26
4.1.2.	Cross Validation Evaluation.....	29
4.2.	Features Selection.....	32
Chapter 5	36
Discussion	36
Chapter 6	44
Method Implementation	44
Chapter 7	47
Conclusion	47
References	50

LIST OF TABLES:

Table 2.1: Squatting related attributes.	8
Table 2.2: Squatting affecting attributes in Yemen.	9
Table 3.1: Summary of collected data.	16
Table 3.2: Sample dataset.	23
Table 4.1: Split data classification results using BayesNet classifier.	26
Table 4.2: Split data classification results using Naïve Bayesian classifier.	26
Table 4.3: Split data classification results using k-NN classifier with k=1.	27
Table 4.4: Split data classification results using k-NN classifier with k=3.	27
Table 4.5: Split data classification results using K-Star classifier.	28
Table 4.6: Split data classification results using J48 classifier.	28
Table 4.7: Split data classification results using the Random Forest classifier.	29
Table 4.8: Cross-validation classification results using the BayesNet classifier.	30
Table 4.9: Cross-validation classification results using the Naïve Bayes classifier.	30
Table 4.10: Cross-validation classification results using k-NN classifier with k=1.	30
Table 4.11: Cross-validation classification results using k-NN classifier with k=3.	31
Table 4.12: Cross-validation classification results using K-star classifier.	31
Table 4.13: Cross-validation classification results using J48 classifier.	31
Table 4.14: Cross-validation classification results using random forest classifier.	32
Table 4.15: Features ranking using Information Gain method.	32
Table 4.16: Effect of features removed based on Information Gain on classifiers evaluated using data split method.	33

Table 4.17: Effect of features removed based on Information Gain on classifiers evaluated using data cross-validation.....33

Table 4.18: Features ranking using the One-Rule method.....34

Table 4.19: Effect of features removed based on One-Rule method using data split technique.34

Table 4.20: Effect of features removed based on One-Rule method using cross-validation technique.35



LIST OF FIGURES

Figure 4.1: The decision tree created by the J48 classifier.	29
Figure 5.1: Classifiers' performance summary illustration for the data split technique.	37
Figure 5.2: Classifiers' performance summary illustration for the cross-validation technique.	37
Figure 5.3: Features removal effect on the classifiers' F-measure using Information Gain ranking method and data split technique.....	38
Figure 5.4: Features removal effect on the classifiers' F-measure using Information Gain ranking method and cross-validation performance evaluation technique.	39
Figure 5.5: Features removal effect on the classifiers' F-measure using One-Rule ranking method and data split method.	40
Figure 5.6: Features removal effect on the classifiers' F-measure using One-Rule ranking method and cross-validation method.	41
Figure 6.1: New seizure interface for an existing squatter.	45

LIST OF ABBREVIATIONS:

Logistics Regression	LR
Artificial Neural Network	ANN
Human Inspection	HI
Conditional Probability Distribution	CPD
Directed Acyclic Graph	DAG
<i>k</i> -Nearest Neighbor	k-NN
Structured Query Language	SQL
Graphical User Interface	GUI

CHAPTER 1

INTRODUCTION

Possessing a piece of land has always been one of the main goals to be achieved by humans and is one of the most precious assets to own. It provides a shelter, a workplace or both for the individuals and their families. The ease of achieving this goal is controlled by many factors, some of them are related to the individuals themselves and the others are related to the country or the area where the land, intended to be owned, belongs to. Thus, when it is quite difficult for an individual to purchase a land, and it is still an important goal to achieve, some people tend to go through other ways to possess a piece of land. One of these ways is the adverse possession.

The remainder of this study is organized as follows. Chapter two reviews the literature related to this study. Chapter three presents the available data collected for the existing seizures and how these data are processed in order to use them to train and evaluate the selected classifiers. The classifiers, selected to be used in the class prediction, are also discussed in chapter three alongside with performance evaluation methods, used to evaluate the performance of each classifier, and the features selectors used for features ranking in order to eliminate the features that have no effect on the classification process. Chapter four shows the results of the experiments conducted to evaluate the performance of the classifiers and the features selectors. Then, these results are discussed in chapter five in order to select the most appropriate classifier for the method implementation using the least number of features to ensure minimum process time. Chapter six demonstrates the method implemented, based on the selected classifier and features, for the municipalities to store and manage the data, and predict

a risk level for new squatters. Finally, the conclusions of this study are summarized in chapter seven.

1.1. ADVERSE POSSESSION

Adverse possession is defined as the occupation of a piece of land without the permission of the lawful owner of it with the intention of possessing this land as an individual's own asset [1].

Adverse possession, as a term, has a general meaning that refers to any land possessing without the permission of the lawful owner regardless whether this adverse possession was intended to happen or not. For example, a person unintentionally builds one of the house's walls few centimeters deep into the neighbor's property. This adverse possession is accidental and the neighbor has the right to accept or refuse such adverse possession. Squatting, on the other hand, is an adverse possession where a squatter *is a person who adversely possesses an unoccupied or abandoned land for residence or work purposes*. This means that the squatter, intentionally, adversely possesses the land with the will to eventually own it [2].

Most of the countries have rules to regulate the adverse possession as it is still ongoing, up today, all around the world but in different rates. Countries that suffer from wars, lack of security, crises or low individual income rate have higher adverse possession rates than other countries. Although it still happens that a squatter adversely possesses a private property, it is more frequent for squatters to adversely possess state's lands [3]. This affects the state's municipality on many aspects. The financial loss is the most obvious effect of squatting state's lands, which is a loss of the value of the squatted land in addition to any profits planned to be earned by investing this land.

Two unplanned changes occur because of state's land squatting, the first is because of the addition of the buildings while the other one is because of the unplanned population growth in some regions. The unplanned addition of buildings will badly influence the appearance of the city, as these buildings are usually built on a very limited budget and without permission from the municipality. These building will also

share infrastructure of the neighboring buildings, which creates a greater burden on infrastructure that may overload it to breakdown. The unplanned growth of the population, on the other hand, affects the quality of services provided to that region such as education, transportation and health care services. Thus, squatting affects the entire neighborhood where it happens and sometimes requires urgent interventions to maintain the infrastructure and services provided to that region. Another important influence of adverse possession is on the projects planned by the municipality, where many projects are halted because there are squatters adversely possessing lands that are assigned to these projects [4].

In Iraq, adverse possession is a very ancient problem that faced many governments through history up today. Recently, squatting rate has grown rapidly in Iraq because of the extremely high cost of lands, low individual income, lack of security, wars and the forced displacement out of some regions because of the terrorist attacks. This rapid growth poses many challenges to the municipalities of the cities closer to regions where people are forcibly displaced and where income rate is low. Thus, Baquba municipality is one of the most municipalities that faced these challenges [5].

Although the city's municipality is the authority that is required to deal with the squatting problem, it needs to coordinate with many other authorities to take preventive action in an attempt to control the adverse possession problem because of the limited authority of the municipality as well as the limited available resources. Thus, Baquba municipality is unable to take the necessary preventive actions, to reduce the rate of adverse possession, which is distributed around the wide area of Baquba city, against the huge number of squatters that exists in the meantime. Thus, it is important to classify the squatters into groups according to the estimated risk of squatting another land in order to allow the municipality to take necessary actions, to control the situation, with the available resources and the coordination with the other authorities.

This classification can be made using data mining techniques to learn, from the existing database, the characteristics of the squatters who have more than one squatted

land and classify the remaining squatters, using the knowledge acquired from the training dataset, into clusters according to the estimated risk of making another squat.

1.2. DATA MINING

Data mining is the process of knowledge extraction from huge datasets by finding interesting patterns, which represent the relationships among the dataset's attributes, and use them to predict the future trends of the new incoming data. Data mining techniques use interdisciplinary methods that may include statistics, machine learning, and database management subfields.

Classification is one of the important data mining techniques, which is widely used to classify each tuple in a dataset into one of the predefined classes. Many mathematical methods are used in classification, such as linear programming, decision tree, statistics and neural networks. These methods are used to develop software that is capable of extract, from a pre-classified data set, the necessary knowledge to use it in order to predict the class that new, unclassified, tuples belong to by applying the extracted knowledge to the new data [6].

Association rules are sets of if/then statements that represent the relationships among, what may seem to be, unrelated data from a relational database. These rules are concluded by analyzing the currently classified data to be applied to any unclassified data in order to predict the class that each tuple, in this unclassified data, falls into. By this classification, the future behavior of this tuple is predicted by knowing the characteristics of the class it belongs to and the overall tendency of the tuples in this class. The most important relations are concluded using the support and confidence criteria. The support represents how frequent the tuples appear in the dataset, while the confidence is the count of how many times the if/then sets are able to accurately predict the result.

1.3. METHODS EVALUATION

As there are many methods proposed for classification in data mining, these methods must be evaluated in order to choose one method to be implemented in the

model. There are many factors that are used to describe how good the classification results are such as purity, accuracy and F-measure. Although there is a slight difference between the calculations of each method, all methods rely on the results of reclassifying the classified data after concluding the association rules. In other words, the training data is used for knowledge extraction, then, these rules are used to classify the same data and compare the resulted class to the original class. These results are distributed in a matrix called the confusion matrix. All performance measures are calculated using this confusion matrix, which contains the original classifications of the training set and the prediction results, distributes in a way that makes it easy to calculate all the performance measures. It is important to use more than one performance measure to compare classifiers as using only one measure may be misleading in some specific cases [7].

1.4. PROBLEM DEFINITION

According to the abnormal situations that Iraq is going through, like high unemployment rates, terrorist attacks and forced displacements. The rate of state's lands adverse possession is growing rapidly. Thus, it became very important to take preventive actions to limit this phenomenon. These actions require coordination and resources that are way beyond the abilities of the municipalities responsible for these lands. Thus, it is important to classify the squatters according to their tendency to squat another land.

1.5. AIM OF THE THESIS

The aim of this thesis is to test the available data mining classification methods and find the method that has best results to be used for classifying the squatters' database according to the risk of squatting another state land. The squatters predicted to be high risk are suggested to the decision makers in Baquba municipality to take the necessary actions that prevent them from going further and squatter another land. This classification reduces the number of squatters that must be processed, by the municipality and other authorities, using the existing limited resources. Then, a method is implemented in order to predict the risk level of any new squatter added to

the database. The method also stores and manages the information of the squatters and the seizures, and provides tables for the existing seizures, processed squatters and the high-risk squatters who are not processed yet. This method makes it possible to process the squatters one by one, as they are entered into the database, within the available limited resources of the municipalities.



CHAPTER 2

RELATED WORK

The *bad faith adverse possessor* is defined, by [3], as the person who intentionally occupies a land that has no lawful right to occupy it and is described, by the law, as an anomalous figure. Furthermore, it considers the bad faith adverse possessors as thieves, as they are willing to own the occupied land as their own. On the other hand, [8] shows that adverse possessors with good faith tend to fare better than those with bad faith. Thus, although good faith is mandatory, it may have heavy influence during the judicial assessment. Moreover, [2] suggests that the refusal of rewarding innocent mistakes helps reduce making these mistakes, pointing that entitling a good faith possessor to the occupied land, adverse possession, is a reward that encourages others to be less careful when they occupy a piece of land.

The lack of governments' plans to accommodate the growing rate of rural-urban migration is, proposed by [9] as, one of the factors that lead to informal settlements in the city. The informal settlement may be divided into two categories, slums and squatters. The slums are areas where the population grows rapidly without any expansion in the architect of that area, leading to a high population density that exceeds the planned rate. The squatted area is an area that has not been occupied before, which usually has no infrastructure or services, and is being occupied without permission from the lawful owner of the land, which is mostly either the municipality or the state. Thus, this lack of plans leads to squatting, which eventually leads to adverse possession. Few more factors are found to have less influence on informal settlements like inappropriate spatial, poor governance and corruption.

The characteristics of a squatter area are well described by [10] using three aspects of description. First are the physical characteristics, which shows that the infrastructure and services provided to these areas are below the acceptable levels. This is attributed to the non-legal status of these areas. Second are the social characteristics, where most squatters and their family members are found to be working on, or near, the minimum wages, but some household incomes may be at high levels as there are many working family members. Squatters are usually found to be either rural-urban or urban-urban migrants, but there also exist some second and third generation squatters. Finally, the legal characteristics of the squatters are described, where the lack of lawful ownership of the occupied land is the major legal issue existed. In summary, internal and external attributes are found to be related to the squatting and the determination of the settlement's size and quality. The internal attributes are related to the squatters, while the external attributes are out of their control. These attributes are summarized in Table 2.1.

Table 2.1: Squatting related attributes.

Internal Attributes	External Attributes
Ethnicity	Municipal/government policies
Origin	Tenure security
Language	Landowner
Housing investment	Length of stay in city
Renters' presence	
Workplace	
Length of stay in settlement	
Construction activity	

Some regions are making the adverse possession laws more stringent than they used to be. For example, [11] explains how Alaska state in the United States of America changed the adverse possession laws to provide more protection to the lands' owners in 2003. According to the unique circumstances of the state of Alaska, prior to 2003, the laws used to be more favorable to the adverse possessor than the lawful owner. This law revision is assumed to eliminate the bad faith squatters from adversely possessing a land in order to reduce the rate of adverse possession in the state.

In developing countries, [12] shows that squatting is considered as one of the most critical problems faced by urban cities. This is resulted from the unparalleled growth of population as well as the low-economy countryside to the main cities with the dream of better life. It also shows that most of the squatted lands are owned by the government. The vast number of squatters and their determination germinated a challenge to the land owner as they worry about personal retaliation in case they attempt to expel the squatters.

The effects of wars and political situation on the phenomenon of squatting are discussed by [4]. The authors studied the rapid growth in the number of squatters in Yemen in the early nineties when a union is established between south and north Yemen and the soldiers had just returned from the second Gulf war. The study investigated the relation between squatting and a set of attributes. These attributes are split into two categories as shown in Table 2.2. The effect of squatting is found to be mostly on environment and health. Thus, the study suggested that the government must step into the matter and take some preventive action to reduce the effects of squatting and adverse possession.

Table 2.2: Squatting affecting attributes in Yemen.

Category	Attribute
Demographic	Gender
	Age
	Marital status
	Education level
	Occupation sector
Physical	Squatting duration
	Renovation
	House extension
	House size

Adverse possession is an ancient phenomenon in Iraq, as stated by [5]. Huge areas of land were adversely possessed by tribes with no attempts to prevent that by the Ottoman government, at that time. Later, these lands are split into smaller units and distributed among the families of that tribe as the government started to weaken

the influence of the tribal groups in favor of the family as the unit of the society. Adverse possession of state lands is organized by a law enacted in 1932. This law is based on Alezma rights, which is an Ottoman system. This legal system is based on the idea that a state's piece of land may be reclaimed by cultivating it. If the person continues cultivating the land for a significant amount of time and keeps improving it, that person may apply to the government for possession and pay only one-third of the actual price.

In order to control the growth of such phenomenon, it is important to predict the actions of the squatters in order to figure out the bad faith squatter, which is considered to be at a high risk of squatting another land. The prediction of human actions is not as difficult as it looks, in fact, these predictions are always used for planning, such as predicting the spread of humans during city planning, [13].

These predictions are also quite accurate, for example, [14] assumes that it is possible to predict a person's location, at any given time, with an accuracy of 93% at the uncertainty of 3 km². Thus, for more specific areas, like airports, malls or train terminals, the accuracy is much higher. Such predictions are used for resources optimization and safety.

Data mining techniques are used to extract domain knowledge from a dataset in order to use this knowledge to predict a matching class for any new entries. These techniques are used by [15] to predict the strategy followed by a person in order to predict the next action to be taken by classifying this person into one of the predefined classes. Then, the overall characteristics of the class are used for that person. A class may represent one or a group of characteristics.

Classification is one of the most important techniques of data mining. There are many methods for classification, but in general, they use a pre-classified dataset for training purposes in order to find relations between the attributes of the dataset and the distribution of tuples in these classes, [16]. There are many proposed methods to classify the data. A method may perform better than another on a certain dataset, while it shows poor performance on another dataset. Thus, it is important to evaluate more than one method on each dataset in order to find the best classifier to be used to classify

the test data. It is important to mention that the training data and test data must have the same structure, the only difference between the training dataset and the test dataset is that the training dataset is used to learn the classification scheme while the test dataset is classified, using the extracted knowledge from the training dataset, and compared to the original classification in order to test how good is the classifier's performance.

There are many classifiers that may be used to extract the knowledge from a dataset and used to classify similar tuples in order to predict their future behavior. These classifiers may be categorized, depending on the methodology used to design the classifier, into different categories. Three of the most popular classifiers' categories are the Bayesian classifiers, lazy classifiers and the decision trees. As their names suggest, the Bayesian classifiers are based on Bayesian networks, the powerful probabilistic representation. The lazy classifiers are the classifiers that do not extract any knowledge until the moment the classification process starts. Moreover, the decision-tree classifiers create paths of if then statements that are shaped like an upside down tree, where the root is on the top and the leaves are at the bottom, [17].

The Bayesian classifiers use computational and statistical methods to calculate the probability of a tuple to be in each class depending on the distribution of the existing training dataset. Then, the class, with the highest probability that this tuple belongs to, is selected as a predicted class for that tuple, [18].

The lazy classifiers, on the other hand, predicts a class for a tuple depending on the distribution of the most similar tuples, in the training dataset, among the existing classes. Thus, it is impossible to extract the related knowledge until the tuple required to be classified is known to the classifier, which means that knowledge extraction is related to the tuple being classified and the knowledge extraction process cannot be started prior to the classification process [19]. The fact the knowledge is extracted per each classification process is the reason behind the huge time consumption, when compared to other types of classifiers, and the permanent need for the training dataset existence. But there is also an advantage of that the knowledge extracted by these classifiers are dynamic and they change as the behavior of the tuples in the training dataset changes, [19].

Decision tree classifiers represent the extracted knowledge as paths of conditions, where each path leads to a single class called leaf. These conditions are distributed in levels from top to bottom, which means that the root of the tree is on top and the leaves are at the bottom. Each level is derived from the upper level based on the value of a specific feature that is specified in the higher level. Eventually, after selecting the directions based on these values from root to leaf, the class in that leaf is predicted as a class for the tuple being classified, [20].

Data mining techniques are widely used in different fields, for example, they are used by the governments to classify suspects according to the risk of committing a crime. The model built by [21] is used to classify vessels according to the probability of being used for smuggling or not. The statistical technology, Logistics Regression (LR), and the information technology, Artificial Neural Networks (ANN) are used to classify and predict the smuggling behaviors. The prediction results of these two methods are compared to the Human Inspection (HI). The results show significant superiority for the prediction results of the data mining techniques when compared to the Human Inspection method. On the other hand, the results show that the better performance, between the two data mining methods, is achieved by the ANN method after making many adjustments to the obtained knowledge during learning.

As the predicted class is concluded using the values in each attribute of the new tuple being classified and the values in the training dataset, the time consumed to classify the new tuple is proportional to the number of features used in the classification process, regardless of the type of the classifier used in the process.

CHAPTER 3

SQUATTERS RISK ASSESSMENT

Assessing the risk of a squatter to go further and seize another land is very important for the Iraqi municipalities according to their limited resources, compared to the required resources, to take necessary legal actions against all squatters to prevent the seizure of more state's lands. As these squatters are distributed over very wide areas, and the security situation in such areas is usually critical, which requires the cooperation of the municipalities with other authorities for protection and execution, it is very difficult for the municipalities to take those actions against all squatters. Thus, only squatters who are predicted to be at high risk of seizing another state land are suggested to the decision makers to make the appropriate legal actions in order to limit the adverse possession of state's lands.

To assess the risk of each squatter, the data of all adversely possessed lands are collected in order to use these data to train a classifier that is used to predict a risk level for each squatter. This procedure enables the decision makers of taking actions against the squatters predicted by the classifier to be with a high risk level of seizing another state land, and may also be used to provide priority levels depending on the prediction confidence of the classifier, where squatters predicted, by the classifier, to be in high-risk level with high prediction confidence are supposed to be processed first.

3.1. DATA COLLECTION

The data collected for the seized lands include information about the seized land as well as information about the squatter. These features are selected based on the

suggestions in the earlier studies, keeping in mind the differences among data logged in different countries, where for example, the “Ethnicity” information is suggested, by [10], to be effective in the adverse possession phenomenon, but this feature is not available in Iraq as it is not collected by the government. The data are collected for adversely possessed state’s lands using the following features:

- **Squatter’s Identification Number.** A unique number per each squatter. This number is used to protect the personal information of the squatter.
- **Squatter’s Gender.** The gender of the squatter whether to be a male or a female.
- **Squatter’s Age.** A numerical value that represents the age of the squatter at the year when this study is conducted, which is 2017.
- **Squatter’s Marital Status.** A categorical feature that holds one of four marital statuses, which are single, married, divorced or widow.
- **Squatter’s Parenthood Status.** A categorical feature that holds information whether the squatter has children or not.
- **Squatter’s Family Members.** A numerical feature that holds the number of family members who are supported by the squatter.
- **Squatter’s Residence Status.** A categorical feature that may contain one of two categories that represent whether the squatter has a residence place other than the seized land or not.
- **Squatter’s Residence Place.** A categorical feature that describes how far is the residence of the squatter, if exists, from the seized land. This feature may contain one of two categories that are near and far.
- **Squatter’s Origin.** This feature is a categorical feature that consists of two categories, which are same and different. These categories represent whether the squatter is originated from the same city or from a different one.
- **Squatter’s Educational Level.** A categorical feature that represents the highest degree acquired by the squatter in education.
- **Squatter’s Employment.** The type of employment is set in this feature, where a squatter may be working for the government, a private company, self-employed or unemployed.

- **Seized Land Type.** The purpose that the seized land is planned to be used for by the municipality. This may be one of three formal purposes that are residential, commercial or industrial.
- **Seized Land Location.** A categorical feature that contains whether the seized land is located in the city center or outside of it.
- **Seizure Purpose.** The purpose of seizing the state's land, whether to be seized for residential purpose or for a commercial one.
- **Seizure Date.** The date when this land is seized.

As these data contain more than one data type, it is important to preprocess these data to produce a more homogeneous dataset that is to be processed easier by the classifier. Although some classifiers may be able to process a dataset that contains numerical and categorical data, the dataset is preprocessed to convert the numerical features into categorical and classify some of the seizures into high and low risks.

3.2. DATA PREPROCESSING

In order to provide a useful dataset to the classifiers, so that these classifiers may extract knowledge from these data and apply the extracted knowledge to predict squatters with high-risk levels, the following tasks are applied prior to using the classifiers.

3.2.1. DATA LABELING

The training data are distributed into two classes, high and low. The high represents squatters with the high-risk level of seizing another land. To provide real examples of squatters with a high risk of seizing another land to the classifier for training purpose, the squatters who have more than one seized state's land are marked as high. It is important to mention that no squatters are found to seize more than two lands in the entire data set. Thus, the second seizure of the same squatter is deleted from the dataset, while the first seizure is considered to hold the characteristics of a high-risk level seizure.

The low-risk level seizures are selected based on the experience and the nature of the Iraqi society, which for example makes it difficult for a single woman to be living alone in a seized land. Only seizures with decisive characteristics are classified as low, while seizures that are not marked as high-risk level and have no decisive characteristics are left unclassified so that the classifier may be used to predict a class for these seizures. The summary of the collected data after labeling are shown in Table 3.1.

Table 3.1: Summary of collected data.

Category	Count
Total seizures	624
Total squatters	587
Labeled as high-risk squatter	37
Labeled as low-risk squatter	31
Total labeled squatters	68
Unlabeled squatters	519

3.2.2. IDENTIFICATION NUMBER REMOVAL

As the dataset has no squatters with more than two seized state's lands, and as the second seizures are deleted from the dataset, because these squatters are already considered to be with high-risk level of seizing another land, the squatter's identification number is now equal to the row number on that seizure in the dataset. This makes the squatter's identification number unique throughout the entire dataset and can be concluded using the row number of the seizure in the dataset. Thus, this feature has no role in the classification task and is removed from the dataset.

3.2.3. NUMERICAL DATA CONVERSION

There are three features in the dataset that hold numerical data, which are the squatter's age, squatter's family member and seizure date, are converted into categorical, where each feature consists of two categories. These features are converted as follows:

- **Squatter’s Age.** Squatters who are younger than forty years old are considered to be “younger”, while other squatters are considered to be “older”.
- **Squatters Family Members.** The number of family members that are supported by the squatter is converted into two categories that describe the size of the family. These categories are “big” for families with more than three members and “small” for families with less than four members.
- **Seizure Date.** Seizures prior to the latest Iraqi war against terrorism, which occurred in 2014 and caused massive forced displacements of families in the affected areas, are considered to be “old”, while seizures after that are considered to be “recent”.

3.3. DATA CLASSIFICATION

After preparing the dataset for the classification process, it is important now to select a classifier that can be used to predict a class for each seizure depending on the characteristics of that seizure, which contain information about the squatter and the seized land. A classifier’s performance may vary from one dataset to another. Thus, it is important to evaluate the performance of many classifiers on this dataset to select the classifier with the best performance.

3.3.1. BAYESIAN CLASSIFIERS

The use of Bayesian Networks has attracted significant attention according to the powerful probabilistic representation of these networks. These classifiers calculate the probability of a tuple to be in each class of the training data. Then, the class with the highest probability is predicted as the class that the tuple belongs to. Bayesian Networks represent the probability relationships among the features of the dataset in a graphical model. Then, the conditional dependencies are concluded from this graphical model using computational and statistical methods. Thus, Bayesian Networks are considered to combine the properties of statistics and computer science [22].

BayesNet is a classifier that uses Bayesian Networks to extract knowledge from a dataset with nominal features, which are categorical features that have no meaningful

order of the categories in it, [23]. The Bayesian Network represents the joint probability distribution of the features in the dataset by representing these features as nodes and the links between the nodes represent the direct influence of one feature on the other. A conditional probability distribution (CPD) is annotated to each node, where the CPD of a feature X_i , which has influence on the feature X_j , is written as $P(X_i|Pa(X_i))$, where $Pa(X_i)$ is the features that have influence over the feature X_i . A unique probability distribution of features joints is calculated from the directed acyclic graph (DAG) generated by the Bayesian Network is factorized as $P(X_1 \dots X_n) = \prod_i P(X_i|Pa(X_i))$.

Naïve Bayes classifier, on the other hand, uses the training data to calculate the probability based on the repetition of values in a feature and the repetition of combinations of values in different features in a specific class. This classifier selects the class C_i with the highest posterior probability $P(C_i|x)$, conditioned on the tuple X being classified, where $P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$, $P(X|C_i) = \prod_{k=1}^n P(x_k|C_i)$ and $P(C_i)$ is the ratio of the tuples in that class to the total number of tuples in the training dataset.

3.3.2. LAZY CLASSIFIERS

The lazy classifiers do not extract any knowledge from the dataset until the moment when classification process starts. Thus, the training dataset must always exist every time a tuple is classified and the knowledge is extracted from the training dataset for every classification process. Unlike other types of classifiers where knowledge is extracted and stored so that when a new tuple is to be classified, the extracted knowledge is directly applied to classify this tuple. The main advantage of the lazy classifiers is that the knowledge is updated within every classification process, while their main disadvantage is the time consumed to extract that knowledge in every classification process, when compared to other classifiers where the knowledge is stored and directly applied to classify the tuple.

k -Nearest Neighbor (k -NN) classifier is a lazy classifier that computes the similarity between the tuple being classified from one side, and the tuples in the training dataset from another side. Then, the (k) most similar tuples, which have the

highest closest distance to the tuple being classified, are used to predict a class for that tuple. The predicted class is selected based on the classes of the selected neighbors and the distance of each neighbor, where classes of closer tuples have more effect on the results than the classes of the farther ones. The number of selected neighbors (k) is predefined and the number of tuples in the training dataset may be limited so that when a new tuple is added, the oldest one is deleted.

The overall distance between two tuples is calculated using the Euclidean distance, which is calculated by

$$\text{Euclidean Distance } (X, Y) = \sqrt{\sum_{i=1}^n (X_i - Y_i)^2}$$

Where X and Y are the tuples that the distance between them is being calculated, which are consist of n attributes. This equation can be easily applied to numerical data, but the existence of nominal values in the datasets makes it difficult to use this equation with these data types. Thus, the humming distance is used with nominal values, which states that the distance between two tuples per each feature is equal to zero when these values are identical, and equal to one when these values are different. Mathematically, it is represented as

$$\text{Hamming Distance } (x, y) = \begin{cases} 0, & x = y \\ 1, & x \neq y \end{cases}$$

where x and y are the values of the same nominal feature in tuples X and Y . Thus the Euclidean distance of a nominal dataset is equal to

$$\text{Euclidean Distance } (X, Y) = \sqrt{\sum_{i=1}^n (\text{Hamming Distance}(X_i, Y_i))^2}$$

K-Star classifier is another lazy classifier that measures the similarity between the tuple being classified and the tuples in each class in the training dataset. The difference between the K-Star classifier and the k-NN classifier is that the K-Star classifier uses entropic means to calculate the distances between tuples [24], while the k-NN classifier uses the Euclidean distance for that purpose. The other difference is

that the K-Star classifier predicts a class for the tuple being classified based on the dominant class among the selected neighbors regardless of the distances of these tuples from the tuple being classified.

3.3.3. DECISION TREES

The classifiers in this type of classification have the capability to break down the complex task of classification into a set of simple conditions that can be applied to any tuple in order to predict a matching class for it, [20]. These conditions are arranged in the shape of a tree where higher priority conditions are closer to the root and condition with less priority are the leaves of the tree. Then, these conditions are applied from root to leaves on a tuple in order to predict a class for that tuple.

One of the popular algorithms to create classification decision tree using the C4.5 algorithm. This algorithm calculates the gain of each feature in the dataset using the Entropy, which is calculated using the equation

$$Entropy(\vec{y}) = - \sum_{j=1}^n \frac{|y_j|}{|\vec{y}|} \log_2 \frac{|y_j|}{|\vec{y}|}$$

where n is the number of categories in that feature, and the conditional entropy, which is computed by

$$Entropy(j|\vec{y}) = \frac{|y_j|}{|\vec{y}|} \log_2 \frac{|y_j|}{|\vec{y}|}$$

then, the gain is defined by

$$Gain(\vec{y}, j) = Entropy(\vec{y}) - Entropy(j|\vec{y})$$

Eventually, the tree is created based on the calculated gains, where features with higher gains are closer to the root of the tree. All the training dataset is used directly to conclude the classification decision tree.

A more complex classification method that is based on decision trees is the random forest. This forest consists of multiple trees that are generated by dividing the

training set into multiple subsets of data randomly and generate a decision tree using each subset separately. Then, when a new tuple is classified using this forest, each tree is used to conclude a class, and voting is used to select one class, among the concluded classes using each tree, as a predicted class for that tuple.

3.4. PERFORMANCE EVALUATION

Suppose there are two classifiers that are used to predict a class for a specific tuple by extracting knowledge from the same training dataset. Each classifier predicts a different class for that tuple with the same prediction confidence. This makes it mandatory to evaluate the performance of these classifiers in order to know which classifier has better performance, thus, its results are more accurate than other classifiers. To achieve that, classifier evaluation methods are proposed to measure the performance of these classifiers.

Performance measures represent how good this classifier's predictions are. This poses a problem that it is impossible to evaluate a future prediction. Because it is still unknown, up to the present, whether this future prediction is correct or not. Thus, to evaluate the performance of a classifier, it is applied to classified data, so that when a class is predicted, it is possible to evaluate this prediction. The classified data exist in the training dataset. Thus, part of the training dataset may be excluded from the knowledge extraction process and is used for evaluation purpose only, while it is recommended to use the entire training data for knowledge extraction, in order to extract all available knowledge, and then use them again for evaluation, to provide better evaluation resolution. To simplify the calculations of performance measure, all predicted and actual classes are summarized together in a confusion matrix, where predicted classes are distributed vertically and actual classes are distributed horizontally.

3.4.1. ACCURACY

Accuracy is one of the simplest performance measures that are widely used to evaluate the classifier's prediction results [25-27]. It measures the ratio of the correctly classified tuples to the total number of classified tuples. This provides a good

representation of how accurate the overall classification results. For example, a training dataset that contains T tuples is classified using classifier that is able to predict Y tuples correctly, the accuracy of this classifier is calculated as

$$Accuracy = \frac{Y}{T}$$

3.4.2. F-MEASURE

The F-measure provides a better illustration of the performance of the classifier in the classes rather than the overall performance as in the accuracy measure. The calculation of the F-measure is based on the precision and recall, which are calculated using the confusion matrix of the classification results [28]. The precision in a class represents the number of tuples that are correctly classified into that class to the total number of tuples predicted to be in that class. For example, a classifier predicted R tuples to be in class C . Only R_c are actually in this class, which means that these tuples are correctly classified by the classifier, then the precision in class C is

$$Precision_c = \frac{R_c}{R}$$

while the recall is the ratio of the number of correctly classified tuples to the total number of tuples in that class. Thus, if class C in the previous example contains C_Y tuples, then the recall of that class is calculated using

$$Recall_c = \frac{R_c}{C_Y}$$

Then, the F-measure of that class is

$$Fmeasure_c = \frac{2 * Recall * Precision}{Recall + Precision}$$

Finally, the overall F-measure is the average of the F-measures calculated for each class, while the weighted F-measure is the summation of the F-measures calculated for each class multiplied by the number of tuples in that class and divided by the number of total tuples in the dataset used for evaluation.

3.5. FEATURES SELECTION

The time required to predict a class for a tuple is proportional to the number of features in that tuple regardless of the method used for classification. There are many methods to evaluate the contribution of each feature in the classification process. In other words, the importance of each feature to the classifier, which is also known as the weight of the feature, may be measured, and some features that have very low, or no, contribution in the classification process may be eliminated prior to classification in order to reduce the required processing time [29]. For a better illustration of the feature weight, the sample dataset shown in Table 3.2 has two features. The first feature (A1) has no contribution in the classification of the attributes because each of the categories of this feature exists in both classes equally. The second feature (A2) has the highest possible contribution or weight, as it has one category value per each class. Thus, it is possible to rely on the second feature to classify any new tuples regardless of the value in the first feature.

Table 3.2: Sample dataset.

A1	A2	Class
A	C	1
B	C	1
A	D	2
B	D	2

3.5.1. INFORMATION GAIN METHOD

Information gain is one of the methods used to measure the contribution of a feature in the classification of the tuples. It computes the weight of the feature by measuring the possibility of classifying a tuple depending on that feature only. This is done by examining the distribution of each category in that feature over the existing classes. A category is considered to have higher weight when its existence in one class is more frequent than other classes [30]. The information gain of a category D in a feature from a dataset that contains m classes is calculated using the Entropy by

$$Info(D) = - \sum_{i=1}^m (p_i) \log_2(p_i)$$

where p_i is the ratio of the number of instances in class C_i that has the value of category D in that feature to the total number of tuples that have that category value in that feature. Then, the weighted average of all (v) categories in that feature is calculated using the equation

$$Info_A(D) = - \sum_{j=1}^v \frac{|D_j|}{|D|} \times info(D_j)$$

finally, the information gain by classifying the dataset based on the feature A is

$$Information\ Gain(A) = info(D) - info_A(D)$$

3.5.2. ONE-RULE METHOD

The one-rule method classifies a dataset using the values in one feature. Thus, although this method is proposed as a classifier, the methodology used for classification in this method makes it very representative to the feature's contribution in the classification process. Thus, it is used for features weighting by using this method to classify the training dataset and then use the confusion matrix of the classification results to calculate the accuracy of the classification. This accuracy is then used as a weight of the attribute as the attributes that result in better classification has more weight and higher classification accuracy [31].

CHAPTER 4

EXPERIMENTAL RESULTS

The experiments conducted in this study may be divided into two main stages depending on the goal set for these experiments. The first stage is to test the performance of the classifiers over the existing dataset in order to be able to select the classifiers with the better performance than the others for further testing. In stage two, the classifiers selected, in stage one, are used to select the best features selector that may be used to minimize the number of features used by the classifier without affecting the classification task. All experiments are executed using the WEKA¹ [32] version 3.8 running on a computer that has Intel® Core™ i5-4200U CPU @ 1.60GHz 2.30 GHz, 6.00 GB of memory and runs using Windows 7 operating system.

4.1. CLASSIFIERS' PERFORMANCE

In this section, the performances of six classifiers are measured using the existing dataset using the overall accuracy and F-measure as performance measures. The six classifiers are divided as two Bayesian classifiers, two lazy classifiers and two decision tree classifiers. The total number of classified tuples in the dataset are 68 tuples. These tuples are used for both training and testing in two different techniques. The first technique splits the dataset into two parts, one for training and the other is for performance evaluation. The other technique, which is known as cross validation, divides the dataset into a pre-defined number of sets, then, each subset is selected as a test set, while all other subsets are used for training, and eventually, the average of

¹ <https://www.cs.waikato.ac.nz/ml/weka/>

these iterations is calculated as the overall performance of that classifier, [33]. In this study, both techniques are used for more accurate results.

4.1.1. SPLIT DATA EVALUATION

The labeled dataset is split into two parts, where 80% of the tuples are selected randomly to be used for training and the remaining 20% are used for performance evaluation. Then, the classifiers are tested using this split and the performance of these classifiers is measured.

The classification results using the BayesNet classifier are summarized in the confusion matrix shown in Table 4.1, alongside with the performance measures calculations.

Table 4.1: Split data classification results using BayesNet classifier.

		Predicted		Precision	Recall	F-Measure	Accuracy
		High-Risk	Low-Risk				
Actual	High-Risk	8	1	0.8	0.89	0.84	
	Low-Risk	2	3	0.75	0.6	0.67	
Overall Average						0.75	0.79

The other Bayesian classifier that is tested using the split dataset is the Naïve Bayesian classifier. The results of the classification process and the calculations of the performance measures are shown together in Table 4.2.

Table 4.2: Split data classification results using Naïve Bayesian classifier.

		Predicted		Precision	Recall	F-Measure	Accuracy
		High-Risk	Low-Risk				
Actual	High-Risk	8	1	0.8	0.89	0.84	
	Low-Risk	2	3	0.75	0.6	0.67	
Overall Average						0.75	0.79

The k -NN classifier is an instance based classifier, where the number of instances, used by the classifier, is preset by the user. Regardless of the number of instances set to the classifier (k), the distance between the tuple being classified and every tuple in the training dataset are computed, then, the (k) closest tuples are selected

for the classification purpose. Thus, this classifier is tested using one and three nearest neighbors, as there are two classes in the training dataset and it is impossible to get a tie between the classes among the selected nearest neighbors. The results of using one nearest neighbor are shown in Table 4.3.

Table 4.3: Split data classification results using k -NN classifier with $k=1$.

	Predicted		Precision	Recall	F-Measure	Accuracy
	High-Risk	Low-Risk				
High-Risk	7	2	0.78	0.78	0.78	
Low-Risk	2	3	0.6	0.6	0.6	
Overall Average					0.69	0.71

Although the performance of the nearest neighbor classifier is relatively good, the use of only one neighbor is very sensitive to noise, where noise is defined as the tuples in the training dataset that has the characteristics of one class but they belong to another, [34]. This makes it important to increase the number of neighbors selected for classification. Thus, this classifier is tested using three nearest neighbors, and the test results are summarized in Table 4.4.

Table 4.4: Split data classification results using k -NN classifier with $k=3$.

	Predicted		Precision	Recall	F-Measure	Accuracy
	High-Risk	Low-Risk				
High-Risk	7	2	0.78	0.78	0.78	
Low-Risk	2	3	0.6	0.6	0.6	
Overall Average					0.69	0.71

The other lazy classifier is the K-Star classifier, which measures the one average distance between the tuple being classified and the tuples per each class in the training dataset, [24]. Thus, this lazy classifier requires no preset values. The classification results of using the K-Star classifier to classify the training dataset are shown in Table 4.5.

Table 4.5: Split data classification results using K-Star classifier.

		Predicted		Precision	Recall	F-Measure	Accuracy
		High-Risk	Low-Risk				
Actual	High-Risk	7	2	0.78	0.78	0.78	
	Low-Risk	2	3	0.6	0.6	0.6	
Overall Average						0.69	0.71

The performance of the K-Star classifier using the same training dataset is noticed to result in the exact same results as the k -NN classifier, which are marginally lower than the performance measures of the Bayesian classifiers.

Furthermore, the decision tree classifiers are tested using the training dataset in order to measure the performance of these classifiers. First, the J48 classifier, which is a C4.5 based decision tree classifier, is used to classify the dataset to measure its performance. The results of this experiment are summarized in Table 4.6 alongside with the performance measures calculated using these classification results.

Table 4.6: Split data classification results using J48 classifier.

		Predicted		Precision	Recall	F-Measure	Accuracy
		High-Risk	Low-Risk				
Actual	High-Risk	6	3	0.75	0.67	0.71	
	Low-Risk	2	3	0.5	0.6	0.55	
Overall Average						0.62	0.64

The decision tree created by the J48 classifier, which represents the knowledge extracted from the training dataset, is shown in figure 4.1. These rules are applied to any tuple in order to predict a class for that tuple, including the tuples in the test data.

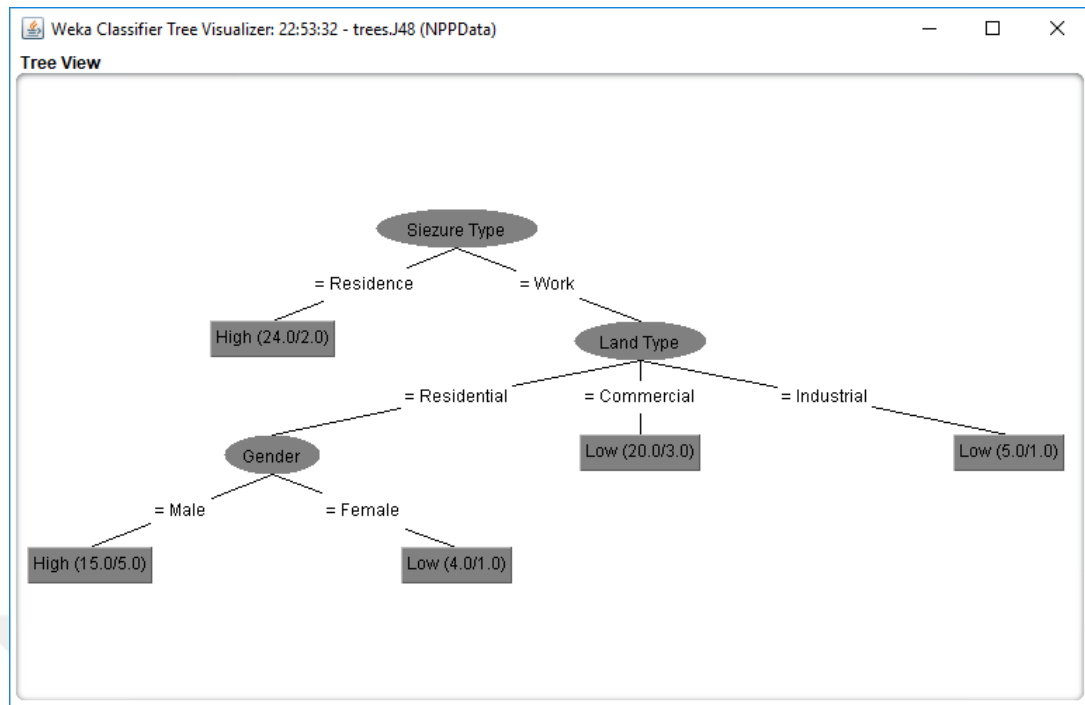


Figure 4.1: The decision tree created by the J48 classifier.

The random forest classifier, which consists of multiple decision trees is also tested using the same training dataset and the performance measures are evaluated using the classification results summarized in the confusion matrix shown in Table 4.7.

Table 4.7: Split data classification results using the Random Forest classifier.

		Predicted		Precision	Recall	F-Measure	Accuracy
		High-Risk	Low-Risk				
Actual	High-Risk	8	1	0.8	0.89	0.84	
	Low-Risk	2	3	0.75	0.6	0.67	
Overall Average						0.75	0.79

4.1.2. CROSS VALIDATION EVALUATION

The labeled dataset is divided into ten folds. Each fold is used once for performance evaluation and nine times for training. Thus, the classifier iterates

through these folds ten times. Then, the average performance of these ten classifications is calculated as the overall measure.

The classification results of the BayesNet classifier are shown in confusion matrix in Table 4.8. This confusion matrix is used to calculate the accuracy and F-measure of the BayesNet classifier using the cross-validation method.

Table 4.8: Cross-validation classification results using the BayesNet classifier.

		Predicted		Precision	Recall	F-Measure	Accuracy
		High-Risk	Low-Risk				
Actual	High-Risk	28	9	0.8	0.76	0.78	
	Low-Risk	7	24	0.73	0.77	0.75	
Overall Average						0.76	0.76

The Naïve Bayes classifier, which is also a Bayesian classifier is also tested using the cross-validation technique. The confusion matrix shown in table 4.9 summarizes the classification results and the performance measures for this classifier.

Table 4.9: Cross-validation classification results using the Naïve Bayes classifier.

		Predicted		Precision	Recall	F-Measure	Accuracy
		High-Risk	Low-Risk				
Actual	High-Risk	28	9	0.8	0.76	0.78	
	Low-Risk	7	24	0.73	0.77	0.75	
Overall Average						0.76	0.76

The lazy classifiers are then tested using the cross-validation technique. The results of cross-validation classification using the k -NN classifier, first using $k=1$, which is also known as the nearest neighbor classification, are shown in Table 4.10.

Table 4.10: Cross-validation classification results using k -NN classifier with $k=1$.

		Predicted		Precision	Recall	F-Measure	Accuracy
		High-Risk	Low-Risk				
High-Risk	30	9	0.73	0.77	0.75		
Low-Risk	11	22	0.71	0.67	0.69		
Overall Average						0.72	0.72

As the use of one nearest neighbor is very sensitive to noise in data, despite the high-performance measures that may result during tests. More reliable results are acquired using more than one neighbor for class prediction. Thus, the same data are used to test the performance of the k -NN classifier by selecting three neighbors instead of only one, to minimize the effect of noise on the results. Table 4.11 shows the confusion matrix of the results of this test and the performance measures calculated from these results.

Table 4.11: Cross-validation classification results using k -NN classifier with $k=3$.

		Predicted		Precision	Recall	F-Measure	Accuracy
		High-Risk	Low-Risk				
Actual	High-Risk	28	9	0.72	0.76	0.74	
	Low-Risk	11	20	0.69	0.651	0.67	
Overall Average						0.70	0.71

The other lazy classifier is the K-star classifier, the classification results are summarized in Table 4.12 alongside with the performance measures computed using these results.

Table 4.12: Cross-validation classification results using K-star classifier.

		Predicted		Precision	Recall	F-Measure	Accuracy
		High-Risk	Low-Risk				
Actual	High-Risk	27	10	0.71	0.73	0.72	
	Low-Risk	11	20	0.67	0.65	0.66	
Overall Average						0.69	0.69

The performance of the decision tree classifiers is also tested using the cross-validation technique. The performance summary, of the J48 classifier, is shown in the confusion matrix in Table 4.13 with the measures calculated for these results.

Table 4.13: Cross-validation classification results using J48 classifier.

		Predicted		Precision	Recall	F-Measure	Accuracy
		High-Risk	Low-Risk				
Actual	High-Risk	30	7	0.67	0.81	0.73	
	Low-Risk	15	16	0.70	0.52	0.59	
Overall Average						0.66	0.68

The performance of the random forest, which is also a decision tree based classifier, is also evaluated using this technique. The confusion matrix and the calculated performance measures are shown in Table 4.14.

Table 4.14: Cross-validation classification results using random forest classifier.

		Predicted		Precision	Recall	F-Measure	Accuracy
		High-Risk	Low-Risk				
Actual	High-Risk	27	10	0.73	0.73	0.73	
	Low-Risk	10	21	0.68	0.68	0.68	
Overall Average						0.70	0.71

4.2. FEATURES SELECTION

Next, features selectors are tested to evaluate the contribution of each feature in the classification process, and then, remove the features that have no effect on the classification results. The features in Table 4.8 are ordered according to their importance, where the first attribute has the highest contribution in the classification process, as resulted using the Information Gain method.

Table 4.15: Features ranking using Information Gain method.

Rank	Feature
0.267148	Land Type
0.249350	Seizure Type
0.097028	Job
0.060792	Has Children
0.056895	Residence Location
0.053198	Marital Status
0.036584	Origin
0.029575	Education
0.027840	Gender
0.025081	Seizure Location
0.024178	Family Size
0.004437	Age
0.002702	Has Residence
0.000744	Seizure Date

Next, the features that are suggested by this method to have the least effect on classification process are removed one feature at a time until a feature's removal affects most of the classifiers. The effect of removing these features are shown in Table 4.16 for the data split evaluation method.

Table 4.16: Effect of features removed based on Information Gain on classifiers evaluated using data split method.

Feature	BayesNet		Naïve Bayes		k-NN		K-Star		J48		Random Forest	
	F	Acc.	F	Acc.	F	Acc.	F	Acc.	F	Acc.	F	Acc.
None	0.754	0.786	0.754	0.786	0.689	0.714	0.689	0.714	0.626	0.643	0.754	0.786
Seizure Date	0.754	0.786	0.754	0.786	0.689	0.714	0.775	0.786	0.626	0.643	0.754	0.786
Has Residence	0.754	0.786	0.754	0.786	0.689	0.714	0.775	0.786	0.626	0.643	0.754	0.786
Age	0.754	0.786	0.754	0.786	0.689	0.714	0.708	0.714	0.626	0.643	0.754	0.786
Family Size	0.689	0.714	0.689	0.714	0.844	0.857	0.708	0.714	0.689	0.714	0.754	0.786
Seizure Location	0.689	0.714	0.689	0.714	0.689	0.714	0.689	0.714	0.689	0.714	0.754	0.786

The features elimination procedure is also repeated using all classifiers, and their performance is evaluated using the cross-validation method with 10 folds per classifier. The effect of the removal of each feature is summarized in Table 4.17.

Table 4.17: Effect of features removed based on Information Gain on classifiers evaluated using data cross-validation.

Feature	BayesNet		Naïve Bayes		k-NN		K-Star		J48		Random Forest	
	F	Acc.	F	Acc.	F	Acc.	F	Acc.	F	Acc.	F	Acc.
None	0.764	0.765	0.764	0.765	0.702	0.706	0.688	0.691	0.662	0.676	0.704	0.706
Seizure Date	0.779	0.779	0.764	0.765	0.718	0.721	0.686	0.691	0.662	0.676	0.689	0.691
Has Residence	0.779	0.779	0.764	0.765	0.718	0.721	0.674	0.676	0.662	0.676	0.719	0.721
Age	0.779	0.779	0.779	0.779	0.704	0.706	0.691	0.691	0.702	0.706	0.719	0.721
Family Size	0.765	0.765	0.750	0.750	0.778	0.779	0.750	0.750	0.696	0.706	0.761	0.765
Seizure Location	0.765	0.765	0.765	0.765	0.733	0.735	0.718	0.721	0.713	0.721	0.747	0.750

The features are, then, ranked using the One-Rule method. The ranks scored by the features are shown in descending order, where more important features are shown first, in Table 4.18.

Table 4.18: Features ranking using the One-Rule method.

Rank	Feature
79.41	Land Type
75.00	Seizure Type
66.17	Employment
64.70	Has Children
60.29	Gender
60.29	Marital Status
58.82	Seizure Location
58.82	Education
57.35	Residence Location
54.41	Origin
50.00	Age
48.52	Family Size
48.52	Has Residence
48.52	Seizure Date

Then, the same procedure, of testing the effect of features over each classifier's performance, is repeated depending on these rankings and features are removed from bottom to top, one feature per iteration, until a feature affects all classifiers' results when removed. The summary of this procedure is shown in Table 4.19.

Table 4.19: Effect of features removed based on One-Rule method using data split technique.

Feature	BayesNet		Naïve Bayes		k-NN		K-Star		J48		Random Forest	
	F	Acc.	F	Acc.	F	Acc.	F	Acc.	F	Acc.	F	Acc.
None	0.754	0.786	0.754	0.786	0.689	0.714	0.689	0.714	0.626	0.643	0.754	0.786
Seizure Date	0.754	0.786	0.754	0.786	0.689	0.714	0.775	0.786	0.626	0.643	0.754	0.786
Has Residence	0.754	0.786	0.754	0.786	0.689	0.714	0.689	0.714	0.626	0.643	0.754	0.786
Family Size	0.689	0.714	0.689	0.714	0.775	0.786	0.775	0.786	0.689	0.714	0.754	0.786
Age	0.689	0.714	0.689	0.714	0.775	0.786	0.708	0.714	0.689	0.714	0.754	0.786
Origin	0.689	0.714	0.689	0.714	0.775	0.786	0.708	0.714	0.689	0.714	0.754	0.786
Residence Location	0.626	0.643	0.626	0.643	0.844	0.857	0.708	0.714	0.689	0.714	0.844	0.857
Education	0.626	0.643	0.626	0.643	0.754	0.786	0.775	0.786	0.689	0.714	0.775	0.786

Moreover, using the ranks in Table 4.18, the response of the classifiers' performance is monitored by evaluating the performance measure at every removed feature using the cross-validation technique. The results are summarized in Table 4.20.

Table 4.20: Effect of features removed based on One-Rule method using cross-validation technique.

Feature	BayesNet		Naïve Bayes		<i>k</i> -NN		K-Star		J48		Random Forest	
	F	Acc.	F	Acc.	F	Acc.	F	Acc.	F	Acc.	F	Acc.
None	0.764	0.765	0.764	0.765	0.702	0.706	0.688	0.691	0.662	0.676	0.704	0.706
Seizure Date	0.779	0.779	0.764	0.765	0.718	0.721	0.686	0.691	0.662	0.676	0.689	0.691
Has Residence	0.779	0.779	0.764	0.765	0.779	0.779	0.674	0.676	0.662	0.676	0.719	0.721
Family Size	0.779	0.779	0.764	0.765	0.775	0.779	0.747	0.750	0.696	0.706	0.749	0.750
Age	0.765	0.765	0.750	0.750	0.778	0.779	0.750	0.750	0.696	0.706	0.761	0.765
Origin	0.778	0.779	0.764	0.765	0.778	0.779	0.765	0.765	0.696	0.706	0.763	0.765
Residence Location	0.778	0.779	0.778	0.779	0.793	0.794	0.808	0.809	0.713	0.721	0.777	0.779
Education	0.764	0.765	0.778	0.779	0.775	0.779	0.777	0.779	0.716	0.721	0.777	0.779

CHAPTER 5

DISCUSSION

The performance of a classifier may vary from one dataset to another. Thus, a classifier may outperform another over a certain dataset, but cannot compete with it in another, [35]. This fact makes it important to test many classifiers using a certain dataset in order to select the classifier that outperforms the other classifiers on that dataset. For a better illustration of the classifiers' performances, the performance measures are represented graphically in figure 5.1. This figure shows the performance superiority of the *BayesNet*, Naïve Bayes and Random Forest classifiers over the remaining classifiers when all the features of the dataset are used in the classification process and the classifiers' performances are evaluated using the data split method. Moreover, when these classifiers are evaluated using the cross-validation technique, the performance of most classifiers remains in the same order, despite the difference in the performance measures as absolute values, except the performance of the random forest which has a huge difference in the calculated measures. A better illustration of the performance measure is shown in figure 5.2.

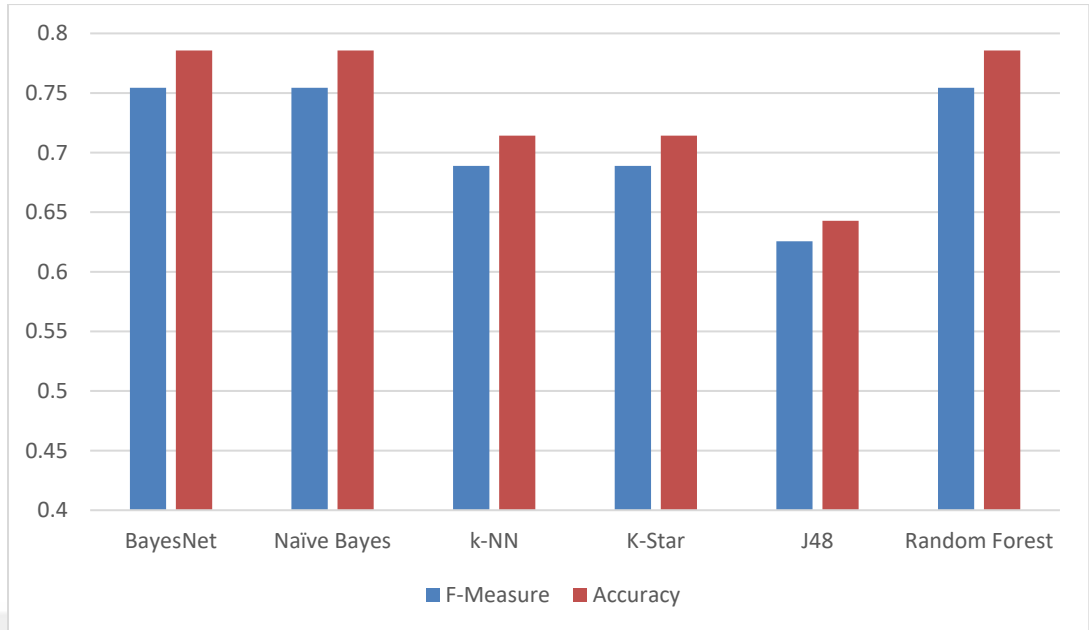


Figure 5.1: Classifiers' performance summary illustration for the data split technique.

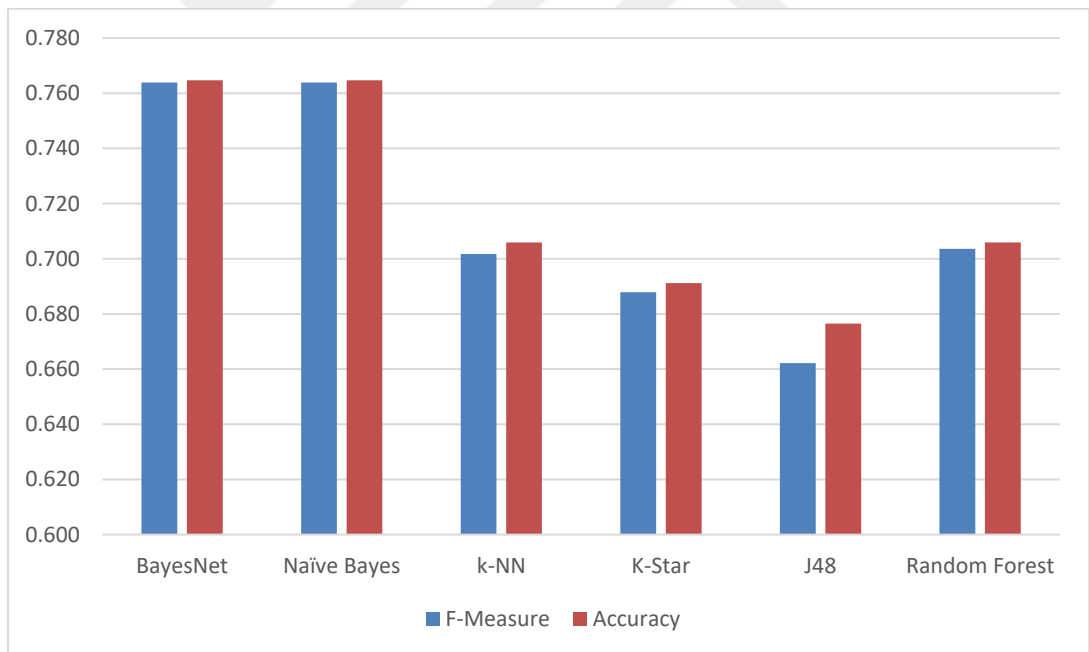


Figure 5.2: Classifiers' performance summary illustration for the cross-validation technique.

The performance of the k -NN classifier is very sensitive to the number of features used in the knowledge extraction, [36]. This sensitivity is according to the use

of hamming distance when the k -NN classifier is used with categorical data. Thus, the removal of low ranked features is mandatory to eliminate their useless effect over the distance measurement. This features removal also assists achieve faster classifications or predictions. Figure 5.3 shows the performance of the classifier, illustrated using the F-measures, when the features are removed one feature at a time, from the least ranked feature up, depending on the ranks calculated using the information gain features ranking method. In this figure, the classifiers are evaluated using data split method.

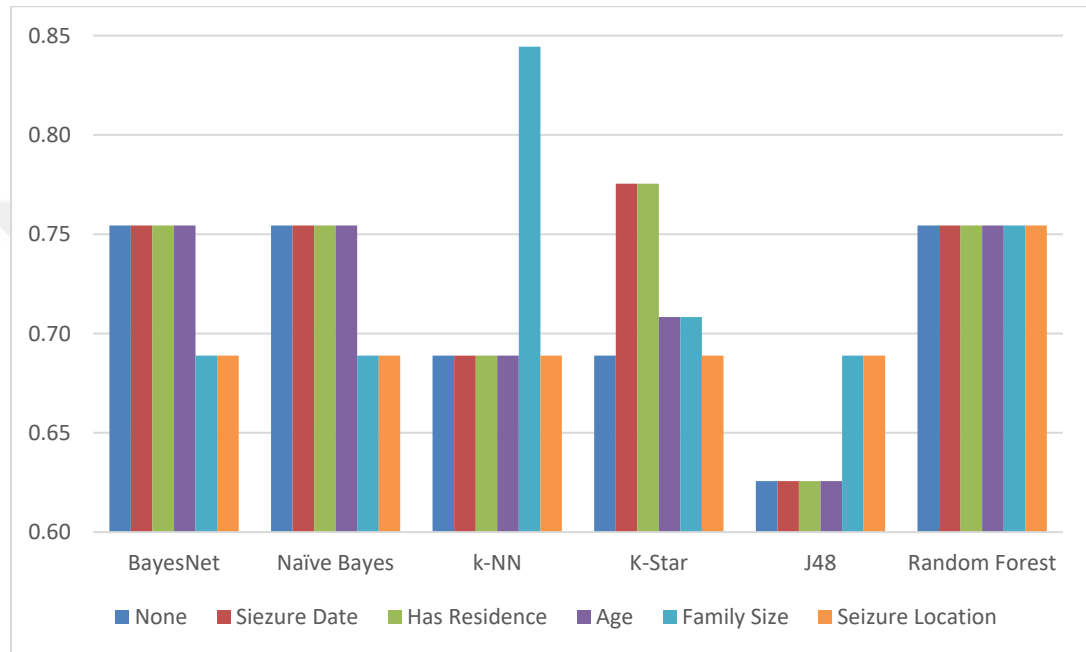


Figure 5.3: Features removal effect on the classifiers' F-measure using Information Gain ranking method and data split technique.

The illustration shows that the performance of the K-star method is improved as soon as the first feature “Seizure Date” is removed, but this improvement lasts only until the third feature “Squatter’s Age” is removed. The performance of the k -NN classifier is not affected by the removal of the first three feature, which are the “Seizure Date”, “Has Residence” and “Age”, but a good significant improvement is achieved by the removal of the fourth feature “Family Size” then it drops down again when the next feature “Seizure Location” is removed. Eventually, by removing the fifth least ranked feature “Seizure Location”, the performance of most classifiers drops. Thus, this feature is considered to have a noticeable effect over the actual classification of the data, and no further features are removed.

By evaluating the performance of the classifiers while removing the features, from least ranked up, using the cross-validation method. The performance of the classifiers affected narrowly until, again, the fourth feature “Family Size” is removed, where the performance of the lazy classifiers as well as the random forest classifier spiked for this test only and drop back when the next feature “Seizure Location” is removed. The evaluation of each classifier’s performance is illustrated graphically in figure 5.4.

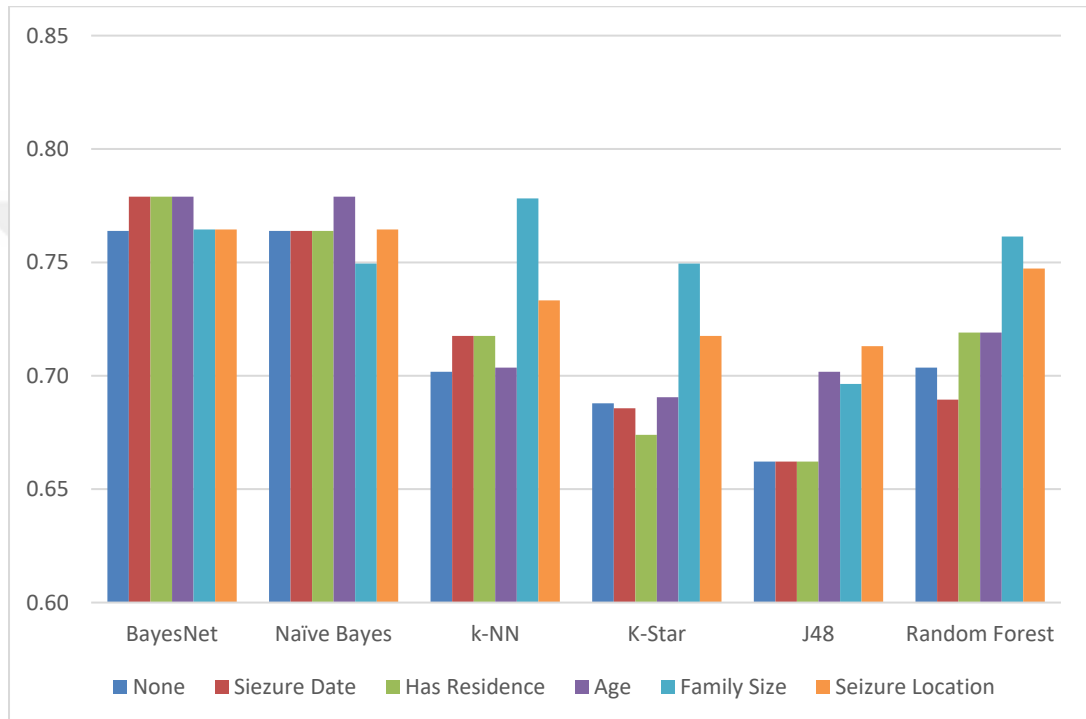


Figure 5.4: Features removal effect on the classifiers' F-measure using Information Gain ranking method and cross-validation performance evaluation technique.

The same procedure is repeated using the ranks calculated for the features using the One-Rule method. The response of the classifiers’ performance is illustrated in Figure 5.5 using the F-measure as an evaluation method of the performance, which is calculated using the data split method.

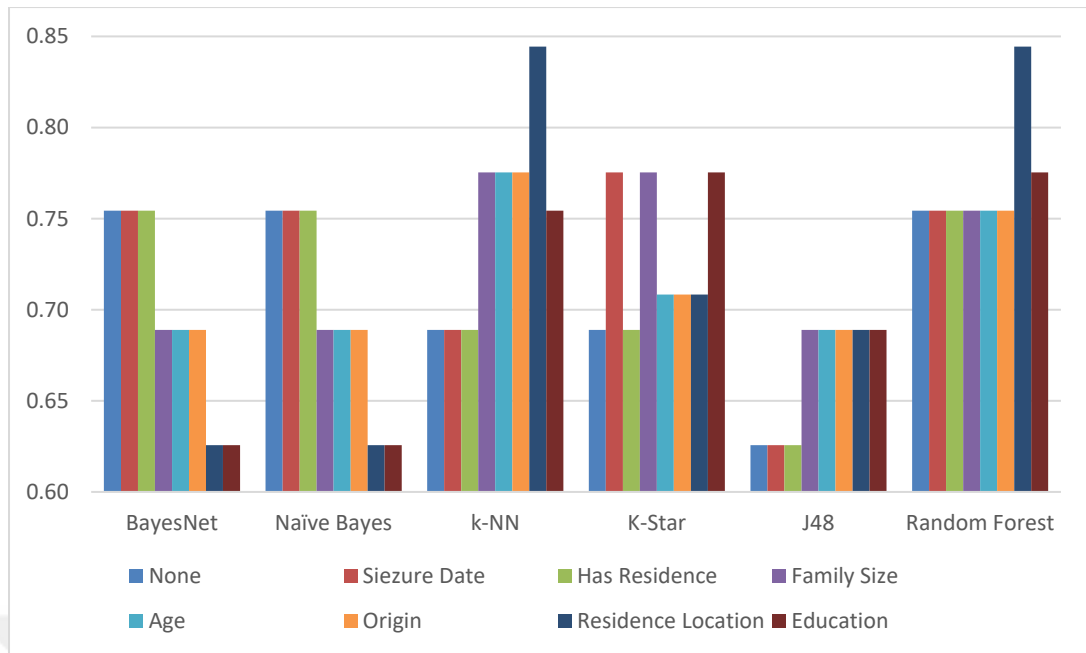


Figure 5.5: Features removal effect on the classifiers' F-measure using One-Rule ranking method and data split method.

The performance of the K-star classifier is noticed to oscillate up and down during the features removal of the features, while the performance of the *k*-NN classifier is improved as the third feature “Family Size” is removed to reach the highest performance among all classifiers until up to the sixth feature “Squatter’s Residence Location: is removed, where the performance of the *k*-NN is improved again, as well as the performance of the random forest classifier, and achieve the highest performance measures. By the removal of the next feature “Education”, the performance of both the *k*-NN and random forest classifiers drops down to lower level.

Moreover, the results of classifiers’ performances evaluation using the cross-validation method, which are illustrated in figure 5.6, show that the performance of the *k*-NN classifier is enhanced by the removal of the two least ranked features, using the One-Rule method, and remains relatively stable, on the top of the performance of other classifiers, until the seventh least ranked feature “Education” is removed, where the performance of most of the tested classifiers drops down.

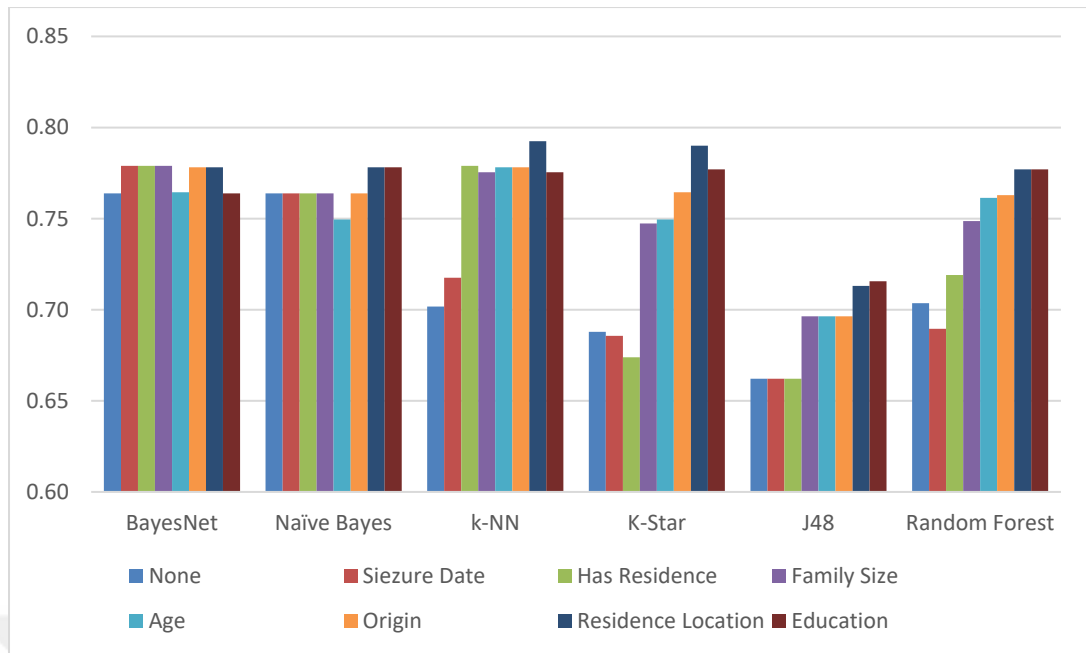


Figure 5.6: Features removal effect on the classifiers' F-measure using One-Rule ranking method and cross-validation method.

Although the random forest classifier, as well as the lazy classifiers, scored relatively highest performance measures, when the six least ranked features are removed based on the One-Rule ranking method, the random forest classifier is based on generating decision trees using the training dataset, [17]. These trees are static, which means that they are generated once and not updated unless the system administrator forces a new training. Thus, such kind of methods are not ideal for similar situations, to the situation being studied, as the human behavior may change according to specific conditions and the surrounding environment. On the other hand, it is important to test these classifiers in order to make sure that the selected classifier is not outperformed by any other classifiers, or in other words, the selected classifier has an acceptable performance.

The other two competitive classifiers, according the performance measures, shown in figure 4.6, are the *k*-NN and the K-Star classifiers. These classifiers are lazy classifiers, which means that they do not start the training task, until a class prediction is required [37]. They are also instance based classifiers, which means that they rely on the instances in the training dataset to predict a class for the new tuple, every time

a new tuple is required to be classified [38, 39]. These two features make the lazy classifiers more dynamic, or more flexible, to the change in the behavior of the new tuples added to the training dataset.

As the k -NN classifier chooses the (k) most similar tuples to the new tuple being classified in order to predict a class for that tuple [40], when the behavior of the squatters changes, the prediction of the classifier changes after the k^{th} squatter, similar to the findings in [41]. Moreover, the K-Start classifier uses the average distance of the new tuple with all tuples in each class [42], in order to predict a class for the new tuple, which means that the prediction for a new tuple is changed only when the average changes, or in other words, when the new behavior becomes more frequent than the old behavior. Thus, the use of the k -NN classifier to build a model is more suitable for the case being studied as it scored the highest performance measure and is dynamic and more flexible to behavior change than other classifiers.

The k -NN classifier, which is selected for method implementation, is one of the classifiers that have achieved high-performance measures until the seventh least ranked feature “Education”, using the one-rule features selection method, is removed, while the performance measure spikes only when the fourth feature “Family Size” is removed and drops back when the fifth feature “Seizure Location”, using the information gain method, is removed. This shows that the features selectors may have different performances depending on the training dataset and it is important to test more than one feature selector on each dataset as there is no absolute better feature selector and it is always related to the training dataset as suggested in [29]. The performance measures response of the k -NN classifier to the features

In the study conducted by [4] in Yemen, males are found to be 92.4% of the squatters. While in this study, 88.6% of the squatter are males, but 89.2% of the high-risk squatters are males. They have also found that the squatters younger than forty years old are 54.2% in Yemen, while in this study, squatters younger than forty years old are found to be 85% of the squatters and 91.9% of the high-risk squatters.

The features with the three highest ranking, when ranked using the One-Rule method, are the land type, seizure type, and employment. Thus, these features are

considered very important in making decisions if the squatter is going further to seize another land or not. Seizing residential lands are found to be 75.1% of the seized lands, while 89.2% of the lands that had another seizure followed seizing them are found to be residential. Thus, it is more likely that a squatter who is seizing a residential land to be high-risk of seizing another land. Furthermore, seizures are found to be for residential purposes in 43.6% of all seizures, while this percentage goes higher in high-risk seizures to reach 59.5% of them. Moreover, the employment feature has 32% of the seizures are made by unemployed squatters, while the high-risk seizures have 43.2% of them made by unemployed squatters. This distribution of percentages shows that the less the ranking of a feature the more similar the percentage of the values in that feature when measured with respect to all tuples, and when measured with respect to the tuples in a specific class.

CHAPTER 6

METHOD IMPLEMENTATION

A method is implemented for data storage, management and risk assessment of squatters upon entering seizures data. This method assists municipalities, like Baquba municipality, to keep the data organized, and enables modifying certain information without the need to access the tables used to store these data, which reduces the risk of mistakenly modifying information that is not supposed to be modified when tables are edited manually. Three tables are used to store the required data. The first table stores the seizures data, while the second table stores the training data used by the classifier, and the third table stores the status of processing the high-risk squatters. These tables are stored and retrieved using structured query language (SQL), while the user graphical interfaces and the codes to execute the SQL commands are developed using C# programming language and visual studio development environment.

The seizures table stores all the information about the squatter and the seized land at the moment of land seizure. These data are entered using a graphical user interface (GUI), which is shown in Figure 6.1, after entering the national ID of the squatter. If the ID of the squatter already exists in that database, the data of that squatter are already filled based on the information retrieved from the stored data. As this is a new seizure, the retrieved data are only related to the squatter, and the old seizure data are not retrieved as they are not important. The data filled to the interface are editable, as it is possible that some data are changed regarding the squatter. For example, a single squatter gets married, or the employment of the squatter is changed.

The screenshot shows a 'New Seizure' form with the following fields and values:

- National ID: 526
- Gender: Male
- Birth Date: 1980-06-10
- Marital Status: Divorced
- Family Members: 3
- Has Children:
- Origin: Same
- Education: High School
- Employment: Private Company
- Purpose: (empty dropdown)
- Has Residence:
- Residence Location: (empty dropdown)
- Land Type: (empty dropdown)
- Seizure Date: (empty text field)
- Location: (empty text field)
- City Center:

Buttons: Save, Cancel

Figure 6.1: New seizure interface for an existing squatter.

When a new land is seized by an existing squatter, the method checks first if this squatter is previously predicted as a high-risk squatter or not. If so, the method then checks the table of high-risk squatters to see if this squatter is processed or not. Then, the status of the squatter and the earlier prediction are shown to the user of the method. It also notifies the user whether this squatter is processed earlier or not. Later, the data of the previous seizure are retrieved and modified in order to be added to the table that holds the data used by the classifier to predict classes for the new squatters.

The training data table contains only the information used by the classifier. In other words, according to Figure 5.5, six least ranked features in Table 4.18 are neither included in the training data table nor in the prediction process, for the reasons discussed in Section 3.5. Thus, when a new seizure of an existing squatter is entered, the eight most ranked features of the older seizure are added to the training data table and labeled as high-risk to be used by the classifier for future prediction.

The method also keeps tracks of the high-risk squatters. When a new seizure is entered, where the squatter who seized that land has no earlier seizures, the method uses the k -NN classifier to predict a class for that new squatter. If the squatter is predicted to be a high-risk squatter, the method notifies the user and automatically add this squatter to the high-risk squatters' table. But as this is the first seizure of this squatter, no data are added to the training data table.

When a squatter is processed, the national ID of that squatter is entered to the method. This squatter is marked as processed in the database, and no longer shown in the unprocessed high-risk squatters. If a processed squatter seizes another land, the user of the system is notified. A summary of the number of existing seizures, total number of squatters, number of actual and predicted high-risk squatter and the number of processed high-risk squatters are shown in the main interface of the implemented method.

The implemented model is used to classify the remaining squatters in the existing dataset. These squatters have only one seized land and experts could not make a decisive call whether each squatter is to be high- or low-risk. The classification of these squatters shows that 395 (76.10%) of the squatters are high-risk of going further toward seizing another land, while the remaining 124 (23.90%) are predicted to be low-risk of seizing another land, and thus, no actions are needed to be taken toward them. These predictions are suggested to the decision makers in order to make the appropriate legal actions to minimize the effect of adverse possession on the municipality of Baquba.

CHAPTER 7

CONCLUSION

Adverse possession is known in most of the countries all over the world. Many countries have a set of laws that regulate these possessions. Adverse possession of state's lands, on the other hand, is more common in developing, especially when such countries go through certain situations, like wars and disasters. In Iraq, the adverse possession is a very old problem faced by many governments that ruled Iraq over the history. Moreover, the abnormal conditions that Iraq is going through have increased the rate of state's lands adverse possession dramatically, which creates many challenges to the municipalities of the cities where such possession occurs.

The increased number of squatters, who are seizing state's lands in order to adversely possess these lands, has a significant bad influence on the city because these seized lands are not being used for the purpose they are planned to be used for. Thus, these lands usually suffer from the lack of the necessary infra structure, which encourages the squatters to overload the infra structure of the adjacent regions. Thus, squatters are not only living in non-acceptable conditions, but they are also affecting the people who are living in the regions where adverse possession is occurring.

Using the data collected from the municipality of Baquba, which is the main city of Diyala governorate in Iraq, for the state's land being seized by squatters for the purpose of adverse possession, the characteristics of squatters are studied in order to find the features that are mostly lead to a state's land seizure. These features are used predict the risk level of the squatters, whether they are going to go further and seize another land or not. These features may also be studied by other authorities in order to take special care of them and consequently reduce the rate of adverse possession.

Six classifiers from three classification schemes, Bayesian, Lazy and decision trees, are tested using these data. The training dataset is collected from the existing seizures' data, where all the seizures, except the last one, of a squatter who has more than one seizure are considered to be of high-risk, while some selected seizure, based on the experience and the traditions of the Iraqi society, are considered to be of low-risk. These data are divided into two parts, the first part is fed to the classifier for training, then the other part is used for evaluation, where the classifiers are used to predict class for each tuple, then by comparing the original class with the predicted class, the performance is evaluated using the accuracy and F-measure.

Then, the features are ranked according to their contribution to the classification using two methods, the information gain and one-rule methods. The performance of the classifiers is monitored after the removal of each feature, starting from the least ranked feature toward the highest. The classifiers with the best performance maintained their performance measure even when the lowest four features, which are ranked using the information gain method, are removed, while these classifiers maintained their performance even when six least ranked features, which are ranked using the one-rule method, are removed.

The k -NN classifier and the K-star classifier are two of the classifiers that achieved good performance measures when tested using this dataset. The k -NN classifier predicts a class for a new tuple depending on the (k) tuples in the training dataset that most similar to the new tuple being classifier. the K-star classifier, on the other hand, measures the average distance between the new tuple from one side and the tuples in each class from another, then, the class with the least average is predicted as a class for that new tuple. Thus, the k -NN classifier is more dynamic than the K-star classifier, as the predictions in the k -NN update after the k^{th} tuple with the new pattern is inserted in the training dataset. Such dynamic behavior is more similar to the human behavior than other classifiers. Thus, the k -NN classifier is selected to develop a model that assists the municipalities to store manage and predict the seizures and squatter's behavior. Such a method enables the municipalities to control the rate of lands seizures within their available resources.

In future work, it is recommended to study the effect of each feature over different circumstances by studying the order of the ranks of these features in different conditions.



REFERENCES

1. **Ballantine, H. W.** (1918), "Title by Adverse Possession", *Harvard Law Review*, vol. 32, pp. 135-159.
2. **Kim, J.-Y.** (2004), "Good-faith error and intentional trespassing in adverse possession", *International Review of Law and Economics*, vol. 24, pp. 1-13.
3. **Fennell, L. A.** (2006), "Efficient Trespass: The Case for Bad Faith Adverse Possession", *Northwestern University Law Rev.*, vol. 100, p. 1037.
4. **Khan, M. A. I. and Mustafaa, R. A.** "A STUDY ON THE IMPACTS OF SQUATTER IN ADEN CITY, YEMEN", Aden: University of Aden.
5. **Link, J.** (2005), "Land Registration and Property Rights in Iraq", USAID Iraq Local Governance Program: USAID Iraq Local Governance Program.
6. **Han, J., Pei, J. and Kamber, M.** (2011), "Data mining: concepts and techniques", Elsevier: Elsevier.
7. **Sokolova, M., Japkowicz, N. and Szpakowicz, S.** (2006), "Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation", in *Australian conference on artificial intelligence*, pp. 1015-1021.
8. **Helmholz, R. H.** (1983), "Adverse possession and subjective intent", *Washington University Law Quarterly*, vol. 61, p. 331.
9. **Nawagamuwa, A. and Viking, N.** (2003), "Slums, squatter areas and informal settlements—do they block or help urban sustainability in developing Contexts", in *9th International Conference on Sri Lanka Studies*.
10. **Srinivas, H.** (2005), "Defining squatter settlements", *Global Development Research Center Web site*, www.gdrc.org/uem/define-squatter.html, viewed, vol. 9,
11. **Morawetz, J.** (2011), "No Room for Squatters: Alaska's Adverse Possession Law", *Alaska L. Rev.*, vol. 28, p. 341.

12. **Manaster, K. A.** (1968), "The problem of urban squatters in developing countries: Peru", *Wisconsin Law Rev.*, p. 23.
13. **Song, C., Qu, Z., Blumm, N. and Barabási, A.-L.** (2010), "Limits of predictability in human mobility", *Science*, vol. 327, pp. 1018-1021.
14. **Alahi, A., Ramanathan, V., Goel, K., Robicquet, A., Sadeghian, A. A., Fei-Fei, L. and Savarese, S.** (2017), "Learning to predict human behaviour in crowded scenes", *Group and Crowd Behavior for Computer Vision*, p. 183.
15. **Weber, B. G. and Mateas, M.** (2009), "A data mining approach to strategy prediction", in *Computational Intelligence and Games, 2009. CIG 2009. IEEE Symposium on*, pp. 140-147.
16. **Nikam, S. S.** (2015), "A comparative study of classification techniques in data mining algorithms", *Oriental Journal of Computer Science & Technology*, vol. 8, pp. 13-19.
17. **Ashari, A., Paryudi, I. and Tjoa, A. M.** (2013), "Performance comparison between naïve bayes, decision tree and k-nearest neighbor in searching alternative design in an energy simulation tool", *Int. J. Adv. Comput. Sci. Appl*, vol. 4, pp. 33-39.
18. **Muralidharan, V. and Sugumaran, V.** (2012), "A comparative study of Naïve Bayes classifier and Bayes net classifier for fault diagnosis of monoblock centrifugal pump using wavelet analysis", *Applied Soft Computing*, vol. 12, pp. 2023-2029.
19. **Vijayarani, S. and Muthulakshmi, M.** (2013), "Comparative analysis of bayes and lazy classification algorithms", *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, pp. 3118-24.
20. **Safavian, S. R. and Landgrebe, D.** (1991), "A survey of decision tree classifier methodology", *IEEE transactions on systems, man, and cybernetics*, vol. 21, pp. 660-674.
21. **Wena, C.-H., Hsu, P.-Y., Wang, C.-Y., Wuc, T.-L. and Hsu, M.-J.** (2012), "E-government Information Application: Identifying Smuggling Vessels with Data mining Technology", *Electronic Journal of e-Government*, vol. 10,
22. **Langley, P., Iba, W. and Thompson, K.** (1992), "An analysis of Bayesian classifiers", in *Association for the Advancement of Artificial Intelligence*, pp. 223-228.
23. **Boriah, S., Chandola, V. and Kumar, V.** (2008), "Similarity measures for categorical data: A comparative evaluation", in *Proceedings of the 2008 SIAM International Conference on Data Mining*, pp. 243-254.

24. **Tejera Hernández, D. C.** (2015), "An Experimental Study of K* Algorithm", *International Journal of Information Engineering & Electronic Business*, vol. 7,
25. **Foody, G. M.** (2002), "Status of land cover classification accuracy assessment", *Remote sensing of environment*, vol. 80, pp. 185-201.
26. **Jain, A. and Zongker, D.** (1997), "Feature selection: Evaluation, application, and small sample performance", *IEEE transactions on pattern analysis and machine intelligence*, vol. 19, pp. 153-158.
27. **Congalton, R. G., Oderwald, R. G. and Mead, R. A.** (1983), "Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques", *Photogrammetric Engineering and Remote Sensing*, vol. 49, pp. 1671-1678.
28. **Lewis, D. D. and Gale, W. A.** (1994), "A sequential algorithm for training text classifiers", in *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 3-12.
29. **Dash, M. and Liu, H.** (1997), "Feature selection for classification", *Intelligent data analysis, Elsevier*, vol. 1, pp. 131-156.
30. **Yang, Y. and Pedersen, J. O.** (1997), "A comparative study on feature selection in text categorization", in *International Conference on Machine Learning*, pp. 412-420.
31. **Holte, R. C.** (1993), "Very simple classification rules perform well on most commonly used datasets", *Machine Learning, Springer US*, vol. 11, pp. 63-90.
32. **Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H.** (2009), "The WEKA data mining software: an update", *SIGKDD Explor. Newsl.*, vol. 11, pp. 10-18.
33. **Kohavi, R.** (1995), "A study of cross-validation and bootstrap for accuracy estimation and model selection", in *International Joint Conference on Artificial Intelligence*, pp. 1137-1145.
34. **Bay, S. D.** (1999), "Nearest neighbor classification from multiple feature subsets", *Intelligent data analysis*, vol. 3, pp. 191-209.
35. **Demšar, J.** (2006), "Statistical comparisons of classifiers over multiple data sets", *Journal of Machine learning research*, vol. 7, pp. 1-30.
36. **Li, T., Zhang, C. and Ogihara, M.** (2004), "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression", *Bioinformatics*, vol. 20, pp. 2429-2437.

37. **Veloso, A., Meira Jr, W. and Zaki, M. J.** (2006), "Lazy associative classification", in *Data Mining, 2006. IEEE International Conference on Data Mining series, ICDM'06. Sixth International Conference on*, pp. 645-654.
38. **Aha, D. W., Kibler, D. and Albert, M. K.** (1991), "Instance-based learning algorithms", *Machine learning. Springer US*, vol. 6, pp. 37-66.
39. **Brighton, H. and Mellish, C.** (2002), "Advances in instance selection for instance-based learning algorithms", *Data mining and knowledge discovery*, vol. 6, pp. 153-172.
40. **Peterson, L. E.** (2009), "K-nearest neighbor", *Scholarpedia*, vol. 4, p. 1883.
41. **Lathia, N., Hailes, S. and Capra, L.** (2008), "kNN CF: a temporal social network", in *Proceedings of the 2008 ACM conference on Recommender systems*, pp. 227-234.
42. **Cleary, J. G. and Trigg, L. E.** (1995), "K*: An instance-based learner using an entropic distance measure", in *Proceedings of the 12th International Conference on Machine learning*, pp. 108-114.