Evaluation of Data Mining for Two Child-related, Social Risk Issues

James Little[1], Hayder A. Waheed[2] and Andy Rixon[3]

[1] CONNECT Centre, Trinity College Dublin, Ireland
[2] Department of Mathematics, Çankaya University, Ankara, Turkey
[3] School of Health, Wellbeing and Social Care, Open University, UK
jalittle@tcd.ie

Abstract. Two child-related social issues are examined using data mining to determine successful ways of predicting risk. The issues of child truancy and child abuse can be considered similar as both are influenced by, the child's characteristics, family and environment. The results show that from an initial portfolio of algorithms, a one-nearest neighbour approach works well. We believe that reflects the nature of the problem, where expert opinion classifies each new pupil /case in terms of similar ones, while the one-nearest aspect, reflects the small amount of data we had access to.

Keywords: Truancy, Child Abuse, Social AI, Data Mining, Risk

1 Introduction

Child truancy and child abuse are both complex social issues affecting modern society. Both are monitored and assessed by teachers and social workers respectively. They are closely related, because non-attendance at school may be an indicator of neglect.

Many studies have been conducted on their causes, but this research looks at how to use these causes in a data mining model, to predict the risk of the child to truancy or abuse. The question we address, is whether a data mining approach can claim an accuracy in prediction and given the common background to the problems, whether this is also reflected in a similar type model. The actual adoption of the model in real-life is the next consideration. We see data mining only as a support to decision makers in structuring the issue, ensuring consistency between pupils/cases and suggesting areas of intervention.

We use two appropriate sources of data – both quite small, but believed to be representative. The first is truancy, where 30 pupil records from primary, middle and high schools in Iraq were used. These schools are male-only so the analysis covers only boys. The schools keep extensive paper records containing many social, personal and academic observations on the pupil throughout their education. These include whether they truanted or not. This data is taken from the thesis of Wheed [1].

The second source of data is taken from the paper by Little and Rixon [2], which had 20 child abuse cases classified on eight factors against risk. These had been taken from structured interviews with social workers. All these factors are subjective and high level, requiring an expert to balance many aspects of the case, before setting a value. In that paper, the authors only looked at a single type of model based on the ID3 algorithm. This paper re-examines the problem, using the same data, but with a portfolio of data mining algorithms; the aim to determine the one with the highest accuracy.

The factors for both truancy and abuse can be roughly be categorized as, the innate characteristics of the child, factors relating to the immediate family and factors relating to the wider environment [3, 4]. One difference though is in the way they measure risk. With truancy, it is the number of days missing, before a child is considered to have 'truanted'. Whereas with social work, the risk level is set by the team or individual social worker.

2 Social Domains

We consider two related areas of child risk, that of truancy and child abuse. Both have been studied for the causal factors to this risk, but there are few data mining papers on the topics. We look at the two domains in terms of the related work that has been carried out.

2.1 Truancy

Truancy can be defined as the absence from school without a valid reason. This a global social issue affecting many countries such as America, UK and Canada [5,6]. All too often, research focuses on the remedies which are imposed after truancy has occurred. Williams [7] describe systems which will "send letters and/or phone parents when students have crossed the threshold set by the school system as an unacceptable number of absences". The causal factors are a well-studied and papers such as Corville-Smith et al [4] identify them, such as students' school perceptions, perception of parental discipline, level of parents' control, students' self-concept, family conflict and social competence in class. No investigation of the relationship is made nor of their weighting. Wi et al [8] identified statistically several significant factors all coming from family and school, namely parents' marital status, space at home, feeling towards school and uncertainty over the reason for school. An interesting mixture of subjective and objective, small and large. Again, inter-relationships are not investigated.

Pupil performance may draw on the same factors as truancy. Pradeep et al [9] studied this area in much the same way we have with truancy, through evaluating different data mining models. Their sample size was of 670 students with 57 factors. The chosen factors came from areas of psychology, family, social and cultural background, scholastic progress, demography and socioeconomic. They did use a selection algorithm to reduce the factors from 57 to just 12, which is much closer to our size. Equally, as an initial study we focussed on a good set of representative data – so equal numbers of truancy as non-truancy. Padeep et al instead generated data to give this balance. The main difference though came in them only focusing on data mining algorithms which produced easy to understand results i.e. rule induction or decision trees. Therefore, they did not choose other common data mining techniques for potentially better accuracy, as we did. We cannot therefore make a direct comparison with our best approach.

Williams [7] alludes to an IBM system in 2000 which collected data including reading scores and from which patterns were traced using data mining. There is little evidence of this being taken further.

2.1.1 Truancy factors

The following seven factors are selected for our research based on 1. available data, 2. literature review and 3. discussions with educational experts in Iraq. One new factor we had not seen elsewhere was the security situation in Iraq. This was considered significant in advance by the education experts and is one which does not crop up in any other study. All the factors could also be measured directly from data in the school report card (see Appendix A).

Personality

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 or not. If the value is **active**, then this means that the pupil is interested in participating in extracurricular activities. If the activity value is **inactive**, then this means that the pupil participates very little.

Personality (introvert, balanced)

This factor relates to the personality of the pupil and his dealings with the teachers and his peers. If the personality is **introvert**, then this means that the pupil is not compatible with most pupils. He does not interact well with teachers

and may feel afraid of them. This factor shows a detachment with school affairs or fear of the people and hence a greater likelihood towards staying away. If his personality is balanced, then this means he is compatible with his fellow pupils and teachers.

Health (sick, well)

This factor relates to the pupil's health. If the pupil's health is **sick**, then this means that the pupil can have one of two conditions. The first is that the pupil has a psychological condition. The pupil may have an inability to deal with friends and sometimes pupils are afraid to get close. The second is that the pupil has a chronic illness, such as diabetes or a heart condition. The illness leads to the need for regular visits to the doctor. It may also require medication at specific times. Although these are quite different conditions, they both create opportunities for truancy. If the pupil's health situation is **well**, then this means that the pupil does not have any of the conditions mentioned above.

Family

Family situation (stable, troubled) – If the family situation is **stable**, then this means that the parents are alive, have few problems between them and the pupil is living at home with them. The parents provide all emotional and material needs. If the family situation is **troubled**, then this means that one or both parents of the pupil are deceased, or they are separated.

Parents' academic level (high, low) - If the parents' academic level is **high**, then this 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, so they encourage their son to study. If the parents' academic level is **low**, then this means that both parents do not have a university degree, so they do not place such a high value on education.

Family income (poor, suitable) - If the family income is **poor**, then this is less than $500 per month. If the family income is **suitable**, then this is greater than $500 per month.

Environment

Security situation (negative, positive) - This factor covers diverse issues of, the distance of the pupil's residence from their school, the parents' career and the potential sectarian threat. These are all related to the deteriorating security situation in Iraq. If the security situation is **negative**, then the pupil's home is far from the school, implying a greater possibility to be exposed to terrorist acts along the route. Alternatively, if one or both parents are working in a security role for the state, then this increases the likelihood of terrorism towards the family. Finally, if the family is suffering from sectarian conflicts, this may impact the security of the family. If the security situation is **positive**, then this means a lack of all the three issues mentioned above.

Class Factor

Risk of Truancy (high, low)

The value of risk is **high**, then it means that there is a strong likelihood of truancy taking place. Truancy is measured in Iraq through the number of days taken off illegally. If the pupil is absent for more than 15 days without reason per year, then they are in a truancy state. The value of class **low**, then this means recorded absence is less than 15 days per year or none. This is an exact, but subjective measure to identify truancy after the event. We are looking to predict and hence prevent truancy before the pupil reaches the 15 days.

2.2 Child Abuse

For child protection social workers, risk assessment is a daily activity of central importance. Yet this activity is carried out manually. Major efforts have been made to create models for risk assessment, but in 2014 Munro et al [10] concluded that "it remains an imprecise science" and further there are still no clear factors which "appear to be neither necessary nor sufficient conditions for maltreatment (abuse) to occur". Research continues, but the feeling remains that it is hard to get an analytic approach adopted, for an area which is commonly believed to only require

human judgement. Munro et al point this out in saying, “Actuarial and consensus-based risk instruments have been developed and used in many jurisdictions. However, most have not been tested and those that have, reveal levels of specificity and sensitivity that are so low that they raise questions about the ethics of using them in practice”. Even identifying significant factors has not lead to any consensus. Again, by Munro et al, “Lists vary in content and differ in length, with most offering little advice on how factors should be weighted.” Dorsey et al [11] even suggest the experts find it hard to judge the risk and they need “to improve accuracy and identification of cases most at risk”.

An early review of risk assessment tools in 2009 identified 74 different ones [12]. Barlow et al [13] made a more recent study of available tools for assessing/analysing data for the likelihood of significant harm to children. They identified many models for Structured Decision-making which are based on statistical approaches (actuarial tools), checklists and scoring systems. The most widely used included the Child Abuse Potential Inventory, the Child Well-Being Scales and the Child Behavior Checklist. However, there are very few examples based on Artificial Intelligence. Vaithianathan et al [14] was the most recent one, where they modelled the risk through a predictive model, based on data mining. However, their approach was of low knowledge whereby 224 predictor variables about the child, from the benefit and child protection records, were initially used to determine risk scores for ‘substantiated maltreatment’. Only one algorithm seems to have been tried. Our approach though could be considered high knowledge by involving the expert practitioner, in identifying the important factors, which may in themselves require many lower level factors to be weighed and aggregated.

Therefore, we believe that the approach by Little and Rixon [2] continues to be of relevance, as the factors are still present. Child abuse may be assessed in slightly different ways now, compared to 1998, due to on-going research. Yet, the data they used reflected the social beliefs of their practice at that time. Their focus on a single classifier (ID3) demonstrated the usefulness of the approach in the form of an easy to understand decision tree, with the most important characteristics nearer the root of the tree. We take this research further to consider other classifiers which may give better accuracy.

2.2.1 Child abuse factors

Eight different factors have been identified and subdivided across the three main categories. All the factors were determined by the social worker and based on subjective knowledge of the case. The data on the cases is taken from descriptions and reports, but requires interpretation of it by the social worker. This contrasts with the truancy where a reporting structure provides a single point of reference.

Type of abuse (physical, emotional, neglect) **-** The type characteristic indicates one of the three categories of child abuse set out in the Department of Health’s Working Together [15].

Personality

Impact of the child’s behaviour (strong, weak, none) - This factor measures to what extent the child's behaviour might add to the risk. For example, a child crying a lot or having frequent tantrums may be having an impact on the situation.

Vulnerability of child (high, medium, low) - Some children are seen to be more vulnerable to abuse for purely passive reasons, such as age, disability, etc. It is thought that the age of the child could be particularly significant.

Family

Attitude of carers to child (positive, ambivalent, negative, strongly negative) - The carers' attitude is the cumulative measure of the attitudes of each carer towards the child. An example of a negative attitude towards children are those who are `scapegoated' within the family, for some reason. This characteristic consequently can vary considerably between each child in the family.

Ability to do the caring (positive, adequate, negative, very negative) - This factor measures cumulatively how well the carers can meet the needs of the child. There is a range of factors which influences this ability: maturity, drug dependence, isolation, carers' relationship to each other, etc. Measuring this characteristic in a single value is extremely difficult when there can be so many potential issues.

Abusers immediacy (high, medium, low, none) – The immediacy characteristic is a measure of how `near' the abuser is to the child. It covered both the physical proximity and access to the child. The immediacy value is reduced if there are other supervising influences present, with the ability to monitor and control the interaction.

Previous record of abuse (highly negative, negative, none, positive, unknown) - This factor relates to the relevant history of the carers before the current assessment. This considers the time, severity and frequency of past incidents. Similar or escalating patterns point to a negative value.

Environment

Seriousness (high, medium, low, no evidence) – This factor is a measure of how serious the incident precipitating the current investigation was. For cases of physical abuse, more serious injuries are obviously seen as `high', but equally important is the way in which the injury was caused. More `deliberate' injuries attract higher ratings. However, the multidisciplinary nature of the assessment process often helped to determine the value, by incorporating information from other sources, e.g. about a child's height and weight or school performance, truancy, etc.

Class Factor

Level of Risk (high, medium, low)

The risk characteristic indicated the outcome of the decision process. Three possible values were identified by the social work team. These were aggregated later for the analysis.

3 Methodology

We adopt the same methodology for each problem domain, whereby we start by discussing with experts in the field and/or reviewing the literature for factors which are relevant to our problems. Next, with a different set of experts (social workers and teachers) we fill in the values for the factors on a set of real examples (cases and pupils).

The data for child abuse is taken directly from the Little and Rixon paper, but to be consistent for both approaches, we consider the predicted risk to be only one of two levels. So, the values were converted to High and Med/Low.

The full set of input data for child abuse cases is shown in Table 1.

Table 1. Child abuse cases characterised by factors and values

Case no	Type	Child Impact	Vulnerability	Carers' Attitude	Carers' Ability	Immediacy	Previous	Seriousness	Risk
1	Physical	None	Medium	Positive	Adequate	High	Negative	Low	Med/Low
2	Physical	None	High	Very negative	Negative	High	None	No evidence	High
3	Physical	Strong	High	Positive	Negative	Medium	Very negative	No evidence	Med/Low
4	Physical	None	High	Positive	Positive	High	Very negative	Low	Med/Low
5	Physical	None	High	Positive	Negative	Low	Positive	High	Med/Low
6	Physical	Weak	High	Positive	Adequate	Medium	Negative	No evidence	Med/Low
7	Physical	Weak	High	Positive	Very negative	High	Negative	High	High
8	Physical	Weak	High	Ambivalent	Negative	High	Negative	Medium	High
9	Physical	Weak	High	Positive	Very negative	High	Very negative	No evidence	High
10	Physical	Strong	Medium	Ambivalent	Positive	High	Positive	Low	Med/Low
11	Neglect	Weak	High	Negative	Very negative	High	Negative	Medium	High
12	Neglect	None	Medium	Ambivalent	Negative	High	Negative	Low	Med/Low
13	Physical	High	Medium	Positive	Negative	High	None	Low	Med/Low

14	Physical	High	Medium	Negative	Adequate	High	None	Medium	Med/Low
15	Emotional	None	High	Negative	Very negative	High	Negative	Medium	High
16	Emotional	High	High	Very negative	Negative	Medium	Unknown	High	High
17	Emotional	None	Low	Ambivalent	Negative	Medium	Unknown	Low	Med/Low
18	Emotional	High	Low	Negative	Negative	Medium	Unknown	High	High
19	Emotional	Strong	Medium	Positive	Adequate	Medium	Negative	Medium	Med/Low
20	Physical	None	High	Very negative	Adequate	Medium	Negative	Low	Med/Low

A set of input data for the child truancy cases is shown in Table 2.

Table 2. Pupils characterised by factors and values related to truancy

Pupil no	Family situation	Parents academic	Security situation	Family Income	Activity	Personal	Health	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

A range of common prediction algorithms were used, representing a cross section of approaches. These were Naïve Bayes, J48 decision tree, SMO support vector machine, IBK-1 & 2 nearest neighbour and the OneR single rule. The software used was Weka [16] and accuracy was measured as the percentage of correctly predicted instances.

4 Results

The results for truancy prediction across six different prediction algorithms are shown in Table 3.

Table 3. Accuracy across different prediction models on truancy data

Evaluation criteria	Prediction algorithm					
	NB	SMO	IBK-K=1	IBK-K=2	OneR	DT J48
Accuracy (%)	70%	66.7%	90%	56.7%	36.7%	73.3%

The IBK-K=1 nearest neighbour algorithm dominates all other approaches. The imposition of clear, measurable truancy levels and a representative set of examples with good measurements, means it is very appropriate for this type of algorithm of selecting from the nearest neighbour. Significantly the same algorithm with increased nearest number of neighbours breaks down, as the data set is so small, the next nearest neighbour can be far away with a completely different risk level. The most important attribute using Select Filter Attribute was family situation and security situation.

The results for child abuse prediction across seven different prediction algorithms are shown in Table 4.

Table 4. Accuracy across different prediction Models for Child Abuse Data

Evaluation criteria	Prediction algorithm						
	NB	SMO	IBK-K=1	IBK-K=2	OneR	DT J48	ID3
Accuracy (%)	70%	70%	75%	70%	55%	45%	50%

The IBK-K=1 nearest neighbour algorithm is again the best algorithm, but not quite so emphatically. There is a case to suggest that the decision-making process for social workers also uses similar cases for guidance. There is less clarity here of what constitutes high risk and some of the factors are very subjective leading to perhaps a lower accuracy than that of truancy. The most important attribute using Select Filter Attribute was carers attitude and carers ability this confirmed with the findings of the original paper as the first nodes of the ID3 tree.

In passing, we included the ID3 algorithm to show that although in the original paper it gave the user a clear understanding of how it made the risk calculation, the predictions were not good.

5 Conclusions

We can conclude that both social areas are amenable to a prediction approach, but deploying it in a real-life context is quite different. The current decision-making is highly subjective, complex, requiring experience and the understanding of human interactions. The best we can offer is that our approach complements the work professionals are doing and can show points of intervention to prevent situations worsening. In the child abuse case such an approach is always going to be used as a backstop to any decision making by providing a framework for consistency, rather than as a first stop solution. In the intervening years since the first paper there is evidence that this approach is still being researched, but deployment or casting in in a real context appears missing.

It is not apparent if truancy is predicted in these Iraqi schools – it is certainly recognised when it has occurred, but not before. Therefore, in this area there may be a stronger case for such an approach, but always to guide the professionals to focus on certain individuals and factors. Unfortunately, one of the main factor is the security environment in Iraq, which is not one the schools have much control over - nor the family situation, although less so.

The success of the IBK-K=1 approach in both problems shows that this model works well across two similar, but distinct social issues. Both domains are highly subjective. The nearest neighbour approach works on the basis that if a test case is much like another, then assign it that risk level. This may be the way social worker approach the child abuse issue and it may be by measuring truancy in number of days also allows a simple comparison between pupils to predict the right outcome. Since there is so little data, expanding the approach to IBK-K=2 breaks down as the example space is too diverse, in other words the second nearest example is quite different to the first and should carry no weight.

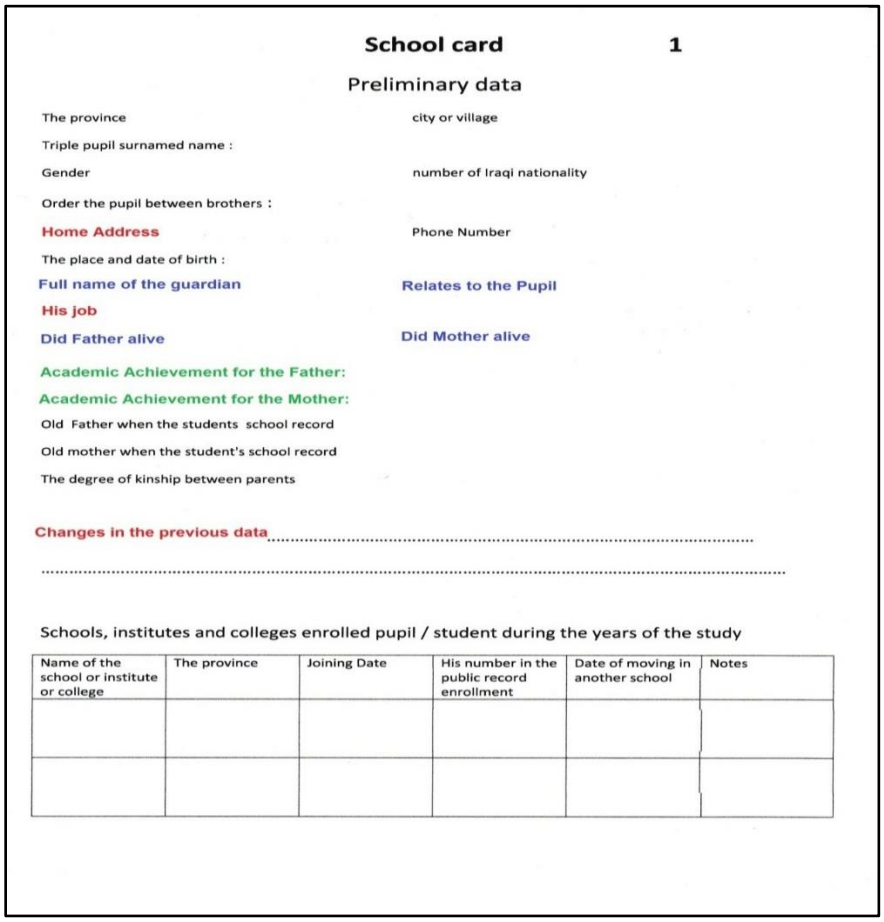
References

1. Wheed, H. A. Identifying the factors influencing truancy and the methods for prediction. MSc Thesis, Department of Mathematics, Çankaya University (2017).
2. Little, J., Rixon, A.: Computer learning and risk assessment in child protection. Child Abuse Review 7(3), 165-177 (1998).
3. Sharon, V. Petch, A.: Understanding child, family, environmental and agency risk factors: findings from an analysis of significant case reviews in Scotland. Child & Family Social Work 22(2), 741-750 (2017).
4. Corville-Smith, J., Ryan, B.A., Adams, G.R. and Dalicandro, T.: Distinguishing absentee students from regular attenders: The combined influence of personal, family, and school factors. Journal of Youth and Adolescence 27(5), 629-640 (1998).

5. Davies, J. D., Lee, J.: To attend or not to attend? Why some students choose school and others reject it. Support for Learning, 21(4), 204–209 (2006).
6. Maynard, B.R., McCrea, K.T., Pigott, T.D., Kelly, M.S.: Indicated truancy interventions: Effects on school attendance among chronic truant students. Campbell Systematic Reviews 10, (2012).
7. Williams, L.L.: Student absenteeism and truancy: Technologies and interventions to reduce and prevent chronic problems among school-age children. Action Research Exchange 1(1),1-4 (2002).
8. Wi, W. S., Iryani, T., Rozhan, M. R., Shamsul, A. S., Zasmani, S.: Psychosocial factors influencing truancy in high risk secondary schools in Kuala Lumpur. Malaysian Journal of Psychiatry 18(2) (2009).
9. Pradeep, A., Das, S., Kizhekkethottam, J. J.: Students dropout factor prediction using EDM techniques. In IEEE International Conference on Soft-Computing and Networks Security, pp. 1-7, (2015).
10. Munro, E., Taylor, J. S., Bradbury-Jones, C.: Understanding the causal pathways to child maltreatment: Implications for health and social care policy and practice. Child abuse review 23(1), 61-74 (2014).
11. Dorsey, S., Mustillo, S. A., Farmer, E. M., Elbogen, E.: Caseworker assessments of risk for recurrent maltreatment: Association with case-specific risk factors and re-reports. Child abuse & neglect 32(3), 377-391 (2008).
12. Daniel, B., Taylor, J., Scott, J.: Recognition of neglect and early response: overview of a systematic review of the literature. Child & Family Social Work. 1;15(2), 248-57 (2010).
13. Barlow, J., Fisher, J. D., Jones, D.: Systematic review of models of analysing significant harm. (2012).
14. Vaithianathan, R., Maloney, T., Putnam-Hornstein, E., Jiang, N.: Children in the public benefit system at risk of maltreatment: Identification via predictive modeling. American journal of preventive medicine 45(3), 354-359 (2013).
15. Home Office, Great Britain: Working together under the Children Act 1989: a guide to arrangements for inter-agency co-operation for the protection of children from abuse. HMSO (1991).
16. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H.: The WEKA data mining software: An update. ACM SIGKDD explorations newsletter 11(1), 10-18 (2009).

Appendix A

School report card, selected pages (translated)



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

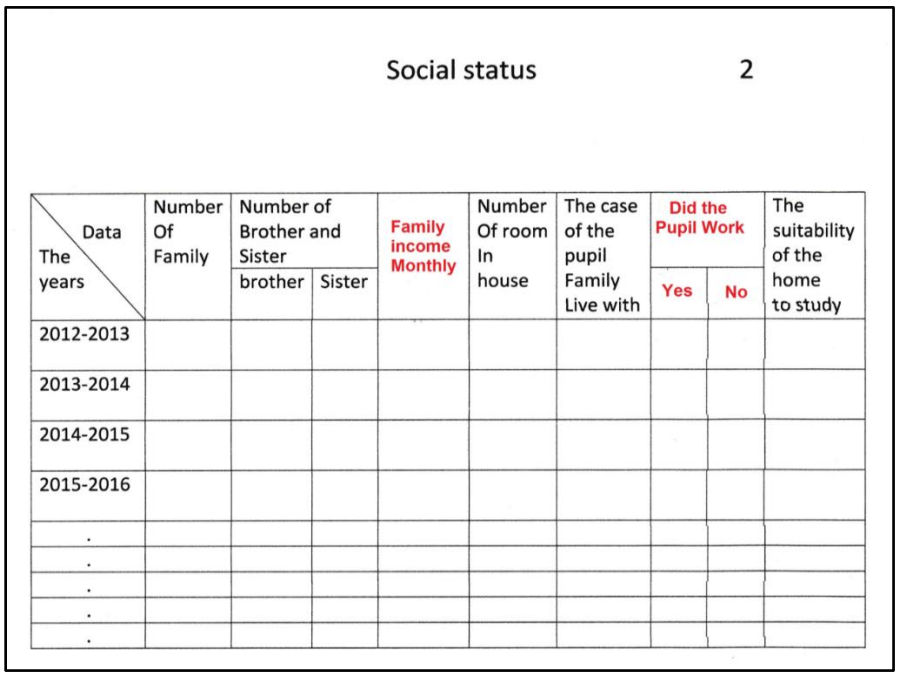
The degree of kinship between parents

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



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									
.									
.									
.									
.									
.									