Introduction to Machine Learning

Reindert Visser

27-9-2021

Exploratory Data Analysis

Research topic

For this project thyroid disease is the main area of research. Problems with the thyroid will result in mental health problems and fatigue. This can be a severe problem and hard to track down. Luckily machine learning can help figure out and calculate patterns. Patterns that otherwise wouldn't be found otherwise.

Which will bring the research question:

Is it possible to predict if a person has thyroid disease by using machine learning, looking at expression values, medicine usage and external factors?

With machine learning, great amounts of data can be loaded up and patterns can be made up from that. When the algorithm encounters new data, these patterns can be identified and based on earlier data; machine learning will become better with more training data. Since the data has a lot of instances, machine learning will be a great tool to use. Since the data has a Sick column, it's either True or False. So this is supervised learning with classification on Sick. Because this is what a patient will look for.

Readying the data

Before the research can start, the data needs to be readied. The data has a lot of information from various topics, such as age, sex but also medicine usage and external factors.

Loading the data

When initially downloaded, the data had an additional column. This was the patient's case, before starting with machine learning; this needs to be removed as soon as possible. The column `Diagnose Letter` was combined with an id of the patient. This was removed using the regular expression `\[[0-9]+\]`, in order to create a true machine learning algorithm.

The labels from the variable `my_labels` were used from the disclosed information in `data/thyroid0387.names`. Except the additional text and explanation.

```
my_data <- read.table("data/thyroid0387.data", sep = ",", header=F, na="?")
my_labels <- read.table("data/labels.txt", sep="\n", header=F)

colnames(my_data) <- as.vector(my_labels[[1]])
(head(dplyr::as_tibble(my_data), n =5))
```

```
## # A tibble: 5 x 30
##     age sex   on_thyroxine query_on_thyroxi~ on_antithyroid_medi~ sick  pregnant
##   <int> <chr> <chr>        <chr>             <chr>                <chr> <chr>
## 1    29 F     f            f                 f                    f     f
## 2    29 F     f            f                 f                    f     f
## 3    41 F     f            f                 f                    f     f
```

```
## 4    36 F     f            f                 f                    f     f
## 5    32 F     f            f                 f                    f     f
## # ... with 23 more variables: thyroid_surgery <chr>, I131_treatment <chr>,
## #   query_hypothyroid <chr>, query_hyperthyroid <chr>, lithium <chr>,
## #   goitre <chr>, tumor <chr>, hypopituitary <chr>, psych <chr>,
## #   TSH_measured <chr>, TSH <dbl>, T3_measured <chr>, T3 <dbl>,
## #   TT4_measured <chr>, TT4 <dbl>, T4U_measured <chr>, T4U <dbl>,
## #   FTI_measured <chr>, FTI <dbl>, TBG_measured <chr>, TBG <dbl>,
## #   referral_source <chr>, Diagnose_letter <chr>
```

The data will be loaded using the `read.table` R build-in function. The data does not have a header and uses a `,` as separator. NA (Not available) is defined with a `?` in the data. Not every instance (patients) have a specific type of thyroid disease, most of these conditions are labelled with `False`. For the hormone measurement columns, these columns form a combination. The hormone name is simply a true or false, while the measured value is the exact amount of hormone expression measured. If there is no hormone expression measured, it will default to NA. This will not form a big issue since the hormone column gives enough information to work with.

For the label file, the data does not have a header either. The separator is a `\n` and there are no NA's. This data is used to overwrite the current column names with the more easy to understand and practical names.

Exploring the data

For the majority of the data, the data is categorical binary data. Which indicates it's either a yes or a no, 1 or 0, True or False. But not all, there are also a few numeric values; which are intervals.

The thyroid plays a big part in the endocrine system, it gives off multiple hormones to regulate the body. There are multiple hormones that the thyroid can give off. The values are tests in order to measure the amount of hormone in the blood. Since most of the hormones transport with blood cells, the thyroid values are testable. If there is a value measured, the default false will become true. With the matching value in the "Measured" column. [1]

- **TSH** (thyroid stimulating hormone): this hormone helps the thyroid release the T3 and T4 hormones. Without TSH, the thyroid would not be able to function at all. If there is too much TSH, the thyroid isn't releasing enough T3 and T4. TSH is released from the hypothalamus and targets the thyroid.
- **T3** (triiodothyronine): is one of the two thyroid hormones, the hormones will influence each other and affect a lot of different processes in the body. Such as metabolism of glucose, the breakdown of cholesterol or increasing the heart rate.
- **T4** (thyroxine): this is the other hormone that will cooperate with the T3 hormone. They have the same function and are involved in the same general processes of the body: Metabolism, growth and increased catecholamine. [2]

All these hormones can be found with tests, in order to prove activity, compare the expression values or prove absence. The test results will be displayed under the measured column in the data, which is continuous data.

- **T3 test**: will measure the amount of T3 in the blood. Since all transport of the thyroid hormones is done via blood cells, this will give an accurate amount of hormones.
- **TT4 test**: will measure the total amount of free T4 and bound to blood T4. Since there are two types of T4, the way that it's used can differ. It can be bound to the blood or roam in the fluid. There for it is also provable by the test.
- **T4U test**: will measure the amount of T4 Uptake, which tests the usage of available T4.
- **TFI test**: will measure the amount of Free T4, Total Free (T4) Index. This can range between 0.7-1.9 nanogram per decilitre of blood.

- **TBG test**: will measure the amount of TBG (thyroxine-binding globulin) enzymes, which is cleaved to produce the T4 hormone. Which means this is an indirect test to measure T4. [3]

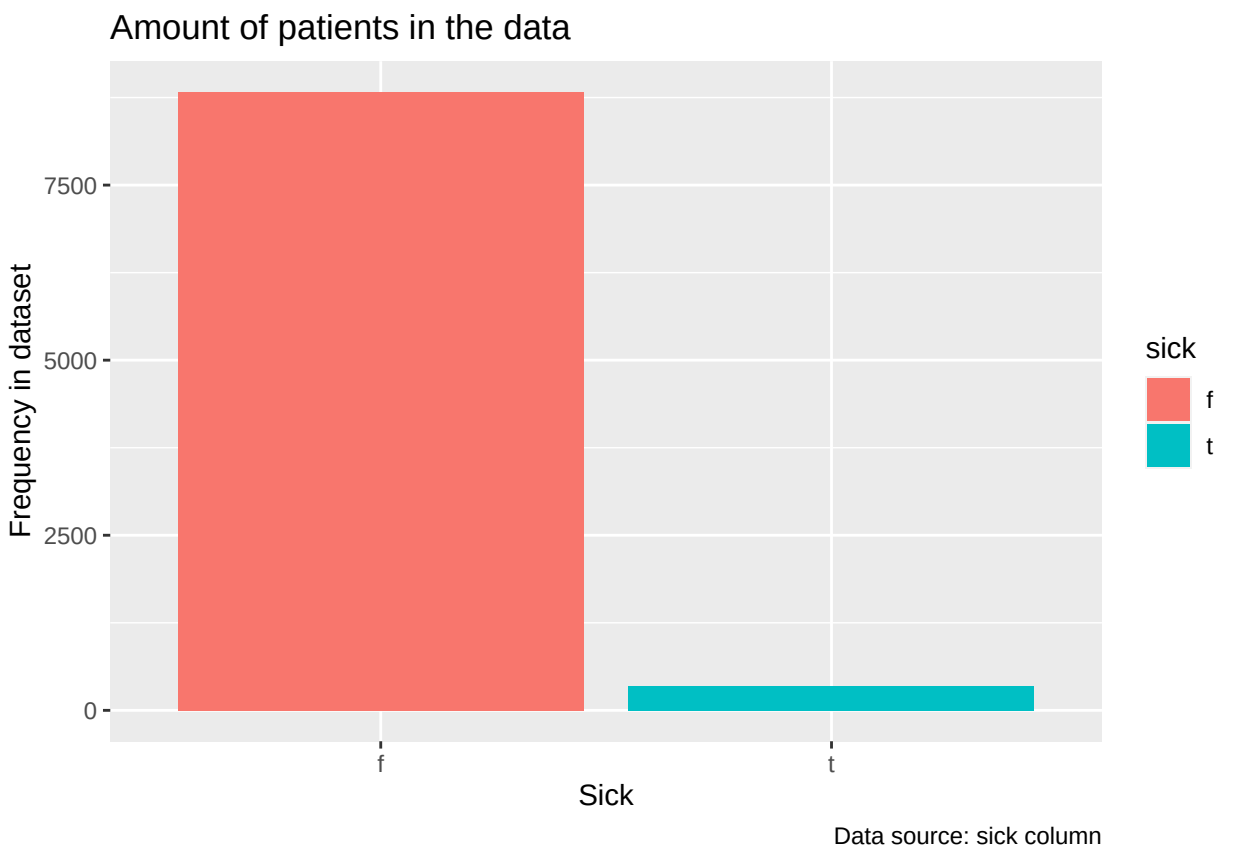
The remaining columns are about external factors. Where the body has either abnormalities or lack of hormones/body resources. These include the following columns:

- **Lithium**: this is an element that circulates in the body and helps different processes. Patients can take extra to stimulate the thyroid and becoming more active. Which will result in true in the data.
- **Goitre**: a major increase in size of the thyroid or local swelling. Which can be diagnosed, if this is the case then the value is true in the data.
- **Hypopituitary**: is a rare case of the pituitary gland does not produce enough hormones. [4]

Visualizing the data

Plotting the data is the fastest way to give a visual view of the data. This can be done by all sorts of ways, since the dataset is mostly categorical; it will be hard to make graphs. Most of the visualization will be done through tables, due to the lack of numeric values.

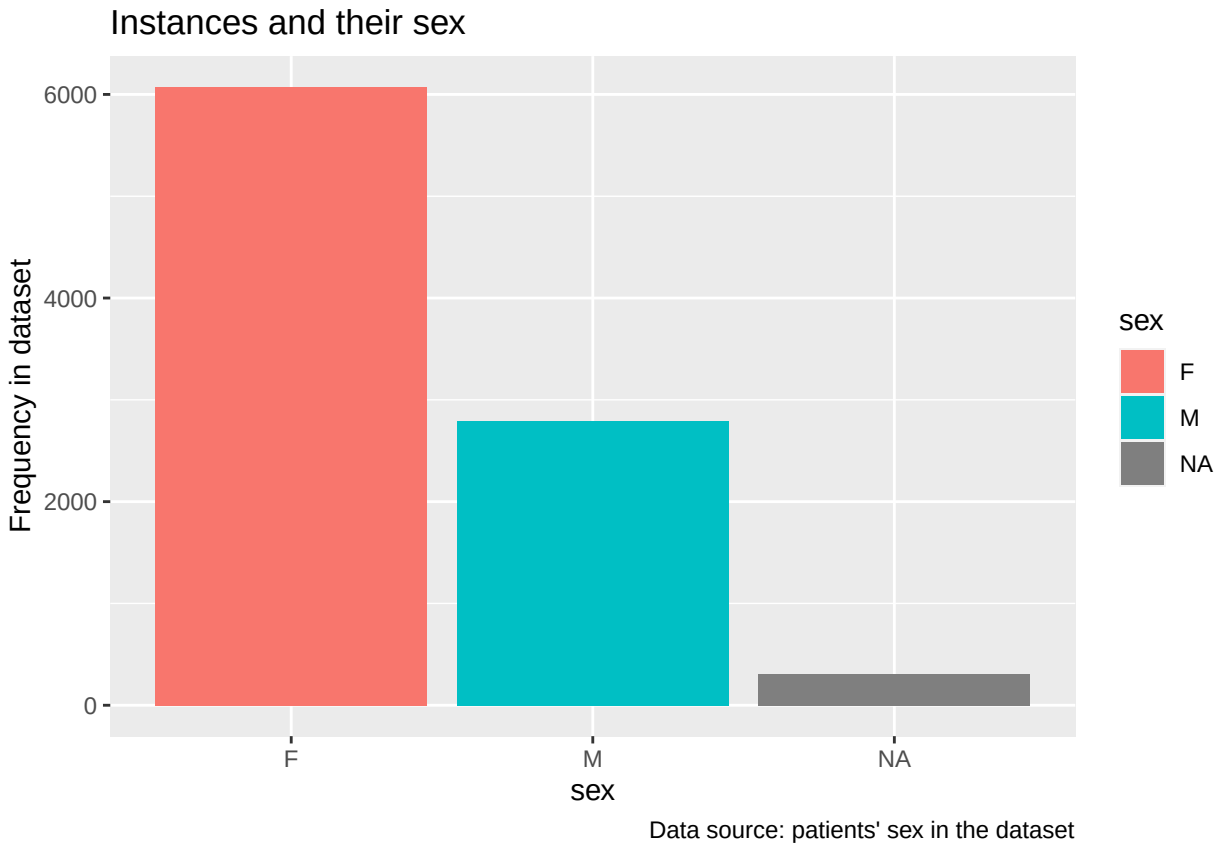
```
ggplot(data = my_data, mapping = aes(x = sick)) +
  geom_bar(aes(fill = sick)) +
  ggtitle("Amount of patients in the data") +
         labs(caption = "Data source: sick column") +
         xlab("Sick") +
         ylab("Frequency in dataset")
```



The great majority of the instances in the data are not patients. Only around 3% of the instances are patients. Which might become tricky to correctly diagnose the real patients according to their thyroid values.

One of the main factors in the data is sex, men tend to go less to the doctor. Since thyroid involves many mental problems, fatigue and eating disorders there will be a difference between sex.

```
ggplot(data = my_data, mapping = aes(x = sex)) +
  geom_bar(aes(fill = sex)) +
  ggtitle("Instances and their sex") +
         xlab("sex") +
         ylab("Frequency in dataset") +
         labs(caption = "Data source: patients' sex in the dataset")
```

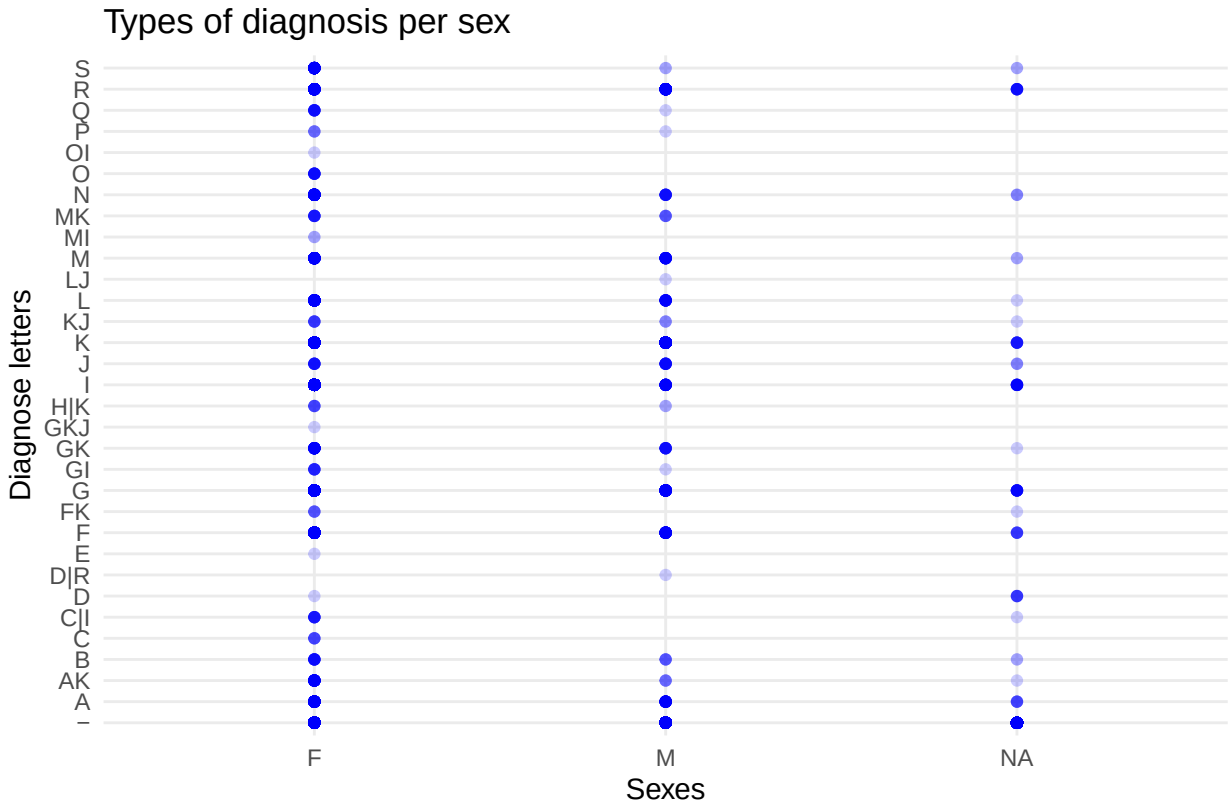


As predicted there are more female instances in the data. There are also people that did not want to specify their sex, these instances are labelled under NA in the data.

Since there are differences between sexes, not only body but also in hormones. The thyroid might act different, this might affect the effects people feel and how they respond to this.

```
ggplot(data = my_data,
  mapping= aes(x = sex, y= Diagnose_letter)) +
  geom_point(col= "blue", alpha = 0.2) +
  geom_smooth(method="loess", se=FALSE) +
  ggtitle("Types of diagnosis per sex") +
  xlab("Sexes") +
  ylab("Diagnose letters") +
  labs(caption = "Data source: sex and diagnose columns") +
  theme_minimal()
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Not all males have the same problems as females, some more frequent then others. NA is added this time to give a general overview of the data.

There are many types of diagnosis, in varying degree of frequency. This will give insight in the distribution of the dataset.

```
pander(table(my_data$Diagnose_letter))
```

Table 1: Table continues below

-	A	AK	B	C	C|I	D	D|R	E	F	FK	G	GI	GK	GKJ
6771	147	46	21	6	12	8	1	1	233	6	359	10	49	1

H|K	I	J	K	KJ	L	LJ	M	MI	MK	N	O	OI	P	Q	R	S
8	346	30	436	11	115	1	111	2	16	110	14	1	5	14	196	85

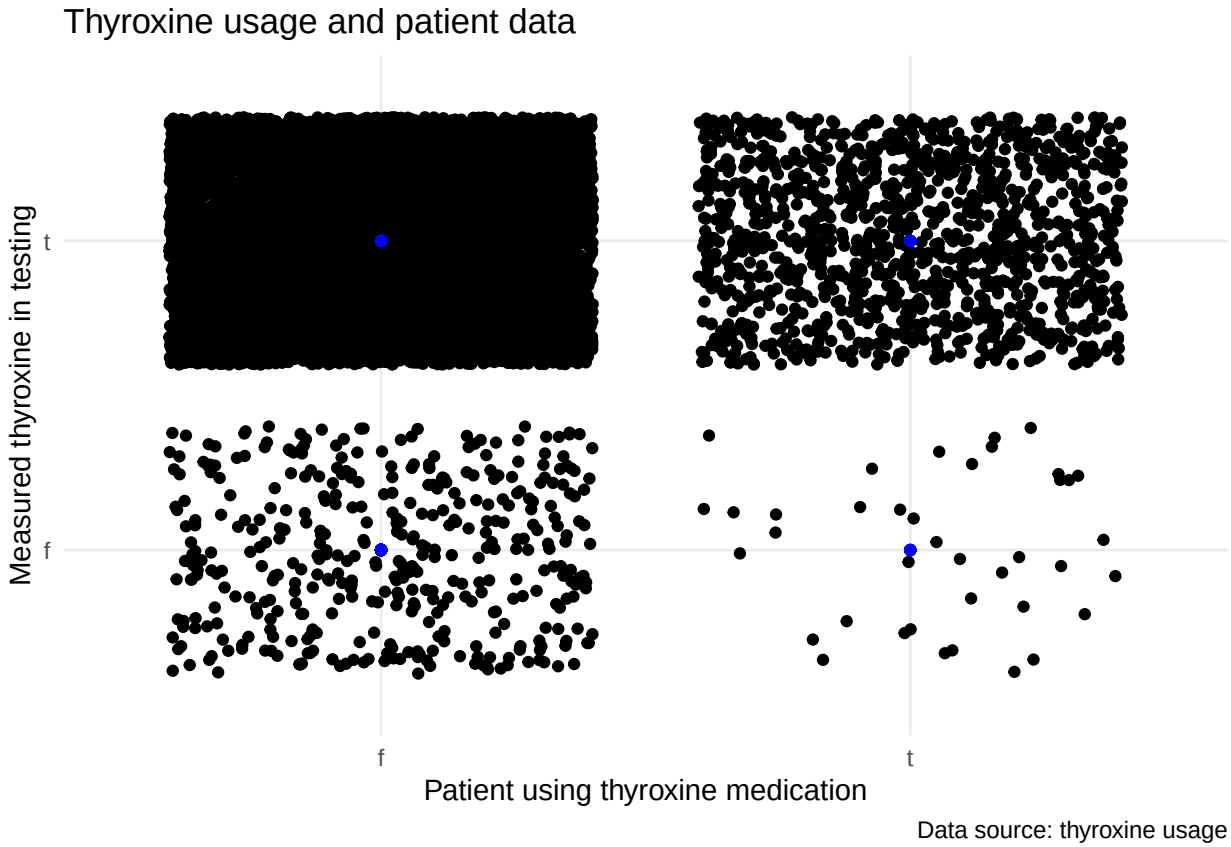
As shown above, not everyone in the dataset is a patient. Some diagnosis are more common than others, a few cases are really rare on a scale of nine thousand instances.

Another common medicine that helps thyroid patients with their problems is thyroxine. This is additional TT4 hormones that will help the body maintain or keep the cycle it is currently in. But how do these effects translate to the patient data set:

```
ggplot(data = my_data,
  mapping= aes(x = on_thyroxine, y= TT4_measured)) +
  geom_jitter(width = 0.4, height = 0.4) +
```

```
  geom_point(col= "blue", alpha = 0.2) +
  geom_smooth(method="loess", se=FALSE) +
  ggtitle("Thyroxine usage and patient data") +
  xlab("Patient using thyroxine medication") +
  ylab("Measured thyroxine in testing") +
  labs(caption = "Data source: thyroxine usage") +
  theme_minimal()
```

```
## `geom_smooth()` using formula 'y ~ x'
```



As seen, the patients that use thyroxine actually translates really well to the data. Due to the scope of the data and the big amounts of instances, it might look insignificant. But medication proves results in our data.

Results

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 originally 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.

References

[1] Boron, W. F., Boulpaep, E. L. (2012). *Medical Physiology Chapter 49, "Synthesis of Thyroid Hormones"* (2nd ed.). Elsevier/Saunders. ISBN 9781437717532.

[2] Basile, L. M. (z.d.). *What are T3, T4, and TSH?* EndocrineWeb. used on 23 September 2021, from https://www.endocrineweb.com/thyroid-what-are-t3-t4-tsh

[3] Shahid, A. H., Singh, M. P., Raj, R. K., Suman, R., Jawaid, D., Alam, M. (2019). *A Study on Label TSH, T3, T4U, TT4, FTI in Hyperthyroidism and Hypothyroidism using Machine Learning Techniques* 2019 International Conference on Communication and Electronics Systems (ICCES). Published. https://doi.org/10.1109/icces45898.2019.9002284

[4] James R Mulinda, Arthur B Chausmer, Francisco Talavera *Hypopituitary Hypopituitarism Causes, Symptoms and Treatment* 2018, January 3rd. EMedicineHealth. *https* : *//www.emedicinehealth.com/hypopituitary/article*$_e$*m.htm*