IDENTIFYING THE FACTORS INFLUENCING TRUANCY AND USING DATA MINING FOR PREDICTION

HAYDER ABULABBAS WAHEED AL-BAYATI

SEPTEMBER 2017

IDENTIFYING THE FACTORS INFLUENCING TRUANCY AND USING DATA MINING FOR PREDICTION

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES OF
ÇANKAYA UNIVERSITY

BY

HAYDER ABULABBAS WAHEED AL-BAYATI

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

SEPTEMBER 2017

Title of Thesis: **IDENTIFYING THE FACTORS INFLUENCING TRUANCY AND USING DATA MINING FOR PREDICTION**

Submitted by **Hayder Abdulabbas Waheed**

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

Prof. Dr. Can ÇOĞUN
Director of Institute

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

Prof. Dr. Erdoğan DOĞDU
Head of Computer Engineering 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.

Assoc. Prof. Dr. James Little
Supervisor

Examination Date: 07.09.2017

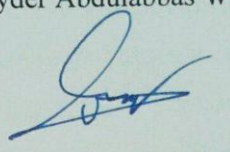
Examining Committee Members

Assist. Prof. Dr. Özgür Tolga Pusatli	(Çankaya Univ.)
Assoc. Prof. Dr. James Little	(Çankaya Univ.)
Assoc. Prof. Dr. Ayşe Collins	(Bilkent Univ.)

STATEMENT OF NON-PLAGIARISM

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, Surname : Hayder Abdulabbas Wheed

Signature : 

Date : 07.09.2017

ABSTRACT

IDENTIFYING THE FACTORS INFLUENCING TRUANCY AND USING DATA MINING FOR PREDICTION

AL-BAYATI, Hayder Abdulabbas Wheed

M.S., Information Technology Department

Supervisor: Assoc. Prof. Dr. James LITTLE

September 2017, 39 pages

Truancy from school is a problem in most societies, including developed societies. This research presents the process of discovering the underlying patterns behind truancy through the use of informed discussions and data mining techniques. The main objective of the study is to identify pupils at risk of truancy, before it occurs. Several models were built using different; classification algorithms. K=1-nearest neighbour showed the highest accuracy across all algorithms. The research also showed the importance of the new factor, security situation in identifying pupils at risk of truancy.

Keywords: truancy, classification, data mining

ÖZ

YOKLUĞUNU ETKİLEYEN FAKTÖRLERİN BELİRLENMESİ VE TAHMİN İÇİN VERİ MADENCİLİĞİNİ KULLANMA

AL-BAYATI, Hayder Abdulabbas Wheed

Yüksek Lisans, Bilgi Teknolojileri Anabilim Dalı

Tez Yöneticisi: Doç. Dr. James LITTLE

Eylül 2017, 39 sayfa

Okuldan kaçma gelişmiş toplumlar da dahil çoğu toplumdaki bir problemdir. Bu araştırma, bilinçli müzakere ve veri tarama tekniklerinin kullanılmasıyla okuldan kaçmanın arkasındaki temel şablonları bulma prosesini sunmaktadır. Çalışmanın ana amacı, meydana gelmeden önce okuldan kaçma riski olan öğrencileri tespit etmektir. Farklı sınıflandırma algoritmaları kullanarak birden fazla model geliştirilmiştir. K=1-en yakın komşu tüm algoritmalarda en yüksek doğruluğu göstermiştir. Araştırma ayrıca, okuldan kaçma riski olan öğrencilerin tespitinde güvenlik durumunun yeni bir faktör olarak önemini de göstermiştir.

Anahtar kelimeler: okuldan kaçma, sınıflandırma, veri tarama

ACKNOWLEDGEMENTS

I first extend my thanks to my supervisor Dr. James Little the Mathematics Department at Çankaya University, who spared no effort in guiding this thesis, who was always ready question about my research and advise me. I would also like to thank all the esteemed teachers from the secondary school and elementary school, school principals, teachers who provided me facilities to access the schools, as well recap their experiences and educational experiences to the success and support my thesis. I would particularly like to thank, Dr. Azhar Al-Saffar. Measuring & Evaluating, Amena-AL-Zubaidi, Dr. Sayed Al-Kaabi. Psychological Counseling, Aieda Hussein Al-Rubaie, Rahim Al-Kaabi, Mohammed Khalaf Al-Muhammadawi, Wafa Bader AL- Delphi, Ayad Al-Lami, Mohammed Madhi Al-Zubaidi, Ali Sattar Al-Adli.Master in Psychological Counseling & Educational, Ahmed Salam Al-Qaisi. To the source of life and love my mother dear, may God prolong your life that did not stop praying to me. Finally, to my lovely wife and my brothers and my friends and all my professors who supported me with their support and their prayers, through them I become more confidence and determination to succeed.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURES

LIST OF TABLES

TABLES

LIST OF ABBREVIATIONS

DM	Data Mining
DT	Decision Tree
NB	Naive Bayes
SMO	Sequential Minimal Optimization
IBK	Instance Based Knowledge

CHAPTER 1

INTRODUCTION

1.1 Background

Truancy from school, or more precisely, truancy without an excuse, has become a problem from which most countries suffer including modern countries such as America, Great Britain and Canada [1]. In developing countries, such as Iraq, where there is currently no clear education strategy or good security, these problems become more acute.

Truancy is not a new problem; it has existed in schools for decades. There are hundreds of thousands of pupils who are truant every year [2] and the rate of truancy can even approach 30% per day [3] in a number of places. Truancy can also be an indicator of a pupil's involvement in crime, in the present or in the future [4]. Clear strategies should be developed for those who commit truancy by monitoring the factors affecting truancy, such as the family, economy, etc. [5].

Understanding the possible underlying causes of truancy is essential to resolving it before, rather than after, it occurs.

Truancy is defined as abnormally frequent interruptions for the pupil from school due to the influence of one or more factors that will be addressed in the following sections.

1.2 Effects of truancy

The continuing absence of a pupil from school can lead to many negative consequences. These include the following:-

1. On the economic level, the absence of the pupil can cost the state in terms of economic loss to book printing companies and to teachers for salary. In the United States, one city estimated revenue loss of $-3.2 billion dollars due to truancy from the school system, and excess of $400 million in social services [Catterall, J.S. 6], which will damage the economy in the long term should it continue.

2. In terms of security, pupils leave school may associate with crime gangs [4], which can lead to instability in the community.

3. The pupil will not keep pace with a lesson, which in turn creates a state of antipathy towards school and spreads illiteracy within society. Education is the only effective tool to eliminate illiteracy in societies [7] and it is through education that societies develop.

1.3 Action against truancy

It is often the effects of truancy which are addressed rather than the causes. Treatment can come too late after truancy has occurred. In the United States, laws have been passed nationally to make school attendance mandatory for all ages, rather than optional for some ages in some states [8]. In number of cases, parents may face fines or imprisonment and possibly even loss of custody of children due to neglect [9]. Parents can also be made to attend parenting classes through legal contracts between parents and schools [10]. In France, some schools in deprived areas have involved parents through awareness meetings to promote the tracking of their

children in order to avoid their truancy [11]. The UK spent £1 billion between 1997 and 2005 to combat truancy [1].

1.4 Reasons for truancy

When finding for reasons for truancy, we discover that there are many factors common across different societies. However, there are countries where the risk of truancy may be higher due to specific circumstances. We believe that the security situation in Iraq is a factor affecting pupils' school attendance.

From the literature, the common causes of truancy in most countries, including Iraq, include family disorders, such as separation of parents or the death of one or both [12]. Another factor is weakness of personality and the lack of integration into the school environment [13].

Yet another possible cause is the parents lack of awareness of the value of education, due to their low level of education. This can lead to a lack of interest in their children's attendance. These children then commit truancy from school [14]. In some cases, families may encourage pupils to leave school to work in order to fill any financial shortfalls of the family. This has a direct impact on regular school attendance [15].

Low family income is also another reason for truancy, with the tendency of the pupil to work to meet the needs the family. The provision of assistance to low-income families by schools and community organizations can contribute to the decrease of such situations [16].

1.5 Iraqi education truancy problem

Iraq has particular truancy problems, which are becoming more widespread than ever, due to the poor security situation and poor living conditions as a result of successive wars. Nevertheless, Iraqi schools and the Iraqi education system still keep

extensive paper records on their pupils, from primary school through to secondary school. Data are recorded on the pupil as well as on the status of the student's family, family income, personality, health status, and security situation. In particular, information is kept on actual truancy. In terms of Management Information Systems, it is well worth speculating as to whether records having been kept electronically would have survived two wars.

1.6 Research approach

We used pupil data from school cards and through data mining to create a predictive model that is able to identify pupils at risk of truancy before it occurs. In the process, we will identify the main factors affecting truancy in terms of statistical significance. The school cards provide information on many factors and most importantly they identify pupils who have reached a certain level of truancy.

1.7 Limitations of research

1- The research was applied to Iraqi schools.

2- The research was conducted using data of male students only: data of female pupils was not used.

3- The research was carried out on a sample consisting of 30 students from different schools.

4- We used only school card data.

CHAPTER 2

STATE OF THE ART

2.1 Introduction

The field of truancy research lies within that of education, which in turn is a part of the social sciences. Therefore, we will consider data mining (DM) in this wider context as being relevant. In addition, the measurement of factors relating to pupils' general performance in education becomes relevant. The reason or this, we believe, is that truancy is but one aspect of an unsuccessful engagement between the child and his/her education and some of the factors may overlap.

2.2 Data mining in social sciences

The social sciences cover a number of areas, including economics, health, social work, geography, and education. DM has been applied to the social sciences, but the instances are few and spread across this whole range.

In economics, the area of bankruptcy has been addressed by DM [17]. Here, researchers collected different financial variables from firms in an attempt to predict the viability and likelihood of bankruptcy of the firms. The results of the study indicated that the method that was used performed well to predict bankruptcy.

Geography is also an area of the social sciences that uses DM to discover new patterns from large volumes of social data. [18] DM is used to discover knowledge from geographic databases; it is difficult for the old methods (handy methods) to

handle this huge amount of data. DM is used to classify soil types because of the inclusion of several geographical variables; difficult used the old methods to determine it.

Child protection within social works is another area that has used DM [19]. There is a degree of similarity here to our approach, as we are also endeavoring to measure subjective social factors impacting on the child regardless of whether those factors include abuse or truancy. The authors also relied on the experiences of social workers to build a model. We hope to follow the same methodology by using the experience of education workers to identify the attributes that affect truancy. They did, however, restrict themselves to only one DM technique, the ID3 algorithm, but they showed identified key factors in the classification of child abuse risk. We intend to test our data using a variety of algorithms for improved accuracy.

2.3 Factors measuring student behavior/performance

Many of the factors that influence behavior and performance in education are the same factors which influence truancy.

The research by Bhardwaj and Pal [20] determine many factors affecting the performance of a student, such as personal, psychological, social, environmental as well as grades. They proved that all these factors have an impact on the performance of the student and can be used to identify students who are likely to have problems.

One study by Avvisati, F. et al [11] has shown that parents' following up their children has a direct influence on their children’s behavior and performance at school. Therefore, the family situation, if it is troubled (such as loss of one or both parents) negatively affects the performance and behavior of their children at school.

2.4 Data mining in academic behavior/performance

DM techniques can be applied to the education sector to predict student behavior, performance and truancy. Educational Data Mining (EDM) is a term for the application of machine learning algorithms to extract knowledge from data in the educational environment.

Kumar and Vijayalakshmi [21] built a model that predicts student performance (pass or fail) by using grades from their courses. This gives teachers an early warning to help students who may be expected to fail. This method proved that helping weak students improved their performance in final exams.

Yadav et al [22] also attempted to predict performance; however, they used not only grades, but also other attributes about the student, such as personal, psychological and social factors. Their model succeeded in identifying students who were expected to fail or pass an exam.

Another study [23] in Mexico used historical data from 670 students and identified several factors to predict failure in their studies as well as truancy. These included family status, social status, personality, psychology, economic status, scientific background and student grades. They selected the best 15 attributes from 77 which were most influential on the accuracy of the models, using the select attribute filter, which determines the most influential attributes in the accuracy of the model. The model succeeded in predicting which students were expected to fail or be truant. It was used to alert teachers and parents to take preventive measures.

In distance learning, a system called Virtual Learning Environment [24] used classification to build a model that predicted students at risk of failure. Their research showed that lack of direct contact face to face (attendance in class) between teacher and student was one of the main factors that affected the student performance. Other factors included lack of time, lack of interaction with the system, financial difficulties and lack of knowledge of modern technology. The study concluded that proper intervention in real-time with students improved their performance.

2.5 Factors affecting truancy

Truancy has been widely studied in the literature and we look here at those studies to identify the factors which may direct our own research.

In a 2003, study by Muhammad [25] in the state of Marwa in the Sudan, three groups (school principals, teachers, and a number of pupils) were questioned about truancy. The study concluded that there were three major factors affecting truancy.

1- Education: Repeated academic failure in some lessons and the students' being older than their peers.

2- Economic: Family poverty such that children are expected to fill any financial deficits of the family.

3- Social: The absence of the father for long periods because of work, the low academic level of parents, cultural behaviors such as the marriage of their children at a young age.

Some of the factors which affect truancy mentioned by the researcher are the same as those used in our research. These include the economic situation, social situation (family disorder), parents' academic level and cultural behaviors such as sectarian conflicts.

Pradeep et al [26] used data from an Indian school between 2011 and 2013 based on 670 pupils aged between 17 and 19. The researchers showed that the factors influencing truancy were related to psychology, family, social and cultural background, scholastic progress, demography and socioeconomic status.

Heredia D, Amaya Y and Barrientos E. [27] used in their research data from 201 pupils and proved several factors, such as the personality of pupils, their birthplace, sex, parents' profession, parents' marital status, educational level of parents, economic variables, and family income to build a truancy prediction model. Many of these factors will be used in our research.

In Turkey, little research has been carried out. Pehlivan, Z. [28] focused on a softer approach to interviewing people for their thoughts on the reasons for truancy. The study concluded that the school staff felt that the pupils and their families were the cause of the truancy. We, on the other hand, take a more rigorous mathematical approach.

The factors affecting truancy mentioned in Arab studies [29] are not very different from the factors in other countries. There are the usual factors of family, economy, society, and pupil psychology and pupil personality. Any difference is reflected by the state of society in Arab countries, which suffer from poverty, underdevelopment, and sectarian conflicts.

2.6 Data mining for truancy (dropout)

We consider that the phenomenon of dropout is similar to truancy because the factors driving it are very similar. Dropout is usually associated with a third level education and means the reluctance of the student to attend lectures or to actually leave the university completely.

Hasbun et al. [30] used DM to predict pupils at risk of dropping out. They created two decision tree models to predict absence, with data from 4,840 pupils. The first model includes the attendance of pupils participating in non- compulsory, extracurricular activities (such as sports, workshops) where truancy can occur without punishment. The second model only had factors relating to attendance in compulsory lessons. The models showed that extracurricular activities influence the absence of pupils. The factor of extra- curricular activity highlighted by the authors has been included in our model through the activity attribute. Although extracurricular lessons are almost non-existent in most Iraqi schools, we felt that they represented the engagement of a pupil with his school out with just lessons.

Perlta et al. [31] analyzed the factors that could be used to construct a predictive DM model to identify students who would drop out from university. They used machine learning algorithms due to their ability to analyze factors to discover the underlying

patterns of evasion as well as to be able to deal with large data. Moreover, it was proved that the variables are relevant to the opinions of human experts working in the field of education.

Many distance learning institutions also suffer from the dropout of students. Santos et al. [32] used DM techniques to detect dropout patterns by creating a model based on pupil interactions with their lecturers. The model succeeded in identifying pupils likely to leave early so that the problem would be addressed before they dropped out.

2.7 Arabic research on truancy

Many schools in Arab countries suffer from the problem of truancy. Few Arab researchers have addressed the problem of school truancy, despite its prevalence in Arab schools. The security situation remains the main factor that is different from other countries. This especially applies to Iraq.

CHAPTER 3

THE METHODOLOGY

3.1 Introduction

In this section, we will describe the stages of creating predictive models for truancy. We will propose a model which has the highest accuracy on truancy, by evaluating several algorithms and filters on real test data. The stages of building the models are shown below in Figure 1.

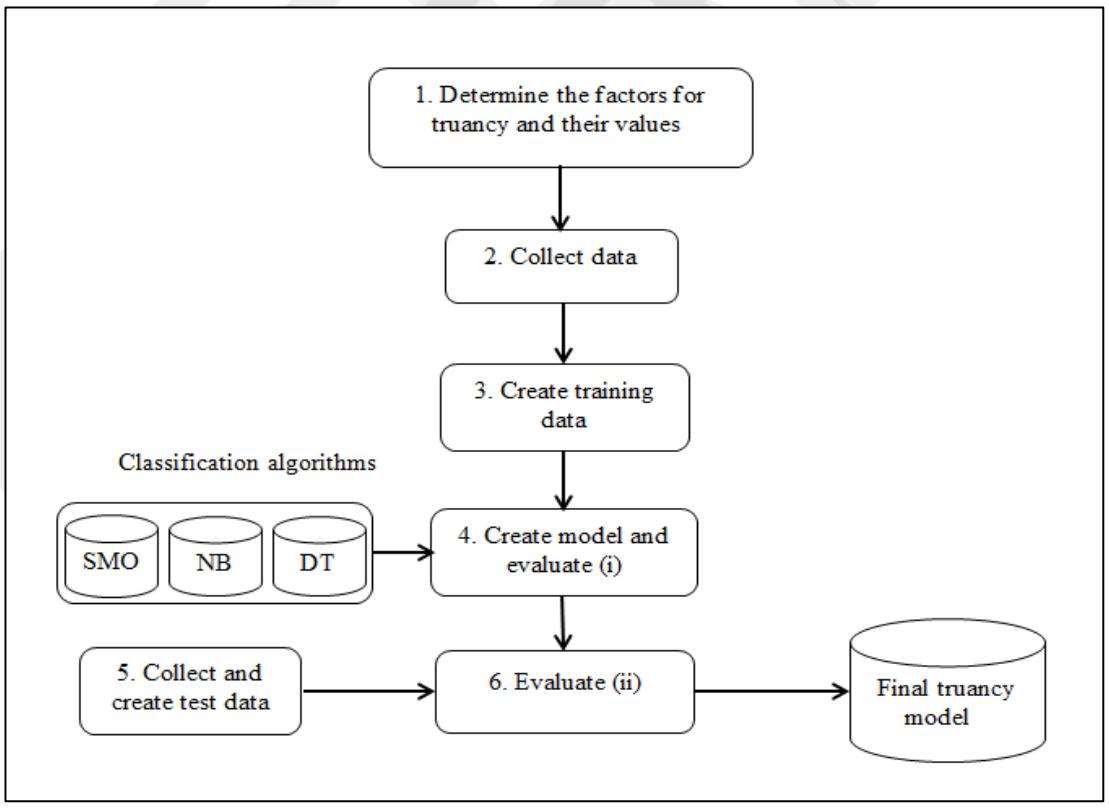


Figure 1 Truancy classification methodology

3.2 Determining the factors for truancy and their values (box 1)

Seven factors have been identified as having a potential impact on truancy. These factors and their values are based on the experience of the author, experts in the field of education in Iraq, as well as research literature. We will adopt a method similar to that used by Little and Rixon [19], to identify factors and their values. In the pupil's school card, there are many fields that contain information written by the teacher and school administration which helps measure these factors. However, these values for the factors cannot be read directly from the school card. It still requires the teacher and school administration to bring different measures together and make one overall assessment of the value. Without consistent data, it makes finding and validating a good model more difficult. The factors and their values are described as follows:

3.2.1 Family situation of pupil (stable, troubled)

This factor relates to the pupil's family and how close they are to a conventional household, of a mother and father, in a stable environment. There are two possible values which are relevant to the analysis.

If the family situation is **stable**, it means that the parents are alive, they have few problems between them and the pupil is living at home with them. The parents are caring towards their son* and provide him with all his emotional and material needs, such as by providing a suitable environment for study at home. This level positively affects the pupil to remain in the school system, since both parents are present to take an interest in him.

---

* The study focuses only on all-boy schools.

If the family situation is **troubled**, it means that one or both parents of the pupil are deceased or they are separated. The pupils will live with one of the parents or with another family member. This reduces the level of care by the parents. Under these conditions, the pupil is believed more likely to be truant.

The required information for this factor is available in the school card highlighted in blue in Figure 2.

3.2.2 Parents' academic level (high, low)

This factor relates to the education level of the pupil's parents. It has two possible values.

If the parents' academic level is **high**, it means that at least one of the parents holds a university degree. It will likely follow that the father or mother understands the importance of education and so they encourage their son to study. This is reflected positively in the continuation of the pupil's attendance at school.

If the parent's academic level is **low**, it means that both parents do not have a university degree, so they do not place high value on education. This leads them to a lower level of academic encouragement, which potentially leads to the pupil being less motivated to go to school.

The supporting information for this factor is available in the school card highlighted in green in Figure 2 as well as the teacher's own knowledge.

3.2.3 Security situation of pupil (negative, positive)

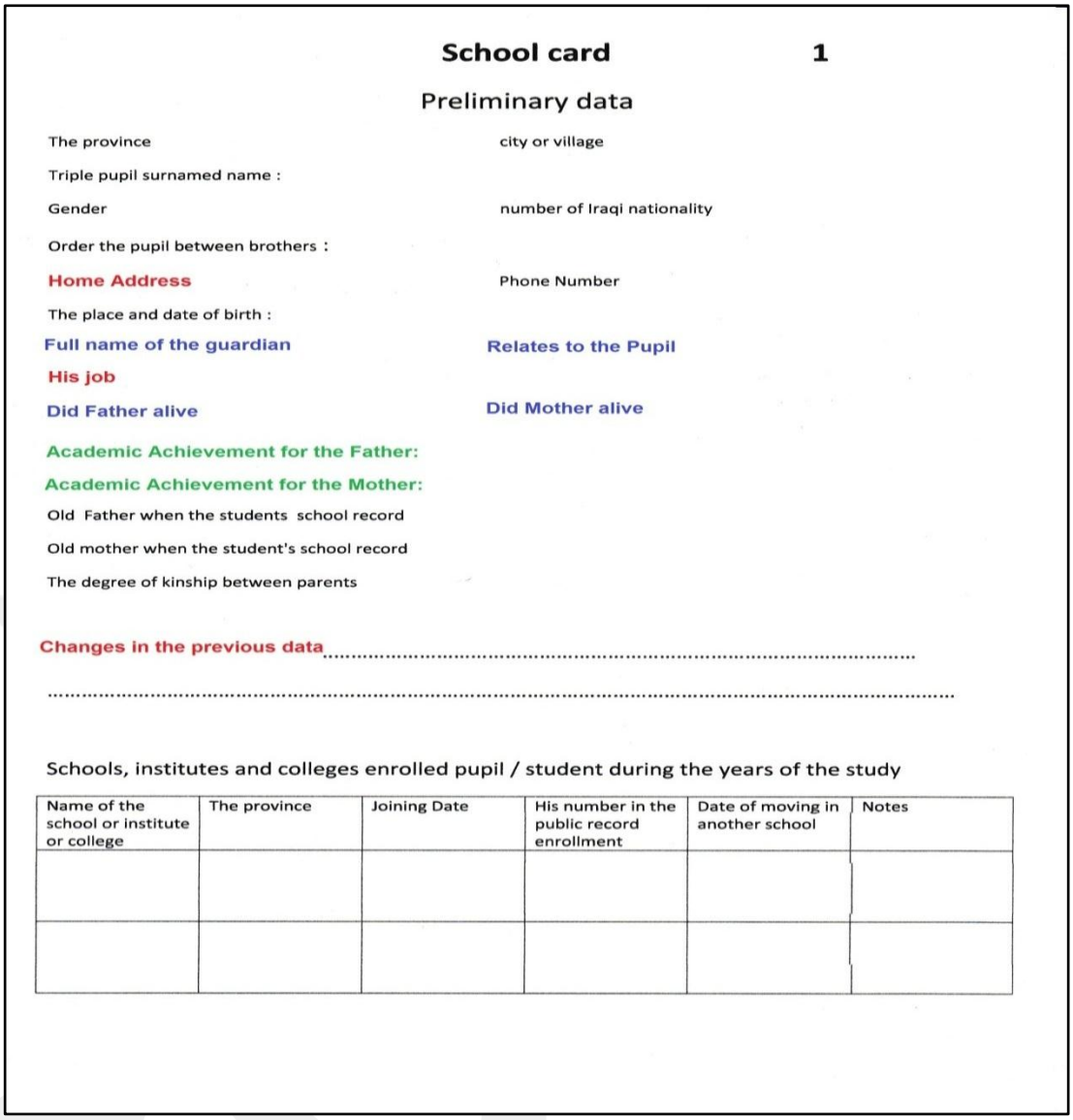
This factor covers issues of the distance of the pupil's residence from their school, his parents' career and any potential sectarian threats. This is a different factor to

those described in the literature because the security situation is particularly dire in Iraq. There are two possible values.

If the security situation is **negative**, then this implies the presence of at least one of the previous three issues. Firstly, if the pupil's home is far from the school, there is a greater possibility of being exposed to terrorist acts along the route and thus being dissuaded from attending school. Secondly, if one or both parents are working in one of the security departments of the state, then this increases the likelihood of terrorism towards the family. Again, this is reflected negatively in the possibility of truancy. Thirdly, if the family is suffering from sectarian conflicts, which causes a disturbance to the security of the family, then this may be reflected negatively on their children, hence an increase in the likelihood of truancy.

If the security situation is **positive**, it means a lack of all the three issues mentioned above. That is, the pupil continues going to school.

This information is written by the school administration on the school card and is highlighted in red in Figure 2.



School card **1**

Preliminary data

The province

city or village

Triple pupil surnamed name :

Gender

number of Iraqi nationality

Order the pupil between brothers :

Home Address

Phone Number

The place and date of birth :

Full name of the guardian

Relates to the Pupil

His job

Did Father alive

Did Mother alive

Academic Achievement for the Father:

Academic Achievement for the Mother:

Old Father when the students school record

Old mother when the student's school record

The degree of kinship between parents

Changes in the previous data.....

.....

Schools, institutes and colleges enrolled pupil / student during the years of the study

Name of the school or institute or college	The province	Joining Date	His number in the public record enrollment	Date of moving in another school	Notes

Figure 2 Fields of school card that contain relevant information about the pupil

3.2.4 Family income of the pupil (poor, suitable)

This factor relates to the economic level or monthly income of the pupil's family. There are two possible values.

If the family income is **poor**, then we consider this as being less than $500 per month. Having low income will increase the likelihood of their son's truancy or even complete absence from school, as the child may be forced to work, full or part-time.

If the family income is **suitable**, it means that the family income is greater than $500 per month. Here the financial situation is better, which reflects positively on the pupil not having to miss school. This information is available in the school card highlighted in red in Figure 3.

Social status 2

Data / The years	Number Of Family	Number of Brother and Sister		Family income Monthly	Number Of room In house	The case of the pupil Family Live with	Did the Pupil Work		The suitability of the home to study
		brother	Sister				Yes	No	
2012-2013									
2013-2014									
2014-2015									
2015-2016									
.									
.									
.									
.									
.									

Figure 3 Information of monthly family income

3.2.5 Activity of pupil (active, inactive)

This factor relates to the pupil's non-curricular activities (e.g. sports and arts). It measures whether the pupil is committed to this additional aspect of school life. It has two possible values.

If the value is **active**, it means that the pupil is interested in participating in extracurricular activities such as sports lessons. This, in turn, is reflected positively in general attendance at the school and hence increases the likelihood of non-truancy.

If the activity value is **inactive**, it means that the pupil participates very little in these extracurricular lessons. This shows a lack of engagement with the school and hence a higher likelihood of truancy.

This factor is not so obvious, but it was suggested because of a study that showed its effect on truancy [29]. In Iraq, there are no school clubs or sports, yet we can still measure involvement through the school card, as highlighted in red in Figure 4.

3.2.6 Personality of the pupil (introvert, balanced)

This factor relates to the personality of the pupil and his dealings with his teachers and peers. This factor has two possible values.

If the personality is **introvert**, it means that the pupil is not compatible with most pupils. He also does not interact well with teachers and in many cases may feel afraid of them. Again, this factor shows a detachment from school affairs or fear of people and hence a greater likelihood of staying away.

If the personality is **balanced**, it means he is compatible with his fellow pupils and teachers; he has no fear of them. This being the case, the likelihood of truancy diminishes. All this information is written by the teacher and the school administration on the school card, as shown in Figure 4 in the balanced field.

Character traits 4

Characters	The Years / Measuring capacity	200 / 200 Class	200 / 200 Class	200 / 200 Class	-------	-------	-------	-------	-------
Activity	**Active in extracurricular lessons**								
Equilibrium	It seems sober								
Leadership	He leads his friends								
Teamwork	Collaborates with his friends								
Self confidence	Trust himself								
The ability to innovate	Gives new ideas								
Attention and focus	Pay attention and concentrate on his work								
Dressed	He cares about his clothes and cleanliness								
Among his colleagues	He loved and respected by his friends								
Dealing with others	Fluent in dealing with others								
Balanced	**Open to life and Compatible with peers**								
Taking responsibility	Bears responsibility in his work								
Courage and boldness	Characterized by boldness and courage								

Note: Place the right word, which comes (frequently, sometimes rarely) in front of capacity, which is characterized by the student / pupil,

Figure 4 Activity and personal characteristics of the pupil

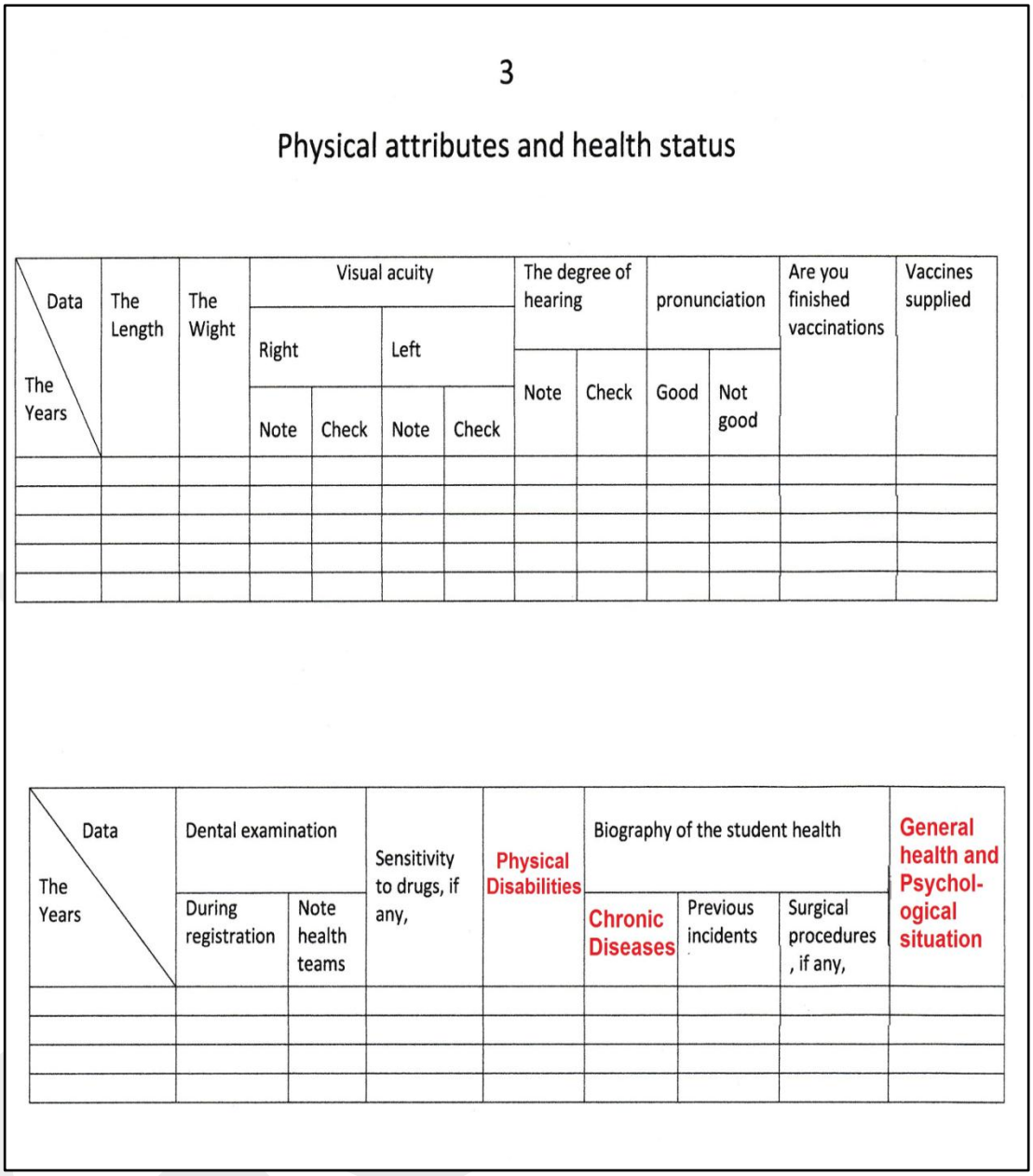
3.2.7 Health situation of the pupil (sick, well)

This factor relates to the pupil's health. It affects his attendance at school by missing classes and falling behind. This factor has two possible values.

If the pupil's health situation is **sick**, it means that the pupil has one of two conditions. The first is that the pupil has a psychological condition, which can be determined by a doctor or through expert observation. The pupil may have an inability to deal with his friends and sometimes other pupils will be afraid to approach him. This leads to his isolation, fear of friends and thus greater deterioration of his condition, leading to truancy. The second is that the pupil has a chronic illness such as diabetes, a heart condition or other chronic diseases. The chronic disease leads to weakness of his body and the need for regular visits to the doctor. It may also require him to take medication at specific times. These issues create opportunities for extra time off, in the form of truancy.

If the pupil's health situation is **well**, it means that the pupil does not have any of the conditions mentioned above. Therefore, truancy is less likely.

Figure 5 highlights in red the health information of the pupil in the fields of chronic illness and mental state. All this information is written by the school administration and is based on any medical reports provided by the pupil or his parents.



3

Physical attributes and health status

Data / The Years	The Length	The Wight	Visual acuity				The degree of hearing		pronunciation		Are you finished vaccinations	Vaccines supplied
			Right		Left							
			Note	Check	Note	Check	Note	Check	Good	Not good		

Data / The Years	Dental examination		Sensitivity to drugs, if any,	Physical Disabilities	Biography of the student health			General health and Psychol-ogical situation
	During registration	Note health teams			Chronic Diseases	Previous incidents	Surgical procedures , if any,	

Figure 5 Information about the pupil's health

3.2.8 Risk (high, low)

This factor represents the value of the class and it is the factor we are attempting to predict. It depends on the values of the other factors in each instance and has two values.

If the value of risk is **high**, then that means there is a strong likelihood of truancy taking place or about to take place. We selected to measure this through the number of days of recorded truancy. Firstly, if the pupil was absent for more than 15 days without reasons, then a warning was sent to the pupil's parents. Secondly, if the pupil truanted for 25 days, then he is considered to have failed the academic year. In both cases, it means the pupil is high risk. This is the official manner of measuring truancy in Iraq

If the value of the class **is low**, then this means that the pupil does not have or is unlikely to exhibit truancy. The pupil is unlikely to exceed the minimum number of days of absence allowed.

We know this value usually for the training data, but the intention is to predict it for new pupils or if the pupil has a new set of circumstances, changing the value of any of the above factors.

The information on the number of days of truancy for the pupil is recorded on the school card. Figure 6 shows in red the actual number of days of truancy by the pupil and the given or supposed reason.

7

Attendance and absence

Data / The Years	Total absences Chapter One		Total absences Chapter Two		The Reasons and Number of days for Truancy	Compulsory licenses granted to students due to *illness* with disease name	Actions taken to address the phenomenon of absenteeism
	Excused	Without Excuse	Excused	Without Excuse			

Figure 6 Number days of truancy excused or not excused

3.3 Collect data (box 2)

We collected data from primary, middle and high schools in Iraq via school cards. For the purposes of our research, we considered the concept of truancy independent of age – as the same basic causes are present. These schools were male only and so the analysis covered only boys. This source of data has been strictly recorded for each pupil and completed by school administration and teachers. Thirty representative samples of pupils were collected. A small example of a completed set of factors/values is shown in Table 1.

Table 1 Set of factors and values in the training set

No	family situation	parents academic	security situation	Family Income	activity	personal	health situation	risk
1	troubled	High	positive	Suitable	active	balanced	well	high
2	stable	High	negative	Suitable	active	balanced	well	low
3	troubled	Low	positive	Suitable	inactive	balanced	well	low
4	stable	High	negative	Suitable	active	balanced	sick	high
5	stable	Low	positive	Poor	inactive	balanced	well	low
6	stable	Low	positive	Poor	inactive	introvert	well	high
7	troubled	Low	positive	Poor	active	introvert	well	high
8	stable	High	negative	Suitable	inactive	introvert	sick	high
9	stable	High	negative	Suitable	active	balanced	well	Low
10	troubled	Low	positive	Suitable	inactive	balanced	well	low

We collected the training data from school cards which we sorted and cleaned as follows.

1- Selection of diverse samples means the factors causing truancy is varied so that to cover the largest number of patterns that cause truancy, the training data is limited.

2- Selecting an equal number of samples of pupils who have a risk of truancy and those who do not means all samples are representative to obtain better performance [33].

3- We adopted various information sources that were transferred by the school administration and teachers to one value per attribute.

3.4 Create training data (box 3)

After establishing a set of examples with factors and class values, we turn this into training data for use with WEKA – the DM tool. We created a standard input "Arff" file of this training data (see Figure 7).

```
@relation truancy

@attribute family_Situation {stable, troubled}
@attribute parents_academic {high, low}
@attribute security_situation {positive, negative}
@attribute family_income {suitable, poor}
@attribute activity {active,inactive}
@attribute personal {balanced,introvert}
@attribute health_situation {well,sick}
@attribute truancy_risk {high,low}

@data
troubled, high, negative, suitable,active,balanced, well, high
troubled, high, positive, suitable,active,balanced, well, high
troubled, high, positive, suitable,active,balanced,well,low
troubled,high,negative,suitable,inactive,introvert,sick, high
stable, low, positive, poor,active, balanced, well, low
stable, low, positive, suitable,inactive,introvert, sick, high
stable, low, positive, poor,inactive, introvert, well, low
troubled, low, positive, suitable,active,balanced,well, low
troubled, low, positive, poor,active, balanced, well, high
troubled, low, positive, poor,inactive, introvert, well, high
stable, high, negative, suitable,active, balanced, well,low
stable, high, negative, suitable,inactive, balanced, sick,high
stable, high, negative, suitable,inactive,introvert,sick,high
troubled,high,positive, suitable,active,balanced,well,low
troubled,high, positive, suitable,active,balanced,well,low
stable, low, positive, poor,active, balanced, well, low
troubled, high,positive,suitable,active,balanced, well, high
troubled, high, positive, suitable,active,balanced,well,low
troubled, high, negative,suitable,inactive,introvert,sick,high
troubled, low, positive, suitable,active,balanced, well,low
```

Figure 7 Training data in standard Arff format

3.5 Create model and evaluate (i) (box 4)

We want to build a prediction model for truancy based on one of several classification algorithms. The models are built with the Weka system; the user simply provides the data and selects the algorithm (plus parameters) to use to build the best model. We selected several classification algorithms most commonly used by researchers for our research. Then, we applied them to the training data and

determined the accuracy level which WEKA provides. We will determine the highest accurate models.

3.5.1 Comparison measurements

The comparison measurements we will make for each algorithm are the percentage accuracy, the number of correctly classified instances, the number of incorrectly classified instances and the Kappa statistic. The Kappa statistic is used to measure the agreement between predicted and observed categorizations of a dataset. Moreover, it is a measure of how better the algorithm works rather than merely a random one. The Kappa statistic when approaching a value one the performance of the model was better, but when the value decreases and approaches zero or less (negative value) the performance of the model is not good. The performance of the classifiers is also based on the confusion matrix. It contains values of, true positives TP is the value of instances is YES and the predict them is YES, true negative TN is the value of instances is NO and the predict them is NO, false positive FP is the value of instances is NO but the predict them is YES. For example, when the pupil has actual high-risk, but the model predicts low-risk. The model indicates, therefore, a false negative FN. These types of agreements/contradictions are shown in Table 2.

Table 2 Confusion matrix for two value class

		Predicted value	
		Yes	No
Actual value	Yes	True positives (TP)	False negative (FN)
	No	False positives (FP)	True negative (TN)

3.5.2 - Classification algorithms

We elected five classification algorithms most commonly used by researchers. They are as follows:

- **Naive Bayes (NB)**

The NB classification algorithm is based on Bayesian statistics of conditional probability. This technique evaluates the probability of each value of a class. It is valid only when events are independent as we believe is the case here.

- **Support Vector Machine (SMO)**

Sequential Minimal Optimization SMO is an example of a Support Vector Machine method which represents the closest instance to maximum margin. The algorithm works based on the creation of a default linear classifier between instances and then it determines the margin of the linear classifier, the width in which its boundary can be increased before hitting the data point. The instance located in the maximum margin is SVM (called an LSVM). Each class always has one Support Vector (instance) and in some cases, has more than one. Moreover, a set of support vectors identifies the maximum margin of the hyperplane to the learning problem.

- **Instance Based Knowledge (IBK)**

The classification algorithm K- Nearest Neighbors is an approach which attempts to find an instance which is close (least distance) and then takes the class value from there. By measuring the distances between the instance (using the Euclidean method), the nearest neighbor is the least distance or has the same factors. The predictions can be weighed for more than one neighbor according to their distance from the test instance. We will use two algorithms, namely the K- Nearest neighbors with value parameters of K=1 and K=2.

- **One rule (OneR)**

OneR is an approach which generates a set of simple rules based only on one attribute and its different values. For example, we could select the attribute of the security situation and split the prediction by the different values to it.

- **Decision Tree (DT) J48**

The decision tree works by moving from the top to the bottom of the tree to reach a value for the predicted class at a leaf. Each move corresponds to a question about the value of the example's attribute. Descending the tree through a series of these questions leads to the bottom and a predicted value. In our case, we seek to predict the risk of truancy by answering questions on the case attribute/value to reach the value of the class. The path down the tree depends on the answers to the attribute/value questions.

3.5.3 - Apply classification algorithms with and without a filter

In this part, we apply each classification algorithm to the training data without using the filters in order to obtain accuracy values from Weka in each classification algorithm. We also used in every experiment a cross-validation of every classification algorithm which decided on a number of folds to obtain the results. This works by dividing the data into three sections at random, using two parts of the data for the training and the last section for the testing. Then it will repeat this process three times so that each section is used each time in the testing.

We then attempt to improve the accuracy of the models by using the Select Attribute filter, which identifies the most influencing attributes and attempts to reduce the size of the problem accordingly. The thought is that a simpler model might have higher accuracy.

3.5.4 - Select initial model (i)

In this part, we will compare the results of the models' accuracy in the tables obtained to determine the best model. We will select the models that hav the highest accuracy in order to test them using test data to determine the best one of them.

3.6 Collect and create test data (box 5)

In this part, we collect new samples of pupil data from different schools. The screening process was conducted according to the steps in box 2. We convert the test data to an Arff format in a similar process to box 3 as input to the WEKA.

3.7 Evaluate (ii) (box 6)

We selected the two best models with the highest accuracy and applied each of them to a new set of test data, which has not been used to build these original models. We evaluate each in terms of its accuracy. We know the risk of the pupils already (ground truth) and to determine the accuracy, we compared the actual with the predicted. We then selected the model (algorithm) with the highest accuracy to adopt this model.

CHAPTER 4

RESULTS

4.1 Introduction

Initially, we provide results obtained without filters by using the default settings. We then considered improving the accuracy of the models by using the filter Select Attribute. Finally, we will be identifying the best prediction model of truancy by testing it on new unseen data.

4.2 Initial prediction models

We built models and validated them with training data of 30 examples on six of machine learning algorithms. Table 3 shows the results for each model across several evaluation criteria.

Table 3 Initial classification algorithms performance

Evaluation Criteria	Classification algorithm					
	NB	SMO	IBK-K=1	IBK-K=2	OneR	DT J48
Accuracy (%)	70%	66.7%	90%	56.7%	36.7%	73.3%
Correctly classified	21	20	27	17	11	22
Incorrectly classified	9	10	3	13	19	8

Table 3 (continued)" on the top.

Kappa Statistic	0.4	0.3	0.8	0.1	-0.3	0.5

In the above table, the algorithm of IBK-K=1 has clearly the highest accuracy and Kappa value compared to the other classification algorithms.

We also obtained a further breakdown of results for the accuracy of class value through the confusion matrix, as shown in Table 4.

Table 4 Accuracy based on the class

Classifier	high	low	Class
NB	9	6	high
	3	12	low
SMO	10	5	high
	5	10	low
IBK-K=1	13	2	high
	1	14	low
IBK-K=2	7	8	high
	5	10	low
OneR	3	12	high
	7	8	low
DT J48	10	5	high
	3	12	low

We conclude that IBK-K=1 is still the best with the highest value of true positives TP=13, true negative TN=14 and the smallest value of false positive FP=1, false negative FN=2 compared to the other classification algorithms.

On the surface, the approach of IBK seems to best solve this type of problem by looking for the closest pupil (has the largest number of similar attributes). This is not unlike how the human works when they think of another pupil similar, to guide their assessment. Also, since we have a small sample here, it is likely the next most nearest example is quite a distance away, so it will not get a good result Hence IBK-

K=2 classification algorithm is worse as we include a ‘bad’ example in the prediction.

4.3 Use of data filters and identification of influencing attributes

We obtain the next set of results by applying the Select Attribute filter. Table 5 shows which attributes are most influential for each classification algorithm.

Table 5 Most influential attributes for classification algorithms (in bold)

Classifier	Attribute influencing
NB	**family situation** **security situation** personal health situation
SMO	**family situation** parents academic **security situation** family income personal health situation
IBK-K=1	**family situation** **security situation** family income activity personal
IBK-K=2	**family situation** parents academic **security situation** family income activity

"Table 5 (continued)" on the top.

	personal
OneR	**family situation** parents academic **security situation** family income activity personal health situation
DT J48	**family situation** **security situation** family income health situation

We see from the table that the most influential factors across all classifiers is the family situation followed by the security situation. This is very plausible and discussions with education experts and teachers* in Iraq bore this out. Dr. Sayed Al-Kaabi, one of the experts in education in Iraq said that the security situation has a significant impact on the pupil and causes psychological instability for him, which reflects negatively in continuing to school and hence truancy.

The accuracy filter suggests a smaller set of attributes may make a difference by reducing the size of the problem. We reduce the input to only include those attributes for each algorithm respectively. In Table 6 we show the change in accuracy to the algorithms.

---

* Dr. Sayed Al-Kaabi. Psychological Counseling.
* Dr. Azhar Al-Saffar. Measuring & Evaluating.

} *The Ministry of Education*

Table 6 Accuracy of classification algorithms using with and without filters

Classifier	% Correctly classified instances with	
	Filter Select Attribute	Original without filter
NB	73.3 %	70 %
SMO	80 %	66 %
IBK-K=1	86.7 %	90 %
IBK-K=2	43 %	56 %
OneR	36.7%	36.7 %
DT J48	76 %	73 %

In the table above, the IBK-K=1 algorithm is still the best performing. However, filtering does not help and indeed reduces its accuracy. The reason could be that when the number of attributes is reduced and the number of cases stays the same, it gives less guidance as to the best nearest neighbor. We still need all the attributes. While SMO algorithm improved significantly. All other classifiers improved with this filtering but still lagged IBK-K=1.

4.4 Determine the best model

We have identified the two best algorithms as IBK-K=1 without filter and SMO with filter. We now test these two algorithms finally against completely new data (15 examples). Table 7 shows the accuracy results we obtained for the two classification models. The table also breaks down the correctly and incorrectly classified instances.

Table 7 The performance of SMO and IBK-K=1 on new test data instances

Evaluation	Classifiers	
	SMO	IBK-K=1
Accuracy (%)	86%	93%
Correctly classified	13	14
Incorrectly classified	2	1

We conclude that the performance of the IBK-K1 algorithm continues to be slightly better than the SMO classifier, in terms of the correctly classified instances. It was also able to classify all, but one correctly, both showed good performance.

Table 8 shows the confusion matrix for the classification algorithms SMO and IBK-K1.

Table 8 Confusion matrix for SMO and IBK-K=1 on test data

Classifier	Predicted class		Actual class
	High	low	
SMO	5	1	high
	1	8	low
IBK-K=1	5	1	high
	0	9	low

We conclude that IBK-K=1 better than SMO. SMO got both a high and low-risk prediction wrong, while IBK-K1 only classified low wrongly. SMO effectively classified a FP, the pupil has high-risk in actual value, but predicted it low-risk from IBK-K=1. This is a costly mistake because of FN not as costly such as FP. As for IBK-K=1 it had one mistake of a FN has a higher ‘cost’ because of we are saying this pupil will not truant when in fact there is a high risk also SMO has this mistake. So SMO is not perfect from IBK.

Since IBK-K=1 is the best, it is important for the reader to understand how it would work based on an example in Table 9.

Table 9 Several known truancy cases and one unknown

No	family situation	parents academic	security situation	family income	activity	personal	health situation	risk
1	troubled	high	positive	suitable	active	balanced	well	high
2	stable	high	negative	poor	inactive	introvert	well	high
3	troubled	low	positive	suitable	active	balanced	well	low
4	stable	low	positive	suitable	inactive	balanced	sick	low
5	**troubled**	**high**	**positive**	**suitable**	**active**	**balanced**	**well**	**?**

In this table, instances 1 to 4 are the training data and instance 5 it is test data – we want to predict the value of risk by building a model from the training data to be used on the test data. In this case the IBK-K=1 model will recommend the value of risk as high for instance 5, because of it is closest with the instance1, whose risk is high. If the instance 1 did not exist, then we will recommend low because the instance 3 is the nearest neighbor. This also shows how dependent the model is on the data. Less data usually means less accuracy in the result.

Looking at the one example where IBK-K=1 misdiagnosed with such few data we can say perhaps that there was no example close enough to choose the correct class. This type of modeling would allow us to always extend the training model with new examples to make it 'better'.

4.5 Additional validation

The research by Little and Rixon [19], is of high relevance to us because it is like our field of research (social/child data). They only reported results using one classifier that based on ID3 decision trees. We want to investigate whether this dataset (Table 10) was also best modeled by an IBK-K=1 approach. We therefore, took their data

and evaluated it against the range of algorithms we had used together with ID3, the one they used. For consistency, we gave the risk level attribute two possible values, High and Med/Low. Like truancy, we are only interested in who will truancy and who not.

Table 10 Original social work data

	Attributes								
Case	Risk	Type	Carers' attitude	Carers' ability	Child impact	History	Seriousness	Vulnerability	Immediacy
1	Low	Physical	Positive	Adequate	None	Negative	Low	Medium	High
2	High	Physical	Very negative	Negative	None	None	No evidence	High	High
3	Low	Physical	Positve	Negative	Strong	Very negative	No evidence	High	Medium
4	Low	Physical	Positive	Positive	None	Very negative	Low	High	High
5	Low	Physical	Positive	Negative	None	Positive	High	High	Low
6	Low	Physical	Positive	Adequate	Weak	Negative	No evidence	High	Medium
7	High	Physical	Positive	Very negative	Weak	Negative	High	High	High
8	High	Physical	Ambivalent	Negative	Weak	Negative	Medium	High	High
9	High	Physical	Positive	Very negative	Weak	Very negative	No evidence	High	High
10	Low	Physical	Ambivalent	Positive	Strong	Positive	Low	Medium	High
11	High	Neglect	Negative	Very negative	Weak	Negative	Medium	High	High
12	Low	Neglect	Ambivalent	Negative	None	Negative	Low	Medium	High
13	Low	Physical	Positive	Negative	High	None	Low	Medium	High
14	Low	Physical	Negative	Adequate	High	None	Medium	Medium	High
15	High	Emotional	Negative	Very negative	None	Negative	Medium	High	High
16	High	Emotional	Very negative	Negative	High	Unknown	High	High	Medium
17	Low	Emotional	Ambivalent	Negative	None	Unknown	Low	Low	Medium
18	High	Emotional	Negative	Negative	High	Unknown	High	Low	Medium
19	Low	Emotional	Positive	Adequate	Strong	Negative	Medium	Medium	Medium
20	Low	Physical	Very negative	Adequate	None	Negative	Low	High	Medium

We obtained the results of the performance of the six algorithms as well as the ID3 algorithm for accuracy (%), correctly classified, incorrectly classified and kappa statistic as shown in Table 11.

Table 11 Accuracy of classification algorithm with social work data

Evaluation Criteria	Classification algorithms						
	NB	SMO	IBK-K=1	IBK-K=2	OneR	DT J48	ID3
Accuracy (%)	70%	70%	75%	70%	55%	45	50%
Correctly classified	14	14	15	14	11	9	10
Incorrectly classified	6	6	5	6	9	11	9
Kappa Statistic	0.38	0.38	0.51	0.4	0	-0.08	0.01

The results show the same pattern of IBK-K=1 being the best performing as we have found for our data. Indeed, the approach (ID3) proposed by Little and Rixon shows relatively poor performance in prediction. These are small datasets though and probably reduced the accuracy of the classifier models. It was important to reduce the risk to two outcomes. Through related experiments, we found that with these few examples it was hard for the classifier to distinguish between three different outcomes. Two outcomes are easier to predict.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

The classifier IBK-K=1 had the highest accuracy in identifying pupils at the risk of truancy. IBK-K=1 worked because it perhaps reflected the decision methodology of the teacher thinking when he observing two similar situations to the new and old pupil having the same risk. As for IBK-K=2, there was not high accuracy, because the second nearest neighbor was remote (common factors are fewer) in such a small dataset. OneR was low accuracy because there was not a single factor which determined risk level – it was more complex and involved in some way, several factors. We identified the most influential attribute was the security situation, verified by teachers and experts in the field of education in Iraq. However, this is a factor we can do little about. It is thought possible to raise awareness for the pupil and his family which may reduce the risk of truancy occurring. The model can be used in practice at the start of the term or when the teacher notices a change in the pupil behavior.

We can use the model we obtained, in other Iraqi schools to predict the risk of truancy. This type of study is also relevant to Turkey which we also believe has a similar problem. Turkish researchers can follow our methodology to develop their own models, to their own family and environment characteristics.

Although there is less evidence of female truancy, it still occurs, but the factors may be different. So we propose as a future work, a separate piece of research is carried out.

REFERENCES

[1] Maynard, B.R., McCrea, K.T., Pigott, T.D. and Kelly, M.S., 2012. Indicated Truancy Interventions: Effects on School Attendance among Chronic Truant Students. Campbell Systematic Reviews. 2012: 10. *Campbell Collaboration*.

[2] Baker, M.L., Sigmon, J.N. and Nugent, M.E., 2001. Truancy Reduction: Keeping Students in School. Juvenile Justice Bulletin.

[3] Trujillo, L.A., 2006. School truancy: A case study of a successful truancy reduction model in the public schools. *UC Davis J. Juv. L. & Pol'y*, *10*, p.69.

[4] Yahaya, A., Ramli, J., Hashim, S., Ibrahim, M.A., Kadir, H.B.H., Boon, Y. and Rahman, R.R.R., 2010. The effects of various modes of absenteeism problem in school on the academic performance of students in secondary schools. *European Journal of Social Sciences*, *12*(4), pp.624-639.

[5] Veenstra, R., Lindenberg, S., Tinga, F. and Ormel, J., 2010. Truancy in late elementary and early secondary education: The influence of social bonds and self-control—the TRAILS study. *International Journal of Behavioral Development*, *34*(4), pp.302-310.

[6] Catterall, J.S., 1987. On the social costs of dropping out of school. *The High School Journal*, *71*(1), pp.19-30.

[7] Okwori, R.O., Ma'aji, A.S., Kareem, W.B. and Attaochu, E.U., 2014. Drop out among Basic Technology Students in Nigerian Educational System: Causes, Effects and Remedies. *Journal of Educational Policy and Entrepreneurial Research*, *1*(2), pp.204-210.

[8] Williams, L.L., 2001. Student absenteeism and truancy: Technologies and interventions to reduce and prevent chronic problems among school-age children. *Retrieved May*, *12*, p.2010.

[9] Smink, J. and Heilbrunn, J.Z., 2006. Legal and Economic Implications of Truancy. Truancy Prevention in Action. *National Dropout Prevention Center/Network (NPDC/N), Clemson University*.

[10] Swansea, K.R., 2010. Finding strategic solutions to reduce truancy. *Research in Education*, *84*(1), pp.1-18.

[11] Avvisati, F., Gurgand, M., Guyon, N. and Maurin, E., 2014. Getting parents involved: A field experiment in deprived schools. *The Review of Economic Studies*, *81*(1), pp.57-83.

[12] van Breda, M.J., 2014. Truants' perceptions of family factors as causes of school truancy and non-attendance. *J Psychol*, *5*, pp.47-53.

[13] Corville-Smith, J., Ryan, B.A., Adams, G.R. and Dalicandro, T., 1998. Distinguishing absentee students from regular attenders: The combined influence of personal, family, and school factors. *Journal of Youth and Adolescence*, *27*(5), pp.629-640.

[14] Marvul, J.N., 2012. If you build it, they will come: A successful truancy intervention program in a small high school. *Urban Education*, *47*(1), pp.144-169.

[15] Considine, G. and Zappalà, G., 2002. The influence of social and economic disadvantage in the academic performance of school students in Australia. *Journal of Sociology*, *38*(2), pp.129-148.

[16] Considine, G. and Zappalà, G., 2002. Factors influencing the educational performance of students from disadvantaged backgrounds. *Competing visions*, p.91.

[17] Foroghi, D. and Monadjemi, A., 2011, June. Applying decision tree to predict bankruptcy. In *Computer Science and Automation Engineering (CSAE), 2011 IEEE International Conference on* (Vol. 4, pp. 165-169). IEEE.

[18] Miller, H.J. and Han, J. eds., 2009. *Geographic data mining and knowledge discovery*. CRC Press

[19] Little, J. and Rixon, A., 1998. Computer learning and risk assessment in child protection. *Child Abuse Review*, *7*(3), pp.165-177.

[20] Bhardwaj, B.K. and Pal, S., 2012. Data Mining: A prediction for performance improvement using classification. *arXiv preprint arXiv:1201.3418*.

[21] Kumar, S.A. and Vijayalakshmi, M.N., 2011, October. Efficiency of decision trees in predicting student's academic performance. In *First International Conference on Computer Science, Engineering and Applications, CS and IT* (Vol. 2, pp. 335-343).

[22] Yadav, Surjeet Kumar, and Saurabh Pal. "Data mining: A prediction for performance improvement of engineering students using classification." *arXiv preprint arXiv:1203.3832* (2012).

[23] Vera, C.M., Morales, C.R. and Soto, S.V., 2013. Predicting school failure and dropout by using data mining techniques. *IEEE Journal of Latin-American Learning Technologie*, *8*(1), pp.7-14.

[24] Wolff, A., Zdrahal, Z., Nikolov, A. and Pantucek, M., 2013, April. Improving retention: predicting at-risk students by analysing clicking behaviour in a virtual learning environment. In *Proceedings of the third international conference on learning analytics and knowledge* (pp. 145-149). ACM.

[25] أبتسام جعفر محمد، ٢٠١٥. *التسرب الدراسي وأثره علي العملية التعليمية بمرحلة تعليم الأساس بمحلية مروي بالولاية الشمالية في الفترة من ١٩٩٩ م -٢٠٠٣ م* (Doctoral dissertation, UOFKK).

[26] Pradeep, A., Das, S. and Kizhekkethottam, J.J., 2015, February. Students dropout factor prediction using EDM techniques. In *Soft-Computing and Networks Security (ICSNS), 2015 International Conference on* (pp. 1-7). IEEE.

[27] Heredia D, Amaya Y, Barrientos E. Student dropout predictive model using data mining techniques. IEEE Latin America Transactions. 2015 Sep;13(9):3127-34.

[28] Pehlivan, Z., 2011. Absenteeism at state high schools and related school management policies in Turkey (Ankara Case). *Procedia-Social and Behavioral Sciences*, *15*, pp.3121-3126.

[29] محمود محمد محمود مصطفى، ٢٠١٣. فاعلية برنامج إرشادي لتعديل الاتجاه نحو المدرسة لدى التلاميذ متكرري الغياب بالمرحلة الإعدادية *CU Theses*.

[30] Hasbun, T., Araya, A. and Villalon, J., 2016, July. Extracurricular activities as dropout prediction factors in higher education using decision trees. In *Advanced Learning Technologies (ICALT), 2016 IEEE 16th International Conference on* (pp. 242-244). IEEE.

[31] Peralta, B., Poblete, T. and Caro, L., 2016, October. Automatic feature selection for desertion and graduation prediction: A chilean case. In *Computer Science Society (SCCC), 2016 35th International Conference of the Chilean* (pp. 1-8). IEEE.

[32] dos Santos Alencar, Márcio Aurélio, Eulanda Miranda dos Santos, and José Francisco de Magalhães Netto. "Identifying Students with Evasion Risk Using Data Mining."

[33] Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J., 2016. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

APPENDIX

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Hayder Abdulabbas Waheed

Nationality: Iraqi

Date and Place of Birth: 11.April.1971, Baghdad, Iraq

Marital status: Married

Phone: 009647722233365

E-mail: hydar767@gmail.com

EDUCATION

Degree	Institution	Year of Graduation
M.Sc.	Çankaya University Mathematics and Computer Science	2017
B.Sc.	AL- Mustansiriya University	1989
High School	Al-Markaziya School	1988

WORK EXPERIENCE

Year	Place	Occupation
2004- Present	Ministry of Education	Computer Department Manager