Results & Discussion/Conclusion

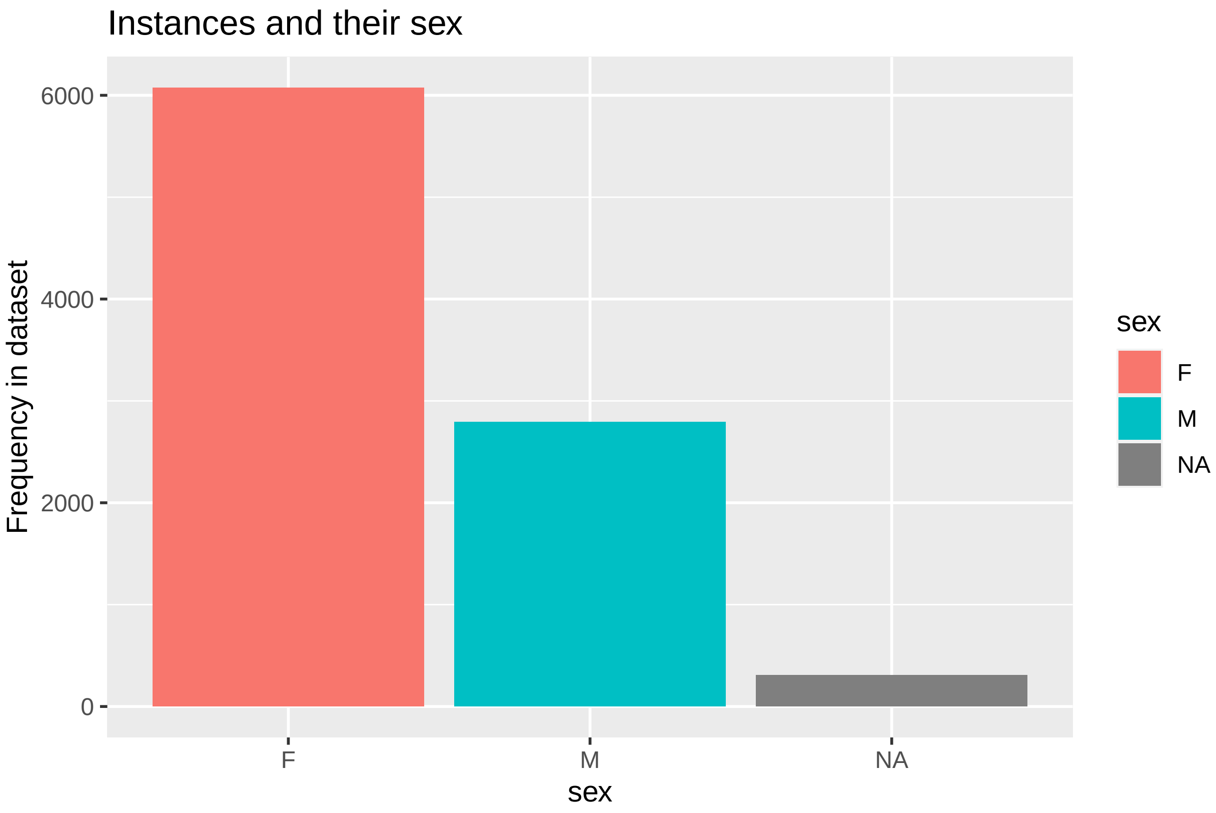
Reindert Visser

2-10-2021

Results

The data is all about people and their thyroid values. All of the hormones have been measured, but also external factors are accounted for. Such as medication, a swollen thyroid or maybe a tumour. These factors affect the hormone cycle of the thyroid, changing a persons mental health. The data has a lot of instances, all of these columns contain a true or false. The patient uses medication or not for example. If the specific hormones (T3 and T4) are measured, the expression value is noted under the _measured column. Of all the 9000 instances, only around the 200 are sick. Since sick is the most valuable column, this will become the class column. The data tells if the patient is either sick or not, so the machine learning algorithm will predict if the patient is sick or not.

Overall the data is of pretty high quality, there are a lot columns that tell you enough information. The data is relatively complete, not extreme amounts of missing values (NAs). All of thyroid hormones are included in the data, and extra hormones like Thyroid stimulating hormone (TSH) are included. So it creates a complete picture of hormones. Most of the patients are healthy, only a small portion of the instances are sick, this might become a problem for the algorithm. Since the algorithm will most likely say "Not sick" and have 80% correct, while the patients with actual thyroid issues might get wrongfully diagnosed.



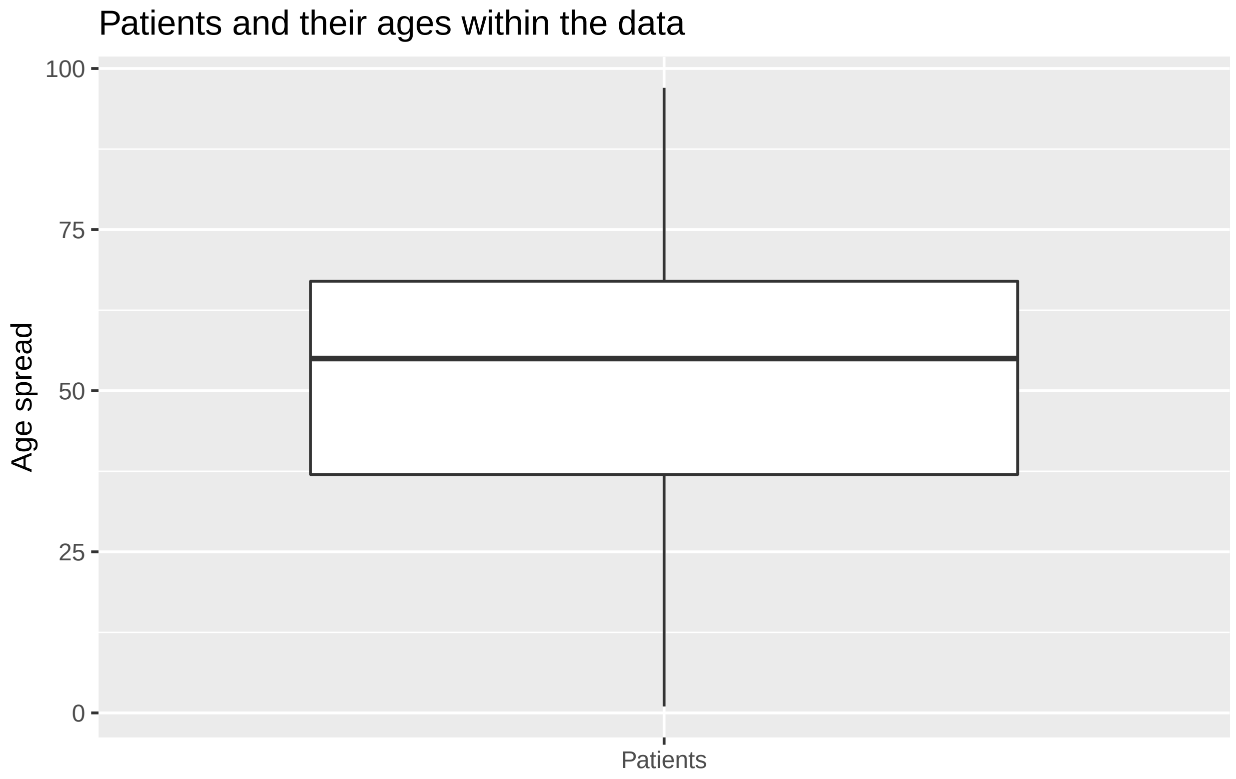
Data source: patients' sex in the dataset

Since sex is a sensitive topic, the NAs have not been filtered out or filled in. This is because the column does not give a lot of information, because the ratio is not near 50/50. Editing the data will result in major integrity loss, so these instances will stay the way they are.

Cleaning data

When it comes to the dataset, which consists of either True or False, cleaning will become an issue. There are no outliers or impossible values, so the data (when orignally downloaded) is mostly clean. There are some measured columns, there are numerical values here, but there are no outliers. So this does not need a lot of work to be done.

The age column has a lot of outliers, there are a lot of people with inhumane ages. For example three people would be in the nine thousand, one would be eight thousand and one would be four hundred years old. This is not possible of course, these values have been modified to only use the first two digits. If someone would say "I'm forty and a half years" and write it down as 40.50. Due to the use of floats it might got mixed up and combined. To not delete the entire instance, the first two digits will be used to fix this problem. Although this is not the most accurate instance. But on a scale of eight thousand instances, this will not be weighted heavily and won't affect the data. The age of the instances makes a jump from 97 to 400 to 8000, the instances above 97 have been trimmed to their first two digits.



Data source: patients' ages from the data

Now that the data has been filtered of the outlying ages, the data looks a lot more believable. Since it is not possible for a person to be 8 thousand years, there must have gone something wrong with the data. This will increase the data's credibility and makes the conclusions of the algorithm move valid. Because there is no need to take these ages into account and split them in the design later on.

Discussion

The data consists mainly of either True or False, so there is little room for outliers and illogical values. This helps the end result in it's integrity. Apart from the age column, which has been modified. Multiple instances have been modified, although this is on a scale on 3 of 9172. This is not significantly for the end result and won't affect the conclusions of the algorithm. Both of the instances are not thyroid patients, so this will fall under the mass of "Not sick". Deleting the instance might result in loss of information, this was the right way to act with these instances. But since this is a human error, it is not certain if this would be the correct for the person. But there is no way to track these individuals down, so there is no other way to either delete or modify the data.

Data ratio

A minor issue with the data, is that the male/female ratio is not 50/50. Females stereotypically tend to go more to the doctor. Since the thyroid involves mental health, an adult man won't go to the doctor unless it's absolutely necessary. This is also represented in the data set. There are a lot more females than males in the data, so hormone expression might be different. But sickness is provable due to hormone expression and external factors, so the male/female ratio won't affect the results of the data set. Since the algorithm does not take the sex of the patient (or instance) into account, since this column gives the least amount of information. The information gain is very little since all instances will most likely be female. This won't score highly in bits per instance; so the information gain will be very low. So the results of the algorithm are valid, even though the sex leans more towards female than male.

Conclusion

The original data consists only of True and False. This has it's benefits, since there is very little room for outliers and typos in the data. On the other side, the data won't have as much value per column. The patient either has it or not. And this goes for a lot of columns, for medication, external factors and hormones measured during the data gathering. So the data won't tell as much as numerical values. Also the visualizing process is harder when there is a limiting factor of True or False values. Most of the data's information is displayed via tables, since graphs aren't an option. This weights the data down in it's value.

Learning pattern

The original data has class labels. Since the instances are patients and Sick is the class column, the data tells if the instance is sick or not. This helps the machine learning section, these labels create a supervised learning environment. This way the algorithm can learn what values the sick patients have and how it differs to the other instances. This has multiple benefits compared to unsupervised, since the criteria will be a lot better and the concept of the algorithm will be better. Supervised learning also prevents the classification errors that unsupervised machine learning has. Since the columns and classes are already readied and human errors have been resolved.

General conclusion

To conclude the research project, the original data was relatively clean. Due to the nature of the data, there was little room for outliers and extreme values. The amount of NAs are also relatively low and all relevant columns for the thyroid have been included in the data. External factors are also included and these columns are a great addition to the credibility of the dataset. So the results of the algorithm are valid, biologically and human errors have been filtered out.

Recommendation

For a possible follow-up machine learning project, the effects of medicine could be taken into account. Of all the sick instances, how could the expression of the hormones be brought into the range of healthy people. This way the dose can be exact for every patient and machine learning can determine what values would be most beneficial for the patient. The same dataset can be used again, only the algorithm becomes way more complex. More maths is involved and the algorithm is tested on more various levels, in more ways than just class label prediction. This algorithm will have a less harder time, since it could be designed to help sick patients. With this research project, an simple algorithm like zeroR will say "Not sick" to all instances and have around 80% correct. This algorithm needs to calculate the hormone up scaling per patient to bring them into more healthy ranges, taking into account all healthy instances.