Biostatistics [SBE304] (Fall 2019)

ProjectBiostatistics [SBE304] (Fall 2019)

Project proposal

Thyroid Disease

Prof. Ayman M. Eldeib **Asem Alaa**

Friday 25th October, 2019

Group num. : 17

Ghada Adel

Gehad Mohammed

Mostafa Tawfiq

Neveen Fathy

1. Background and motivation:

Recently, thyroid diseases are more and more spread worldwide. In Romania, for example, one of eight women suffers from hypothyroidism, hyperthyroidism or thyroid cancer. Various research studies estimate that about 30% of Romanians are diagnosed with endemic goiter. The thyroid gland has a large effect on all organs of the body and it is the secretion of many important hormones of the body, for example the hormone calcitonin, which regulates the level of calcium in the blood and hormone thyroxine which is Responsible for the organization of burning in the body ,therefore the thyroid diseases have a great risk.

For example:

1- heart diseases:

Is irregular heartbeat with0 acceleration or decline leading to muscle weakness in the heart

2- Goiter:

it shows by abnormal swelling in the neck area and affects breathing and swallowing ability naturally

3- Infertility and lack of fertility:

for women affected by the process of ovulation negatively and for men affect the production of sperm.

4- Impact on mental health:

a sense of tension and anxiety and permanent depression

5- problems of adulthood:

the impact of puberty and poor mental growth and irregular PMS

6- Neuropathy:

Peripheral nerves that connect the brain, spinal cord and other organs are damaged

7- Exophthalmos and vision problems:

When neglecting thyroid treatment, exophthalmos, corneal complications and vision problems occur

Thyroid diseases:

1-Hypothyroidism:

It is a common endocrine disorder, in which the thyroid gland does not produce enough thyroid hormones.

This can lead to a number of symptoms, such as fatigue, poor cold tolerance, and weight gain. In children, hypothyroidism leads to delayed growth and intellectual development.

Hypothyroidism occurs in 3% of the population.

Women of pregnancies get hypothyroidism in 0.3-0.5%.

After childbirth, about 5% of women develop postpartum thyroiditis, . It is characterized by a short period of hyperthyroidism followed by a period of hypothyroidism; 20-40% of these cases remain with permanent hypothyroidism.

2-Hyperthyroidism:

It is an overactive thyroid tissue causing excessive production of thyroid hormones.

Hyperthyroidism causes thyrotoxicosis, which is caused by increased thyroid hormones in the blood.

About 5% of patients with myasthenia also suffer from hyperthyroidism

Postpartum hyperthyroidism occurs in about 7% of women during the year after childbirth

The death rate of thyroid storm approached to 100%. But now, with the use of intensive therapy, the death rate from thyroid storm is less than 20%.

Statistics show that about 20 million Americans suffer from thyroid disease, in addition to the millions who have some form of thyroid disease, representing more than 12% of the US population will develop a thyroid condition during their lifetime. Up to 60% of people with thyroid disease are unaware of their condition, and women are five or eight times more likely than men to have thyroid problems.The thyroid gland is the fastest-growing form of cancer in the United States, and the American Cancer Society estimates

that 58,670 new cases of thyroid cancer will be diagnosed in 2017, caused the death of nearly 2,000 people. When thyroid cancer is identified and treated early, most patients can be treated completely. In Australia, statistics show that 1 in 7 Australians have thyroid disorder, mainly due to iodine deficiency in our diets

These statistics help to quick recovery and early diagnosis and make suitable decisions to find better treatment methods.

2. Project objectives:

The goal of this project is highlighting the importance of discovering patterns, relations between various and to make predictions based on volumes of data.

This work is considered a multiclass classification to predict whether a patient is normal (1) or suffers from hyper thyroidism (2) or hy pothyroidism (3).

Additionally,the goal of this study is to find the best classification model in order to make future classification of new patient data more accurately.

3. Data:

Data are unkown patterns that are extracted from an enormous volume of data involving different methods and algorithms which exist at the intersection of fields such as machine learning, statistics and database systems .

The patient data provide a basis for the analysis of risk factors for many diseases (various types of cancer, heart diseases, diabetes, hepatitis etc.).

They are thyroid diseases data set links:

http://archive.ics.uci.edu/ml/datasets/Thyroid+Disease?fbclid=IwAR1oooR-5hq4yorLjDqGBYIOrG75pf9ZxxC0YgHqw6TOaCKTffeBjte9qyY

http://archive.ics.uci.edu/ml/machine-learning-databases/thyroid-disease/?fbclid=IwAR2PYkiEeVRsEmSCc7e_4ymnVitQ3tvZ40_tVCOYmcRpW1Fx3SzZ3tzFL8A

•Feature selection

We need feature normalization because we have 5 feature different ranges. and the model is not able to learn. Because different features do not have similar ranges of values and

hence gradients may end up taking a long time and can oscillate back and forth and take a long time before it can finally find its way to the global/local minimum. To overcome the model learning problem, so we normalize the data. To We make sure that the different features take on similar ranges of values so that gradient descents can converge more quickly it's not good to give all the features to the Machine learning algorithm and let it decide which feature is important because more quickly.

1. Curse of dimensionality — Overfitting

as the dimensionality of the features space increases the number configuration can grow

exponentially and this the number of configurations covered by on observation decreases

2. Occam's Razor

We want our models to be simple and explainable. We lose explain ability when we have a lot of feature.

3. Garbage In Garbage out

Most of the times, we will have many non-informative features. For Example, Name or ID variables. Poor-quality input will produce Poor-Quality output.

So, we do We select only useful features.

Fortunately, Scikit-learn has made it pretty much easy for us to make the feature .selection

There are a lot of ways in which we can think of feature selection, but most feature selection

methods can be divided into three major buckets

1. Filter based: We specify some metric and based on that filter features. An example of such

.a metric could be correlation/chi-square

2. Wrapper-based: Wrapper methods consider the selection of a set of features as a search problem. Example: Recursive Feature Elimination

3. Embedded: Embedded methods use algorithms that have built-in feature selection

.methods. For instance, Lasso and RF have their own feature selection methods.

• Feature selection

we will use feature selection because we may have not necessary feature.

• Data imputation

we don’t need to use Data imputation because we don’t have missing value.

5. Exploratory Data Analysis(EDA):

Is the method of data discovery in order to understand deeper and better and check the quality of data ,if there are errors or spaces and how we can deal with these spaces and test questions from the assumptions that we make on a set of data to look for relationships between variables and imagine this data by drawing Graphical In statistics, exploratory data analysis (EDA) is a method to analyzing data sets to summarize their main characteristics, often with visual methods, Exploratory data analysis, robust statistics, nonparametric statistics, and the development of statistical programming languages facilitated statisticians' work on scientific and engineering problems. Such problems included the fabrication of semiconductors and the understanding of communications networks. An example of a data analysis is the work of a group of amateurs who analyzed many space data collected and found a solar system of four planets by analyzing the properties of light.

There are a number of tools that are useful for EDA.

Typical graphical techniques used in EDA are: Box plot Histogram Multi-vari chart Run chart Pareto chart Scatter plot Stem-and-leaf plot Parallel coordinates Odds ratio Targeted projection pursuit Glyph-based visualization methods

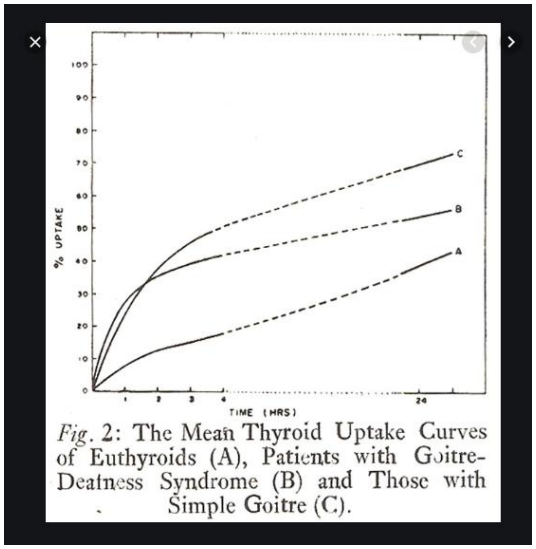
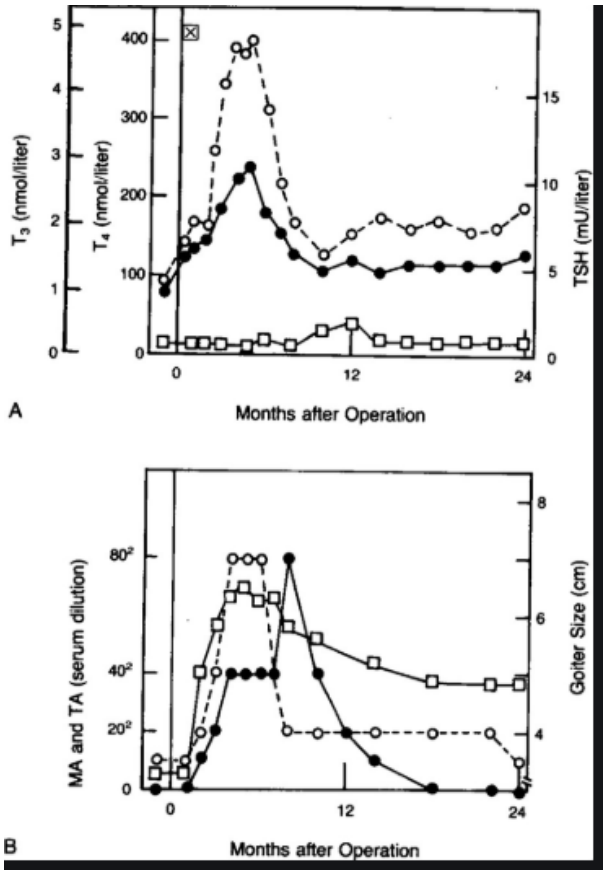
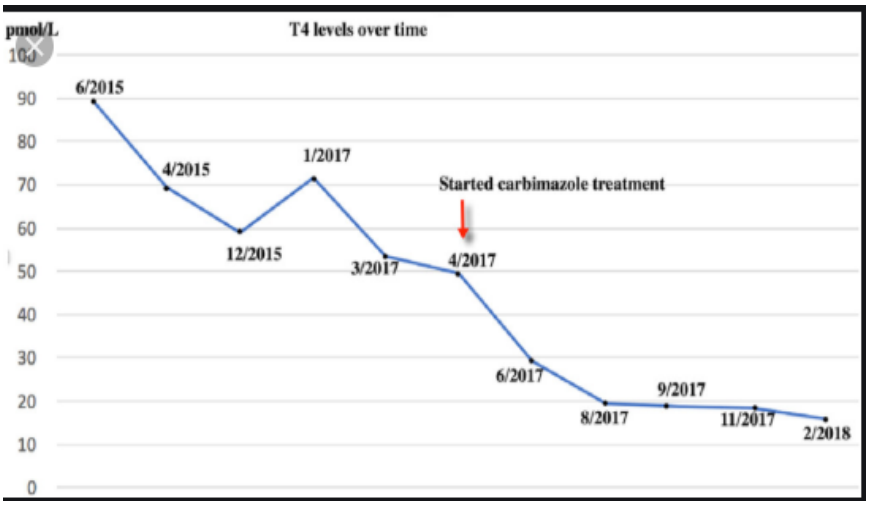
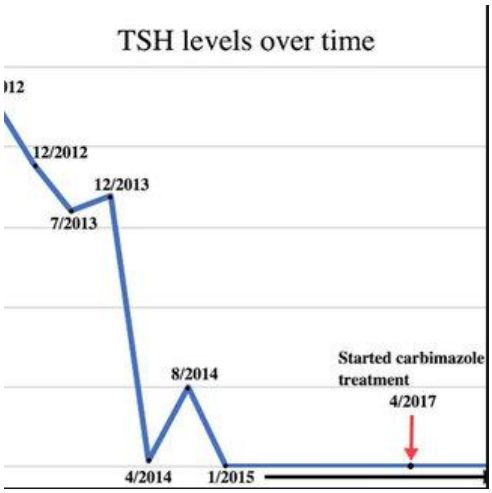


Fig. 2: The Mean Thyroid Uptake Curves of Euthyroids (A), Patients with Goitre-Deafness Syndrome (B) and Those with Simple Goitre (C).







6. Methodology:

We wil use these methods:

• Naive Bayes (NB) Classier (or Gaussian NB Classifier

• Decision Trees

• K-nearest neighbors (KNN) model

Those suitable methods for data sets and features we use .

The method which is with the highest accuracy is Decision Tree ,which is followe by NB, then KNN.

KNN method has a little noise.

7. Project schedule:

Our plan schedule

We are 4 members in the team , each two member will take a separated part ,we illustrate the main 2 scoops that will take our concern and care as observed below with time needed.

What is our data with its issue going to discuss?

It needs to be done in parallel to learning R programming and methods that are used in machine learning (it needs to be in first days) ,(before mid term beginning as a maximum).(11hrs)

1. Analyzing our data about thyroid disease into its 5 features(all information for them) .(2 hrs)
2. re-arranging the sources of data which we need.(30min)
3. understanding the data which we have to deal with (columns refers to what raws refers to what). (20min)
4. classifying the data according to supervised type, unsupervised or another type .(4hrs)
5. searching for the best(needed) algorithm for solving our problem.(1hr)
6. probabilities that we need to calculate for getting the unknowns.(3hrs)

How to use R programing language in machine learning ?

will take 8h in watching : 16 hours-→ 22hours watching with practice .

1. What is machine learning ? methods, jobs,and skills.
2. Decision tree with R .
3. Support vector machine with R-classification and prediction example .
4. Random foster in R-classification and prediction (Definition &steps).
5. Logistic regression with R: categorical response variable at two levels.
6. Multinomial logistic Regression with R: categorical response variable at three levels.
7. Ordinal Logistic Regression or Proportional Odds Logistic Regression with R.
8. Partitioning data into training and validation datasets using R.
9. ROC Curve & Area Under Curve (AUC) with R - Application Example.
10. Forecasting Time Series Data in R | Facebook's Prophet Package 2017 & Tom Brady's Wikipedia data.
11. Neural Networks in R: Example with Categorical Response at Two Levels.
12. Handling Class Imbalance Problem in R: Improving Predictive Model Performance.
13. Linear Discriminant Analysis in R | Example with Classification Model & Bi-Plot interpretation.
14. eXtreme Gradient Boosting XGBoost Algorithm with R - Example in Easy Steps with One-Hot Encoding.
15. Deep Learning with Keras & TensorFlow in R | Multilayer Perceptron for Multiclass Classification.
16. Image Recognition & Classification with Keras in R | TensorFlow for Machine Intelligence by Google.
17. Convolutional Neural Network wirh Keras & TensorFlow in R | Large Scale Image Recognition.
18. Naive Bayes Classification with R | Example with Steps.
19. Fast & Frugal Decision Trees with R | FFTrees | Example using Apple Stock Buying/Selling Decisions.
20. Ridge, Lasso & Elastic Net Regression with R | Boston Housing Data Example, Steps & Interpretation.

21. Competing on Analytics at Kaggle using R | Improving Machine Learning Skills with Real World Data.
22. Self Organizing Maps in R | Kohonen Networks for Unsupervised and Supervised Maps.
23. Deep Neural Networks with TensorFlow & Keras in R | Numeric Response Variable.
24. Feature Selection Using R.
25. Handling Missing Values using R.
26. K-Nearest Neighbour (KNN) with R | Classification and Regression Examples.
27. Working with R in Cloud | RStudio Cloud.

8. The personal websites of the team members:

GHADA:
https://ghada-adel.github.io/ghada.github.io/?fbclid=IwAR1A3SJ0v--n4HKdamZ93zoiXrzmy-LbCbAmP3YJDuiV0BKZZfkmqqL7khI

https://ghada-adel.github.io/Ghada-Divergent/

GEHAD:
https://gehad1999.github.io/gehad.git.io/?fbclid=IwAR0whzUePa2D5PFygfLL-qtFHjOtXIAARn0TKjZsm-0s9Vo6PNg4Hgcf94U

MOSTAFA:
https://mostafa15397.github.io/mostafa.gethub.io/?fbclid=IwAR2ae-ZBDwQZtnZ_JmL9vmttvlklrAvLie7whYpuf2yg5T0ptawjBn24X8g

NEVEEN:
https://neveenfathy1234.wixsite.com/mysite?fbclid=IwAR0UzvCJ6li2MhNRN4-8xE5UAZ90k57w1XW7KTmSV1dbmT45BJxJ24jJ7iI