**ADM3308: Business Data Mining**

**Data Mining Project Using IBM SPSS Modeler**

**(Team work)**

**Important Note:  Read the following academic integrity statement, type in your full name and student ID, and include a copy in your submission. Submitting this form electronically by the team representative is considered as signing the document by BOTH members of the team.**

**Personal Ethics & Academic Integrity Statement**

Student name:   Ankica Basar                                                    Student ID: 7342324

Student Name:   Blake Bolkovic                                                     Student ID: 7852594

By typing in my name and student ID on this form and submitting it electronically, I am attesting to the fact that I have reviewed not only my own work, but the work of my team member, in its entirety.

I attest to the fact that my own work in this project adheres to the fraud policies as outlined in the Academic Regulations in the University’s Undergraduate Studies Calendar. I further attest that I have knowledge of and have respected the “Beware of Plagiarism” brochure found on the Telfer School of Management’s doc-depot site. To the best of my knowledge, I also believe that each of my group colleagues has also met the aforementioned requirements and regulations. I understand that if my group assignment is submitted without a completed copy of this Personal Work Statement from each group member, it will be interpreted by the school that the missing student(s) name is confirmation of non-participation of the aforementioned student(s) in the required work.

We, by typing in our names and student IDs on this form and submitting it electronically,

·         warrant that the work submitted herein is our own group members’ work and not the work of others

·         acknowledge that we have read and understood the University Regulations on Academic Misconduct

·         acknowledge that it is a breach of University Regulations to give or receive unauthorized and/or unacknowledged assistance on a graded piece of work

**ADM 3308 M - Business Data Mining Project**

**Dataset 11: Nursery**

**Professor: Bijan Raahemi**

**By:**

**Ankica Basar, 7342324**

**Blake Bolkovic, 7852594**

**April 12, 2018**

**Abstract**

For a while now, early childhood education has been on the rise. There are many instances in history where children are sent to elite schools, and how the deciding factors in their acceptance relied heavily on their social standing . Our model involves a dataset which  ranked nursery school applications based on whether or not children were eligible for certain schools. This eligibility was dependent on several factors, such as parental income and family structure. This nursery data was previously used to create a hierarchical decision tree to make the process of accepting and recommending applications easier in a time when  enrollment to preschools was in high demand. The purpose of our report is to see if we could find any relation in the data about the deciding factors that would indicate any inequality between children recommended. In addition, we attempted to find any other purposeful relation between the attributes and the target that could be useful to people today, when creating new models in regards to  accepting various applications. Our chosen data sample was made up of 12960 applications from families in Slovenia in 1980. The use of the decision tree, association rules, and neural network allowed us to have a deeper understanding of the deciding attributes and how the target values were determined. Our results went against what we originally expected, however both the directed and undirected methods helped provide a wide range of insight that would prove useful to the way profiles for schools could be designed today.

**Introduction**

Before the data and its analysis is explained, some background information on the dataset will help in understanding why it was created. This nursery database was derived from a hierarchical decision model originally developed to rank applications for nursery schools. During the 1980’s, there was an excessive enrollment to nurseries in Ljubljana, Slovenia. Since these schools could not accept all of the children, many had to have their application turned down. To make sure the selection process was fair, rejections needed an objective explanation as to why they were turned down. By analyzing the data, we will be able to determine which factors were the most important in deciding whether a child was to be admitted to a nursery or rejected.

The data set for our model contained 8 attributes that were based off of a decision model that was used to reject or accept preschool applications.  The 8 attributes represent the following:

1. Parents:    Parents' occupation, values:  **usual, pretentious, great pret**
2. has\_nurs:       Child's nursery, values: **proper, less proper,improper, critical, very         critical**
3. form:         Form of the family, values: **complete, completed, uncompleted, foster**
4. children:     Number of children, values: **1, 2, 3 or more**
5. housing:        Housing conditions, values: **convenient, less convenient , critical**
6. finance:        Financial standing of the family, values: **convenient, less convenient**
7. social:             Social conditions, values: **non problematic,  slightly problematic,**

**problematic**

1. health:         Health conditions, values: **recommended, priority, not recommended**

  Target value:        Nursery recommendation, values: **not recommended, recommended,**

**very recomended, priority and special priority**

Our data was not difficult to process and clean. We did not have any missing values and the attribute values and target value were very straightforward. For clarification, the target values are ranked by their order of importance from not recommended to  special priority. We would like to note that we were not entirely clear on what all of the rankings meant for the attributes and the target value but it is clear that they are weighted on order of importance. The 8 attributes fell under three main categories described by the creators of the database as employment, structure and social health. It should be noted that we noticed that our health attribute did not solely mean the health of the child it could possibly include the child’s competency and skills. (This is quite important and  crucial to our model, as you will see.)

**Preprocessing Data**

For the chosen dataset, there were 12960 complete samples made of qualitative data. Normalization was not done on any of the data since the variables that were being dealt with were all categorical. It was decided that the main goal was to determine which combination of the 8 attributes would bring the greatest probability of getting into the nursery schools. Since the data provided the outcome for each scenario of whether or not the student was admitted, normalizing would bring no benefit to the analysis. Missing values were not a concern since the database obtained did not contain any. However, there was a need to delimit the data points with spaces in excel before uploading the data.

Once the data was loaded into spss modeler,  the type of the data values needed to be changed from nominal to ordinal.  Each data type was labelled as ordinal because each of the attributes were rankable. As an example, the social health of the family attribute had 3 possible values: non problematic, slightly problematic, and problematic. These values have a hierarchy that puts a children's application with a problematic family environment at a higher chance of receiving admission than an application that had a non problematic family environment. As it could not be determined exactly how much greater these values were compared to each other, it was clear that the values were neither intervals nor categorical (not ordered).

**Methods**

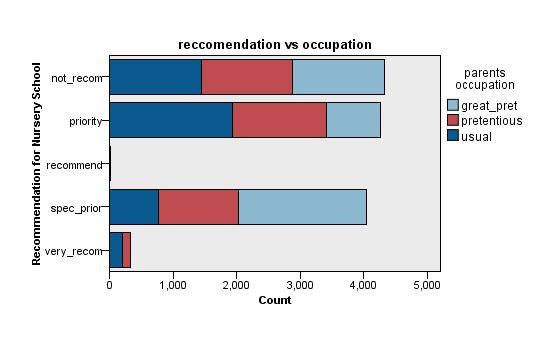
We decided to use three methods, two directed methods and one undirected method, so we could asses patterns in the data from both angles. Since we are looking  to determine what factors (attributes) influence our target value (who is accepted and who is not), we thought it would be best to use classification and clustering techniques so we could get an indepth look of  any patterns in the behavior of different family types, and how it might affect their acceptance to preschool. Our first method is a directed method called the decision tree. It was useful to us as it provided us with more insight as to which factors were weighted as more important when deciding to accept or reject applications.

At first, we had tried clustering and found that clustering itself was not useful enough to interpret our data. This brings us to our second method, the neural network. In this directed technique, the focus is on prediction and classification. We therefore knew that it would handle categorical data effectively. We then decided to try association rules to better understand the behavior within families that were accepted over those who were not.

The third method we decided to use was association rules. It was essential to our understanding of the overall data. We found the description on the collection of the data quite vague, so these rules helped us gain clarity as to how the attributes were related. Below, you will find an in depth description of the results of our methods.

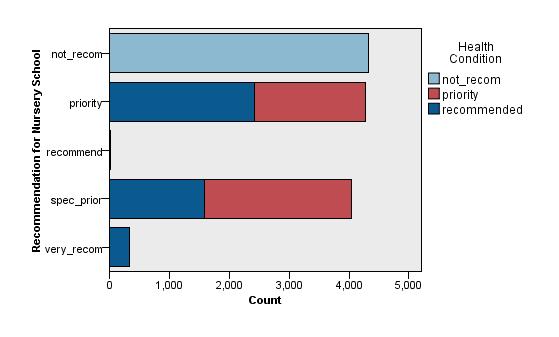
Aside from our 3 modelling methods,  we made three graphs to sort of test our prior assumptions before we began modelling.  We attached two graphs below to see how much of an effect the factors we believed would be important had on the decision making process. We assumed the type of occupation parents vs health/competence of the child,  held more weight in regards to recommendation of pre school. In graph 1 and 2 ,we thought the occupation of parents would hold more weight on the final decision than it did.

**Graph 1**

****

As you can see, even the children who were very recommended for preschool came from homes that did not have any outstanding type of occupation (meaning the income from the job was not anything special but the child was financially cared for). It can be seen that no particular job type in this graph seems to give any proper indication of whether or not the application to the nurseries gets accepted.

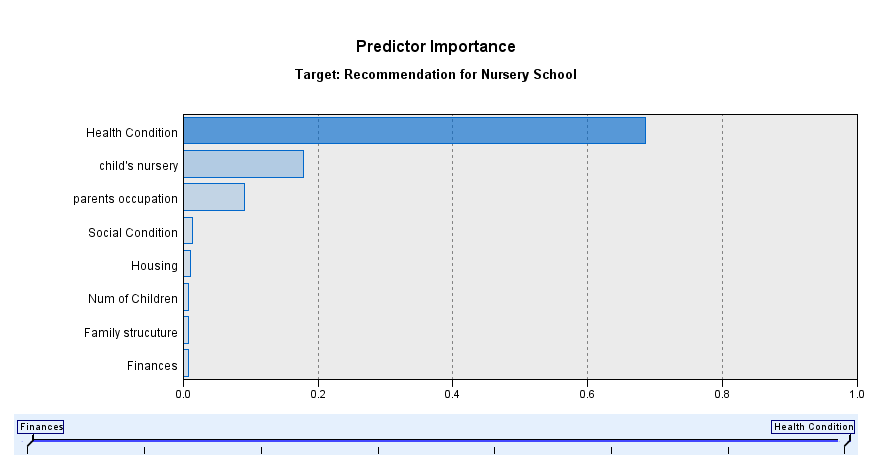
**Graph 2**

****

We then looked at the health attribute and realized that if the children’s health condition was not recommended to justify their presence at the nursery, they would not be recommended for the preschool. It could also be deducted that is the child’s health condition was considered a priority, there would be a greater chance that they would obtain special priority on the acceptance list compared to simply “priority” on the same list. The opposite was true for a child with a “recommended” health condition, as seen in the graph above. The other factors were not so black and white as we thought originally, as the data needed to be processed further to gain more insight.

**1. Decision Tree**

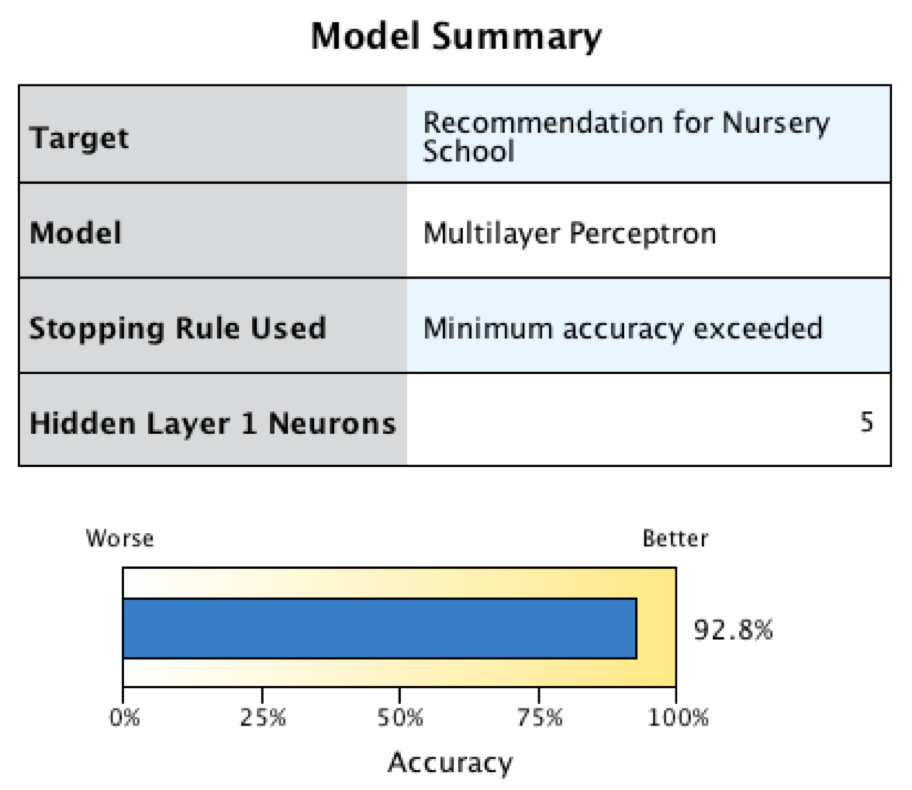
In regards to our model, this task was performed on our data  as mentioned earlier to give us a clear indication of which attributes were used to make the acceptance/rejection decisions.  We found we had a very hard time understanding how the data was used to rank the 8 different attributes. So with the tree we wanted to be able to understand which  children were classified as accepted or rejected based on the attributes. For the purposes of our model,we split the data 70% training and 30% testing. [we had trouble inserting the actual tree, please see our stream for the tree] Below we have a copy of the Predictor importance gained from the decision tree.

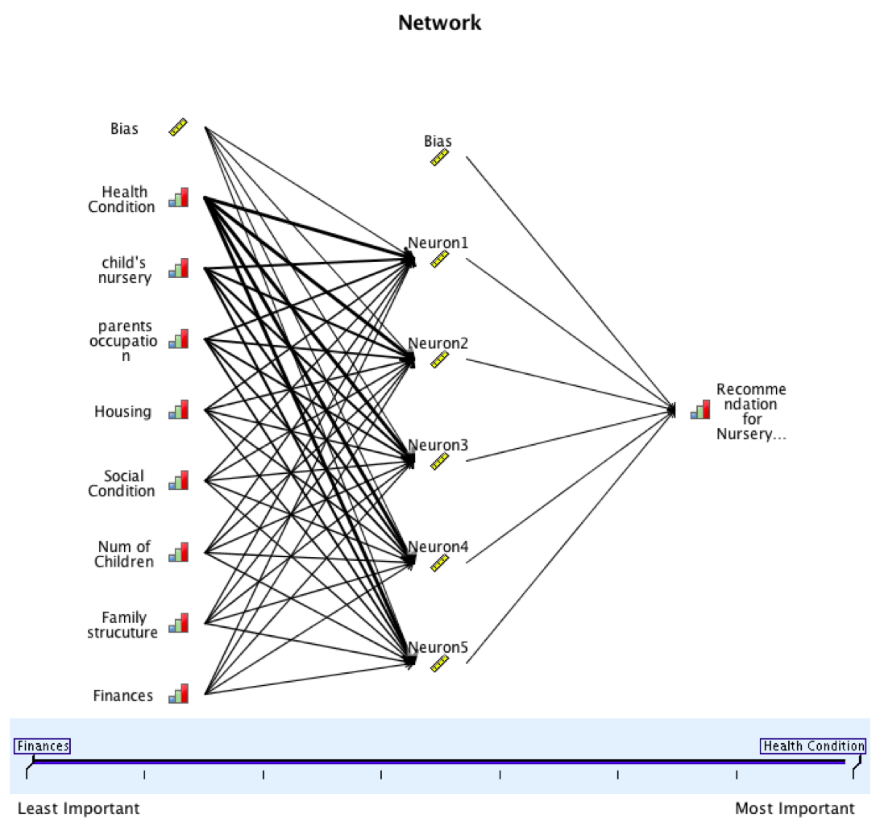


From this chart we are able to understand how the  applications were accepted and rejected since the health condition is most important followed by the conditions of the child’s in home nursery, and the occupation of parents as somewhat important, while every other attribute is not of great importance. The ranking of these attributes became more important than the actual percentage of counts like we had originally looked at in graph 1 and 2.  Moving forward we knew to focus on health as a major predictor of the application status. Moreover, our decision tree model had a correctness of 88.36% . From this we know that our classifiers are accurate since the percentage is so high. Looking at the tree in the stream we can see exactly how much each percentage weighs on the condition and category of each attribute. We found that the decision making process overall was quite fair as you can see in the tree that occupation, social structure and  condition do not place any weight on the final decisions as a deciding factor. The process was focused solely on the terms of the health of the child and if they had an in home nursery (we assume this means that the child had a place to play and was taught some form of schooling at home) . We found that relying on the health of the child gives a fair chance to any applicant. The full layout of the tree made much more sense in terms of what our classifiers were but we still wanted to look for patterns that connected any of the attributes of the families that were accepted. As we modelled other methods we used our tree as a base predictor to see if this correctness would hold in other methods we modelled.

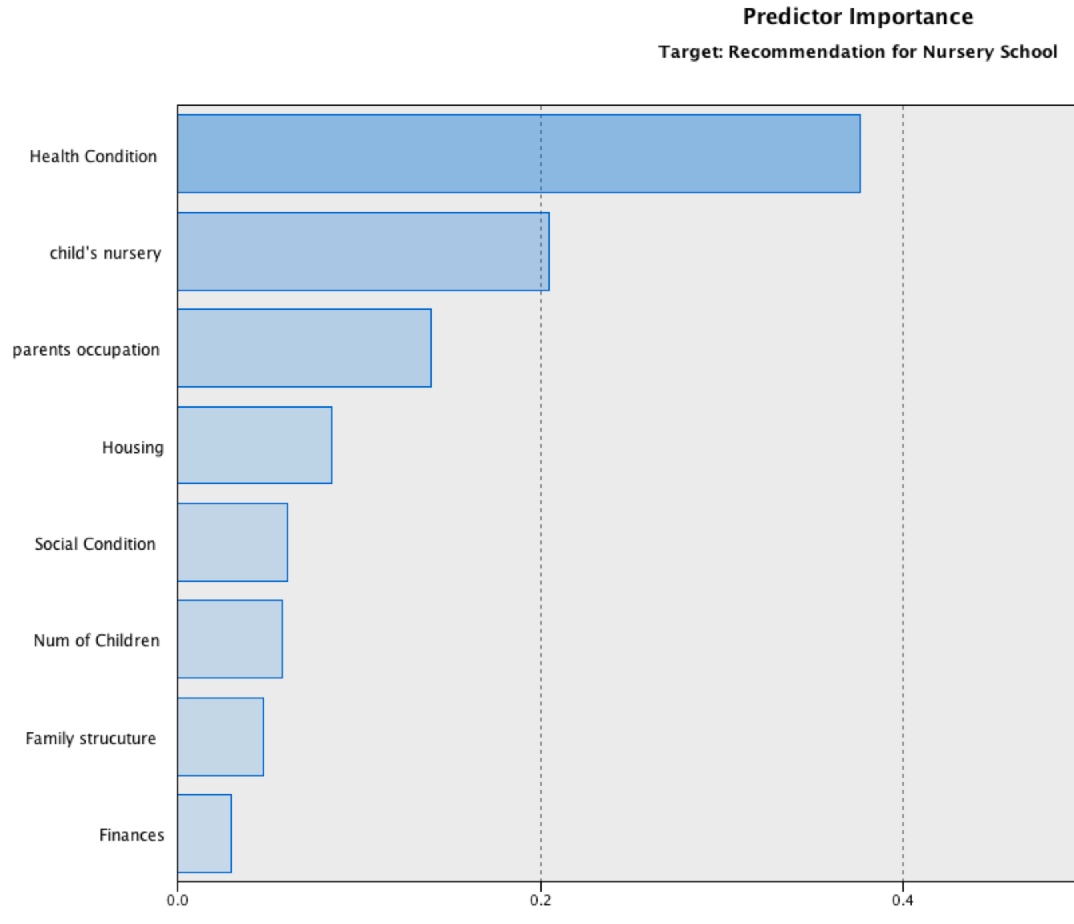
**2. Neural Network**

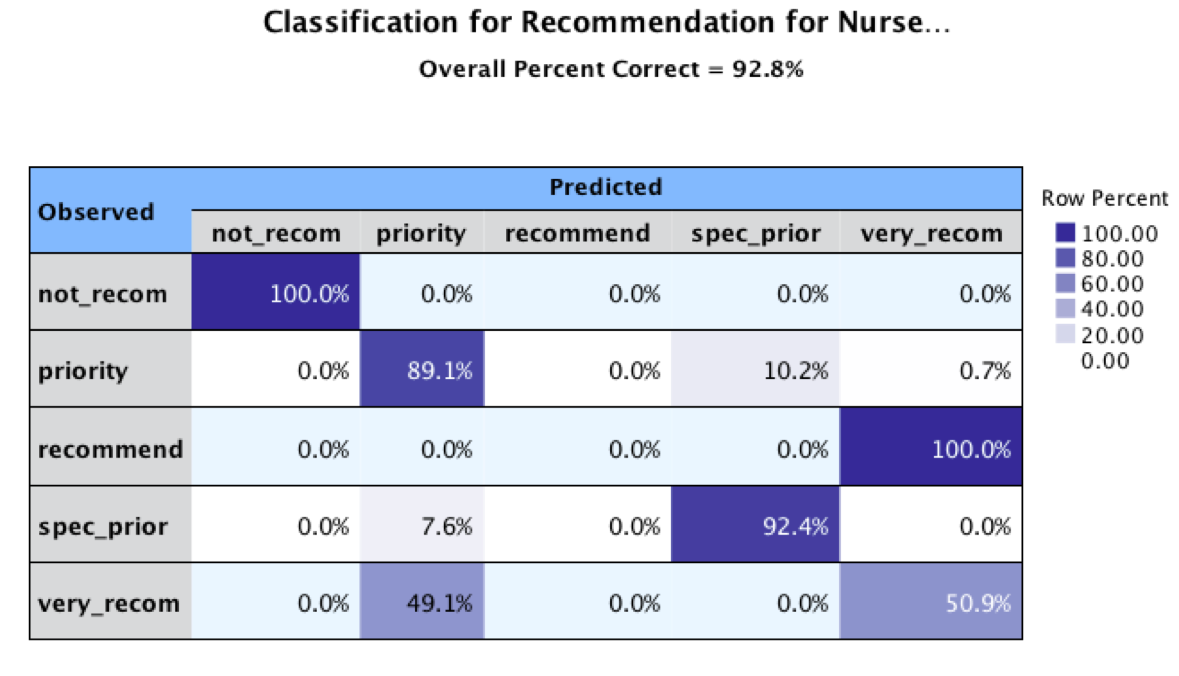
Another data mining task that we used to interpret the data was a predictive neural net model. We used a multilayer perceptron because unlike a simple neural network (no hidden layer), the multilayer perceptron has the extra middle layer. This makes the network more powerful by enabling it to recognize more patterns. When the neural network was initially set up, we let the software decide the number of units in the hidden layer. When the system was done running, it had given us around 20 units for the hidden layer. This was more than double the number of inputs that we had given the network. This increased size of the hidden layer made the network more powerful, but we knew that this would introduce the risk of overfitting. The first neural network that was created had an accuracy of 100%. We knew that this was a sign that the model had been over fitted. We then headed back in to the settings of the NN, and customized the number of neurons to be slightly less than the number of inputs. We also used a stopping rule that would try and keep the accuracy of the model at a minimum of 80%, since this would be an accurate model as well as a model that did not over fit to the data.

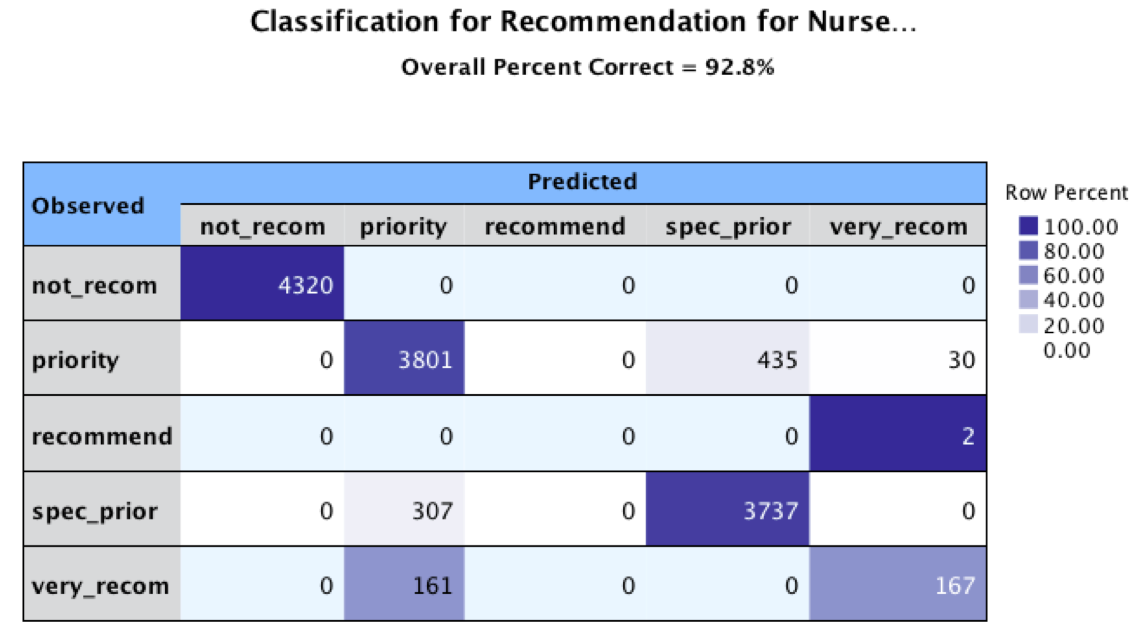




The predictor importance for this model concurred with our other models in the sense that they ranked the attributes in the same order of importance when it came to predicting the recommendation for the nursery school (with the exception of housing and social condition). The children’s health condition was the most important predictor of the outcome for the output as usual, followed by the properness of the child’s nursery. Next in line was the parent’s occupation. These were the three that contained the bulk of the predicting weight, with the rest of the 5 attributes have much less predictor importance.





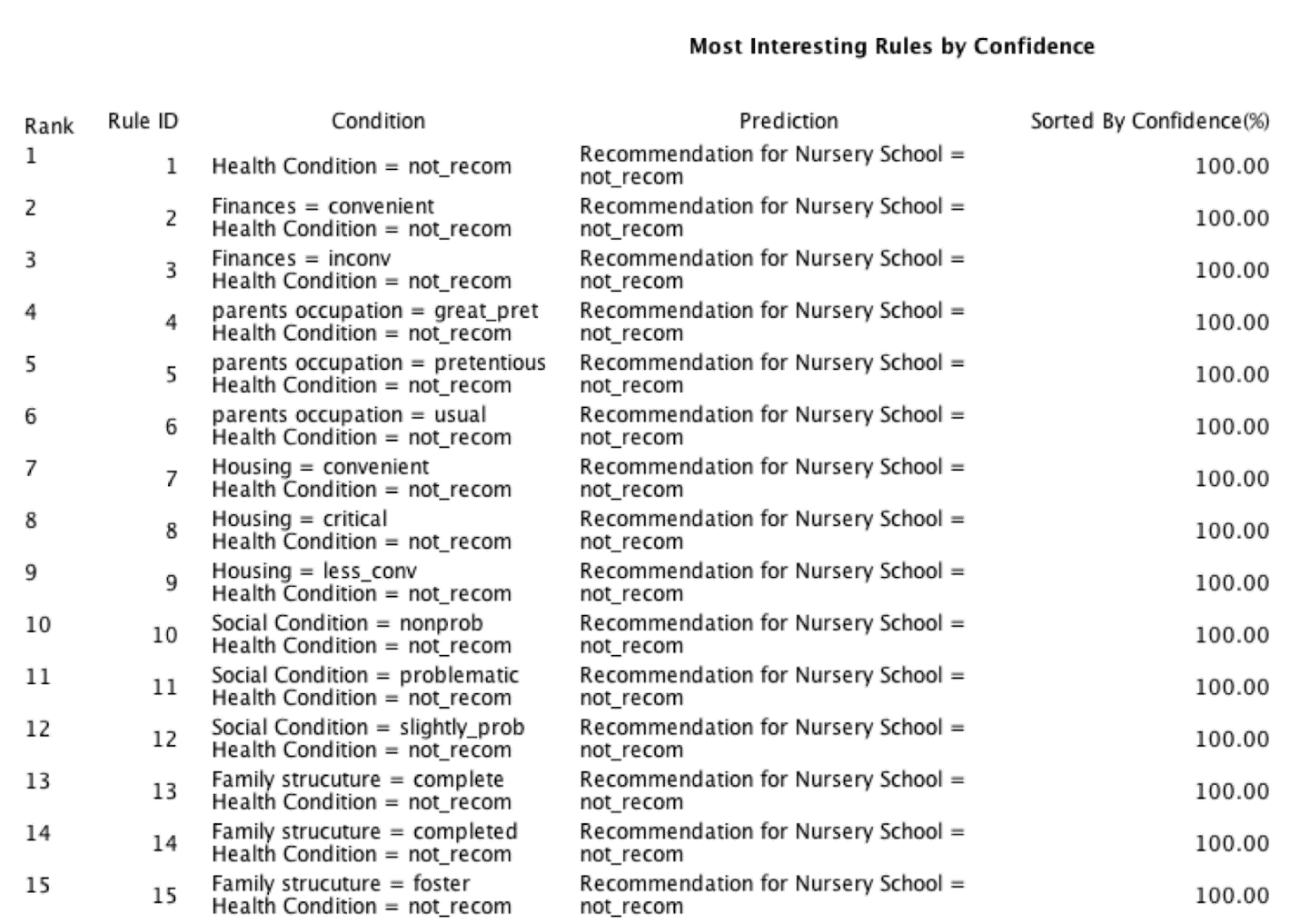


Shown above are the row percents and cell counts for the neural network. In the case of the 100% for the “not\_recom”, it was clear that since the health condition attribute was the most important factor in deciding admittance that a child that’s health was not recommended for the school would have no chance of getting in. For the “recommend” section, there were only 2 values in the whole data set that the health condition for a child was ever recommended. This indicated to us that that 100% did not carry much weight since it was only based on 2 examples. The last case here that will be discussed concerns the “very\_recom” cells. The accuracy seems to be lower than the average here because there were not enough values to give an accurate prediction. In the tables above, we see that the cells that scored high values also contained thousands of values. Since the “very\_recom” cells only have around 200 values, there is less information for the software to work with to create better predictions.

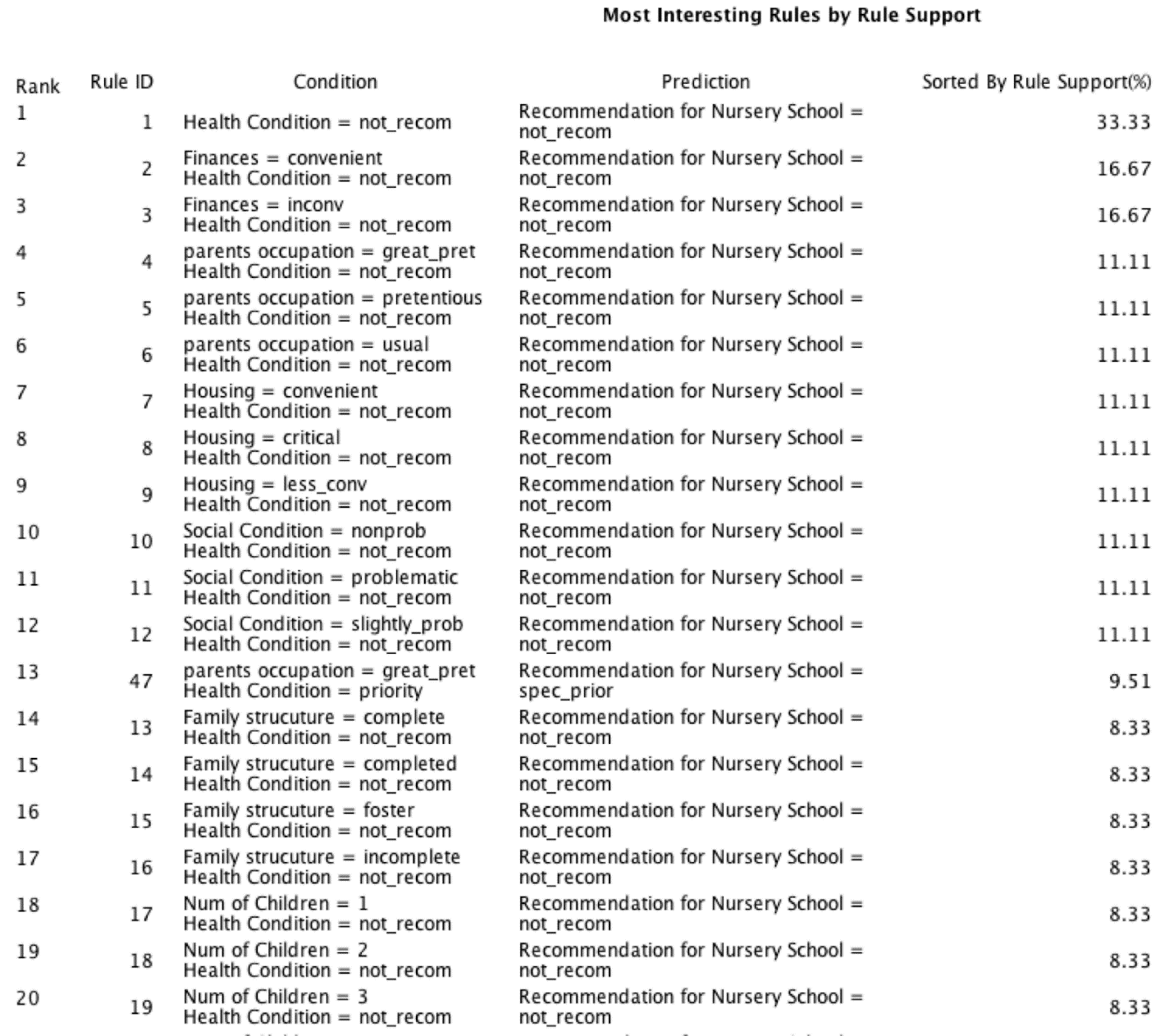
**3. Association Rules**

When the data was being explored in excel, we observed that there were certain patterns that seemed to lead to certain results for the rejection or acceptance to the nursery schools. Finding these patterns would be helpful in any future attempts to decide in exactly what kinds of scenarios children would become accepted into these schools. One way of finding these scenarios is through association rules. In this case, these are clear and actionable rules that can be used to determine the value in the target variable based on the other input variables for each child. For this analysis, the association rules node was used. Once the data was connected to the node, minimum values for the three measures of the rules (support, confidence and lift) were chosen. The eight attributes of the data were chosen as the conditions for the analysis, and the recommendation for nursery school data was chosen as the predictive value. Once the algorithm was finished running, there was valuable information about the most interesting rules from the model. We set the maximum number of rules possible at 1000 to see how many important rules the model would give us. We ended up receiving 55 rules. They were then sorted three times, giving the 30 most important rules in terms of their value for support, confidence and lift. In the tables that will be shown below for the three measures, the rank column indicates the rank for each rule given the rule type, and the rule ID column gives each specific rule a certain ID. These ID’s will be the same for the same rules across the three rankings.

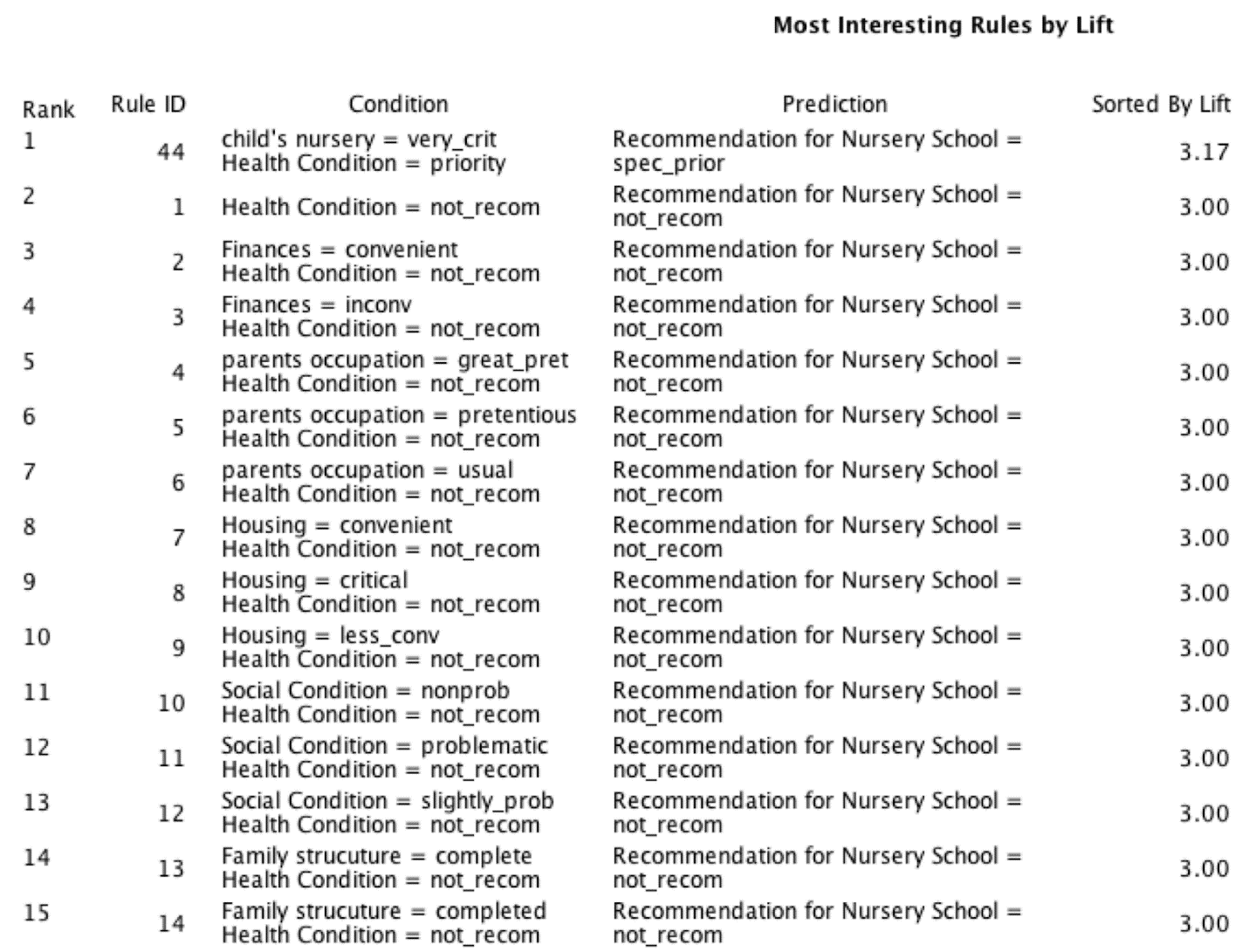
With the rules given for confidence strength, all of them had a score of 100%. This meant that the number of transactions where the conditional part of the rule holds was always the same as the number of transactions where the rule holds. Put simply, if the specific attributes in a child’s file were certain values, you would be able to predict with 100% accuracy that the prediction would be correct. As an example, the *condition* of the first rule ranked in term of confidence was that the child’s health condition was “not recommended” (assuming to mean not suitable for the nursery). If this was the case, the rule would predict that the child would not be recommended for nursery school. Since there was no case where this rule did not hold, it was given a confidence score of 100%. As with the decision tree results explained above, the health condition attribute seemed to have the biggest impact on the prediction. In this case, all of the 30 rules for confidence contained the health attribute. Just like the tree model, the association model confirmed that “health condition” was the most important attribute in deciding whether or not students would be accepted into the nursery schools.



As for the rules given in terms if highest support, most are the same as the confidence rules, with a few exceptions. The level of suppose does however give insightful information on how many transactions support these rules. Using the same rule example of the health condition, the table beneath shows that a third of all transactions were not recommended to the nursery schools because of their health condition. Many rules have the exact same support percentage, and this indicated that certain attributes were linked to each other. We will use the example of the rules which all had a support of 8.33%. Other than the health condition attribute present in all of the rules, these 8 rules contained data on family structures and the number of children. While this might seem like trivial information to those familiar with the nursery business (since it could be taken for granted that these attributes would be related), it proves that the data mining technique is working.



Lastly, the rules were sorted by lift. It should be noted that the lift for all of the rules was above or equal to 3. This is extremely high for lift values since rules with lift over one are considered useful. These rules are therefore much better at predicting the acceptance outcomes of children then random guesses.



All in all, these tables show how good the software was able to pick out affective rules for this dataset, with high values in all the measurements used on the rules.

**Conclusion**

In summary, data was obtained concerning the acceptance of children to certain nursery schools in Slovenia. The original goal was to assess if there was any unfair bias that was giving some children the upper hand in acceptance compared to others. After insightful information was pulled from each model and compared to each other, we came to the conclusion that there was no unfair advantage given to certain students that might have been given an advantage in life. These advantages included aspects such as finance or housing condition. The data revealed that such attributes were ranked by the importance of attributes that were out of the child’s control, such as health and their current nursery. In total, three methods were used, two directed techniques and one undirected.

The tree method was our base that we used to compare the other methods to. It was an effective way to categorize the categorical data that was in the data set. It showed us that certain attributes carried a lot more weight than others in the application decision. The neural network method was used to let the software figure out how to connect the data. The importance of the attributes was very similar to the tree data, and this confirmed that the models were effectively analysing the data. Finally, the rules that were obtained from the association rules model showed us again how the health attribute was integral to the final decision of acceptance, as the attribute was in every single rule. With the lift, confidence and support of each rule giving higher than average scores, it was clear that the rules were effectively describing what was happening with the data.

From our findings, we would like to rule out the theory of any inequality in acceptance of applications. Originally we thought that occupation and social structure would have a greater importance on priority/recommendation. We found that through the classification methods we had a much clearer picture of what determined  acceptance; the health of the child was the final deciding factor as suggested by the confidence score and tree. This increased the fairness of judging applications. In the future when schools need a system to make their application process, we suggest that to increase fairness, the focus should lie on attributes such as competence, health and ability of the  applicant rather than other factors. The rules our models gave for the final decision were logical in the sense that attributes such as number of children, social standing etc did not carry more weight on the final decision. We hope that our model helped bring clarity to the data set and sets a precedent for the set up of future application processing systems.

**Part 2**

**Three pre processing methods used:**

1. We used  the mode or mean depending on the value in the data analysis node, to fill in age,gender, income etc. We individually changed the values  in the data analysis node to ‘fixed’, to fill in the missing value.
2. We also used the  algorithm option to predict the income values that were most probable.
3. We manually changed the options in the type node, as to only have values of yes and no and we  remove any blanks we could.

Once we cleaned the data using the modeler we still had some discrepancies.  We had to clean the one outlier for age manually and had to fix the two negative values for income. We  changed these values to the mode value of 43 suggested by the modeler. After multiple trials with the generator, the spss modeler did not detect the outliers and extremes. Also for some reason the modeler did not discard  the three outliers for age, and income, so we did it by hand.

Moreover, we had to change the value by hand for number children, one was 22 and one was -1. We replaced it with the mean considering we only needed to do it twice, it would not bias the data.

Finally we exported the new data to excel and edited the ages that were filled in to 43.111 from the modeler we changed them to 43. The rest of the data was good to go from the modeler except for a few values that were negative in age we changed them to the mode  value 43.