Data mining project

Science

Objective There are many websites and newspapers giving predictions in this direction, but there is no tool which can give mathematical analysis about the races. For my Data Mining Project I will use a database collected from www. Greyhound-data. Com, then I will use this data in Reprimanded to generate a random race sample and finally I will predict the winner of the race using the same tool. Database The database collected is comprised of 100 examples with 11 dimensions: 1. Place - which represents the national rank 2.

Name - II/II represents the land of standing/land of 3. Land of Birth 4. Land of Standing 5. Year of birth 6. Sex - male or female 7. Sire - father's name 8. Dam - mother's name (the last two dimensions are considered important in ambling) 9. Races - the number of races for 2014 10. Points - how many points each dog heave accumulated in 2014 11. Bag Didst - the average distance of races. All the details are based on 2014 statistics collected from the website up mentioned. On top of these dimensions I manually added three more: 1. Weight - in Keg 2. Owner 3.

Color The last three heave missing data, which make the dataset noisy but I will try to find the best way to recover the missing data. After importing the dataset in Dynamiting from an Excel file, first I analyses the data, then I separated clean data from dirty ATA (no_missing_attributes function). As a result, only 29 items were perfect data, while 71 had missing values (noisy). As we can see in the picture the missing values are highlighted in red. Removing Noise First method used to remove the noise is using the " average" function provided by Reprimanded.

A graphical representation of the design of this method can be seen in the next picture. With this method I replaced " all" missing values with the " average". Generate a Sample Next step is to generate a sample of six items because this is the number of dogs competing in a race. This sample is random generated and the result is: As we can see highlighted in red the national rank is close, which means that the race will be very tight and very hard to predict as well. In the last results I noticed that there is some data that I do not need to use for my final analysis and I decided to remove it.

To do this I used " Remove Useless Attributes" as shown in the next picture: Then the results will look like this: Now is more simple to read data, with only 12 dimensions left. Phase 3 - The Results In this part I will try to predict which of the six dogs will win the race. I will use two ethos, one is the " Aggregate" function and the other is " Attribute Generation". First, I decided to remove some of the attributes as not all of them are actually needed for this operation.

To do this, I used " Select Attribute" function, as shown in the picture below. Six attributes will be enough for the next operation and final operation to find the winner. Next, I will use " Aggregate" operator and I will use the attribute " points" to generate the winner. After I add this operator in the design window, one click is needed to display its functions on the right hand sand. After I clicked on " Edit List", a Indo opened, where I selected the attribute " Points" on the left and the " maximum" function on the left (next picture).

Now we can run the process to see the result: As we can see, based on " Points", the possible winner is the number one dog on the list because he

has the highest number of points. This result can be considered, as the points accumulated are the most important decisional factor when we want to check the " favorite" for a dog race. But because the points are not the only factor to consider, another method has to be found. Next, I will present another solution, which looks even more interesting. It involves weighting the more than one attribute and this is why this method looks better.

I removed " Aggregate" operator and I added another two instead: " Set Role" and " Generate Attribute". I used Set Role attribute to generate a label (picture below - on the right), in this case I choose name. In the next picture is described the Generate Attribute operator. I clicked " Edit List" (number 1) on the right hand side and a new window opened. In this window, new attributes can be generated. At number 2 is defined the new attribute name which is " Winner" in my case, than at number 3 a formula is introduced. The formula weights three attributes " Weight", " Races" and " Distance".

Based on them, Reprimanded will calculate a score for each dog. The results are shown in the next picture In red is highlighted the winner, number one - Austrian Lisa, and in black is the new generated attribute - " Winner", which shows the results for all the competitors. Conclusions This model can be used betting companies like Powdery for example to generate odds for example, but it can be used as well by people who have a passion for gambling. It can be also used to build a website which calculates the winners for future races and attract visitors this way.