MVA - Project

Hongping Feng, Pau Rodriguez, Wangyang Ye

June 2018

1 Abstract

Thyroid function diagnosis is an important measure for detecting thyroid related issue. The thyroid gland is one of the most important organ in our body. It secretes thyroid hormones which are responsible for controlling metabolism. The lower and higher level of secretion of thyroid hormone causes hypothyroidism and hyperthyroidism respectively. We used a publicly available thyroid disease data set with raw records and labels indicating whether the patient is healthy or has some thyroid related disease. We analyzed the data set using multivariate analysis to: firstly, visualize and describe the data and, secondly, build a predictive model of thyroid disease. We evaluated the prediction model based on an independent test set and found an generalization error of XX%.

2 Introduction

The control on the digestion system helps organs, like the heart, to work appropriately. In any case, the disease of thyroid must discharge a fitting measure of hormones into the circulatory system. This movement controlled by the pituitary organ, which produces Thyroid Stimulating Hormone (TSH), Stimulates not animated the arrival of both T4 and T3. At the point when thyroid over produces and exorbitantly secretes hormones, this condition is hyperthyroidism (overactive thyroid). In contrast, hypothyroidism (under dynamic thyroid) is the situation when hormones are inadequately created and discharged. Hypothyroidism causes the majority of the body's procedures to back off and may expand heart assault hazard due to uplifted cholesterol levels. A portion of the more regular hormonal issue is connected with the thyroid organ, which is a piece of the endocrine framework. this framework is about gathering of organs that emit chemicals called hormones specifically into the circulation system.

The dataset we used in this project is the *Thyroid Disease Data Set* in the UCI Machine Learning Repository. This directory contains the latest version of an archive of thyroid diagnoses obtained from the Garvan Institute, consisting of 9172 records from 1984 to early 1987. Each record has 29 attributes and 1 response variable (diagnosis to multi classes). The attributes are given in order and separated by commas. Unknown attribute values are indicated by question marks. The variables are:

Attribute Name	**Possible Values**
age	continuous
sex	M, F
on thyroxine	f, t
query on thyroxine	f, t
on antithyroid medication	f, t
sick	f, t
pregnant	f, t
thyroid surgery	f, t
I131 treatment	f, t
sick	f, t
query hypothyroid	f, t
query hyperthyroid	f, t
lithium	f, t
goitre	f, t
tumor	f, t
hypopituitary	f, t
psych	f, t
TSH measured	f, t
TSH	continuous
T3 measured	f, t
T3	continuous
TT4 measured	f, t
TT4	continuous
T4U measured	f, t
T4U	continuous
FTI measured	f, t
FTI	continuous
TBG measured	f, t
TBG	continuous
referral source	WEST, STMW, SVHC, SVI, SVHD, other
class	-, A-T

The diagnosis consists of a string of letters indicating diagnosed conditions. A diagnosis ”-” indicates no condition requiring comment. The conditions are divided into groups where each group corresponds to a class of comments.

Letter	Diagnosis
	hyperthyroid conditions
A	hyperthyroid
B	T3 toxic
C	toxic goitre
D	secondary toxic
	hypothyroid conditions
E	hypothyroid
F	primary hypothyroid
G	compensated hypothyroid
H	secondary hypothyroid
	binding protein
I	increased binding protein
J	decreased binding protein
	general health
K	concurrent non-thyroidal illness
	replacement therapy
L	consistent with replacement therapy
M	underreplaced
N	overreplaced
	antithyroid treatment
O	antithyroid drugs
P	I131 treatment
Q	surgery
	miscellaneous
O	discordant assay results
P	elevated TBG
Q	elevated thyroid hormones

We analyzed this data set with the objective of, first, reducing the dimensionality and finding latent variables and, second, predicting whether the patient has some kind of thyroid diseases using a random forest classifier.

3 Data pre-processing

Our pre-processing procedure consisted in setting continuous variables as numeric, and categorical variables as factors.

3.1 Errors

We detected 3 individuals with an incorrect value for the age variable (a value of 65535 which is clearly an error). We removed those individuals as their proportion in the dataset is very small.

3.2 Missing values

Several of the continuous variables related to measurements on blood tests have missing values. One of them, called TBG, is specially important since 8823 individuals out of 9169 have this value as missing. Looking on information about the test associated with this variable, we see that it is an important measurement but that it is not usually done. We also learn that the range of values for a healthy person is between 10 mg100 mL and 24 mg100 mL. Looking at the summary of this variable within individuals where the value is not missing, we see that the mean in the dataset is 29.87, and that the mean for individuals with the class ”-” (which means they don’t have a condition) is 22.9, whereas for the individuals that have any of the conditions it is 47.72.
At that point we had three different options to proceed: to remove the variable, to impute the individuals that do not have a condition with the mean for the TBG variable of the individuals that don’t have a condition and the rest with an iterative imputation algorithm like Mice, and finally to impute all with the Mice iterative algorithm. We execute the three approaches and compared the resulting first factorial plane in PCA and the contribution of that variable TBG to the first principal component, to see if the the variable TBG carries an important amount of inertia. If this variable carries an important amount of inertia, it cannot be remove from the dataset.

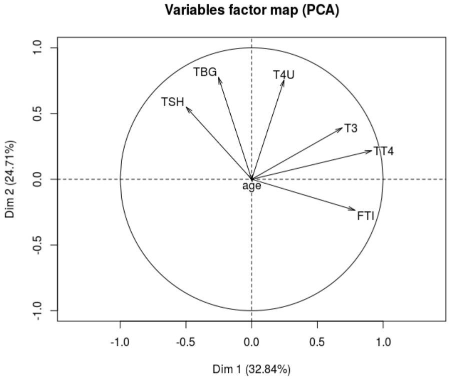
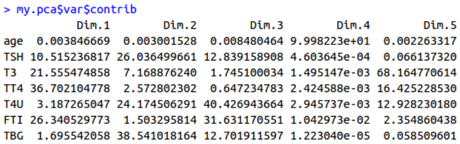


Figure 1: Imputation with Mice



```
> my.pca$var$contrib
          Dim.1        Dim.2        Dim.3        Dim.4        Dim.5
age  0.003846669  0.003001528  0.008480464 9.998223e+01  0.002263317
TSH 10.515236817 26.036499661 12.839158908 4.603645e-04  0.066137320
T3  21.555474858  7.168876240  1.745100034 1.495147e-03 68.164770614
TT4 36.702104778  2.572802302  0.647234783 2.424588e-03 16.425228530
T4U  3.187265047 24.174506291 40.426943664 2.945737e-03 12.928230180
FTI 26.340529773  1.503295814 31.631170551 1.042973e-02  2.354860438
TBG  1.695542058 38.541018164 12.701911597 1.223040e-05  0.058509601
```

Figure 2: Combined imputaton

Figure 3: Different approaches on imputation (all with Mice or combined) , then showing contribution of TBG variable to PC

With the result of the contributions to the first principal components and the amount of inertia explained by both principal components we decide that this variable must be considered in the analysis as it carries an important amount of inertia.

We also compared the two approaches where TBG is totally imputed or just imputed on the individuals that suffer from a condition, to see which of the two imputation strategies affect or disturb the final result or they are in fact similar. We suspected that, since healthy individuals are usually not tested by TBG, using the mean of the TBG variable for healthy individuals on healthy individuals that have this value missing is the most correct thing to do.

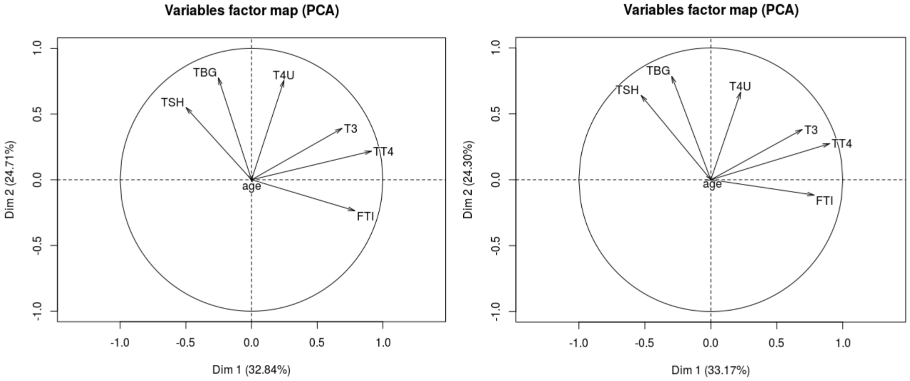


Figure 4: Imputation with Mice

Figure 5: Combined imputaton

Figure 6: First factorial plane with different approaches on imputation (all with Mice or combined)

We observe that the imputation with Mice for all individual is very close to the combined imputation. Therefore, to stay closer to real values we impute all the missing values with Mice.

3.3 Outliers

Including outliers in the training data may invalidate the results, we used the Moutlier function from the chemometrics R package to obtain unbiased results. We Mahalanobis distance to calculate an outlier score and excluded the top 113 outliers. it is NOT acceptable to drop an observation just because it is an outlier. They can be legitimate observations and are sometimes the most interesting ones. It's important to investigate the nature of the outlier before deciding. Therefore, we weighted the individuals inversely to outlying degree of individuals, to diminish its importance. Then we used these weights as metric in the principal component analysis.

4 Data Set and Validation Protocol

The data was retrieved from the “Thyroid Disease Data Set”[**thyroid0387**]. The different data sets and a detailed description of the experiment are available at the UCI machine learning database. After removing the errors, the size of valid data records was slightly reduced to 9169. Then the data was split into a training (70%, N=6418) and a test set (30%, N=2751), randomly assigning subjects to either the training or the test set.

4.1 Principal Component Analysis

We performed a principal component analysis based on the centered and scaled data matrix. We decided to retain 7 principal components based on the Kaiser rule. In this particular dataset, to rotaion 90% of the inertia we need to keep 17 principal components out of 22 variables, which is not suitable. Therefore, we keep 7 principal components that correspond to 46% of the total variance.

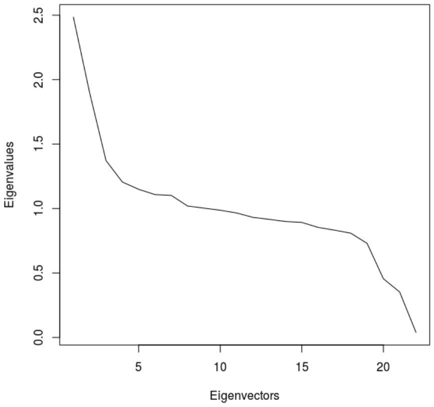


Figure 7: PCA eigenvalues

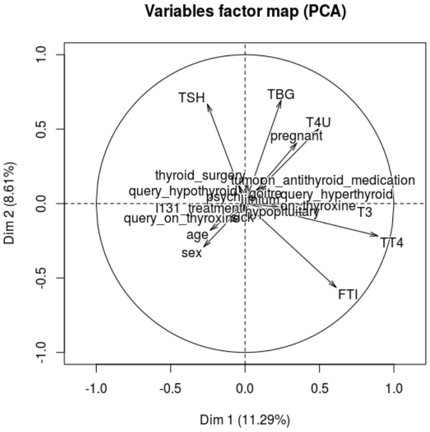


Figure 8: PCA Variable map

Figure 9: PCA plot with row.weights

From the plot of the variables in the first factorial plane, we observe that the most well represented variables are TSH, TBG, T4U, T3, TT4 and FTI. This is validated by looking at the contribution of each variables to the first two principal components, where for the first dimension TT4, T3, FTI and T4U are the most well represented and for the second dimension TBG, TSH and FTI are the most well represented ones. About the contribution to the inertia of first two the principal components, TT4, T3, FTI and T4U are the variables that contribute the most to first component, and TBG, TSH and FTI to the second. A more important result is the most well represented modalities in the first plane. For the target with all its original modalities, the I, A, C, S and M are the most well represented modalities. For the target with 7 modalities (the previous ones grouped in 7 categories) the most well represented modalities are binding, hyperthyroid concurrent_i and negative ("-"). For 4 modality target, hypothyroid is the worst represented modality in the first factorial plane. We finally show a plot in R^p of the individuals with a color indicating it's modality within the target variable. We observe that the data is separable to some degree in the 2 and 4 modalities target scenario.

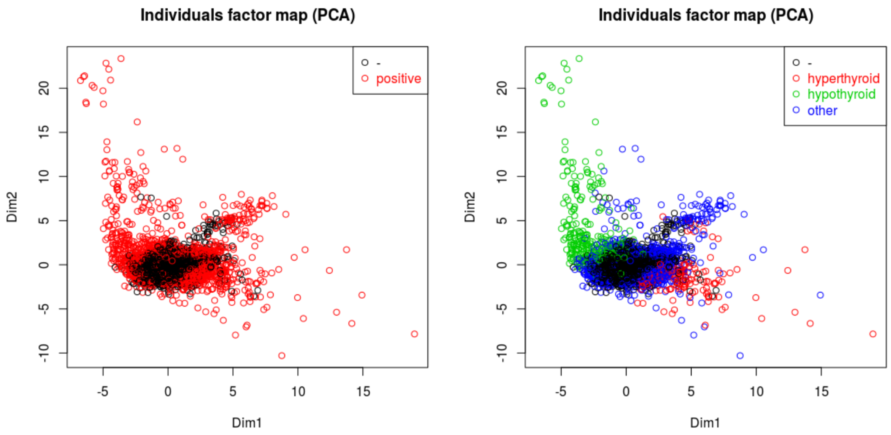


Figure 10: Two modality target

Figure 11: Four modality target

Figure 12: PCA plot with modalities of the target

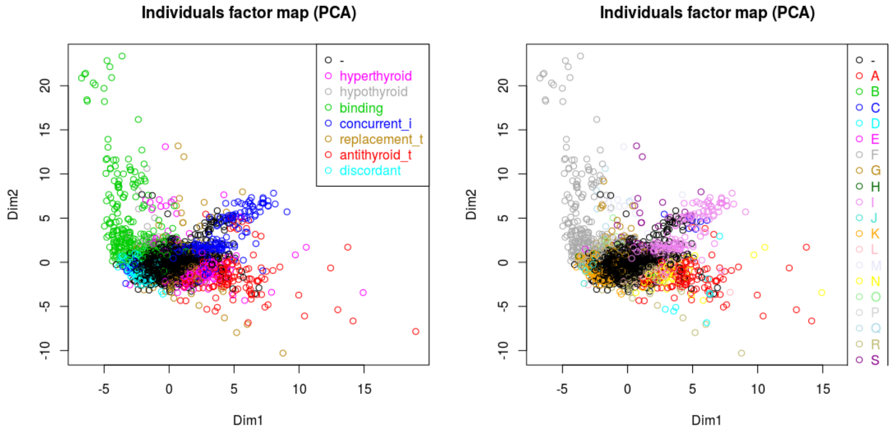


Figure 13: 7 modality target

Figure 14: All modality target

Figure 15: PCA plot with modalities of the target

We perform the Varimax rotation and we plot the rotated variables on figure 16. The plot of the rotated variables projected on the first factorial plane is not well represented nor easier to understand. This is due to the fact the percentage of inertia that the first 2 dimension retain is low (less than 20%). Therefore, as the variables are not well represented in this first factorial plane after the rotation, we determine this result is not to be taken into account.

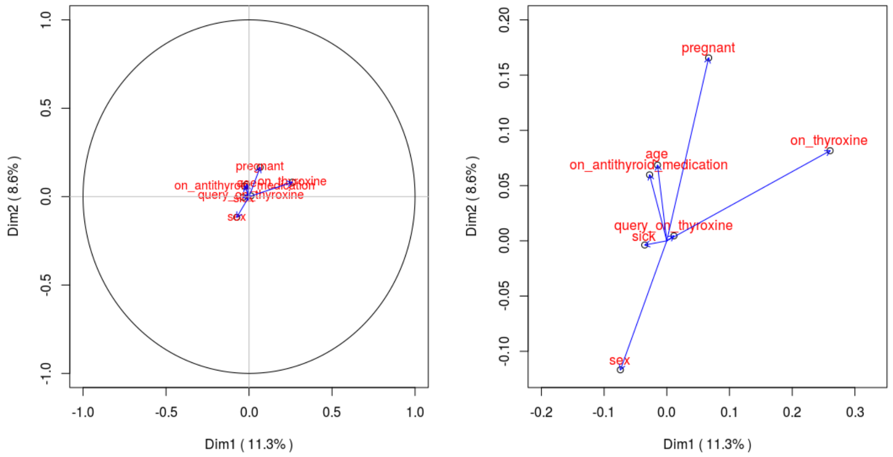


Figure 16: Varimax rotation

4.2 Interpretation of Latent Concepts

We observe from the PCA that the most important variables in the first factorial plane, TSH, TBG and T4U on one hand, and T4, TT4 and FTI on the other hand, seem to be related to 2 different important latent concepts (expressed by the first two principal components). We also observe in the R^p plot that for 2 and 4 modalities of the target, that 2 of the modalities seem to be separable without transformation which suggests at least 2 latent concepts are represented by that structure. We still can't link the variables or the latent concepts to the most separable modalities of the target. Finally, the low inertia explained by the significant components can be an expression of fact that the data comes from different groups or clusters.

5 Clustering

Based on the 7 significant components, we performed a hierarchical cluster analysis with consolidation. We first transformed the new features into an euclidean distance matrix and then used the Ward aggregation method to perform a hierarchical cluster analysis using the hclust function with method argument ward.D2. By observing the jump in height of each consolidation (fig. 18), we determine that the highest jump is in the consolidation number 9 from the right. We conclude that there are 10 clusters in the data. We further consolidate the partition by using the centroids of the 10 clusters as starting point for the k-means algorithm (fig. 19).

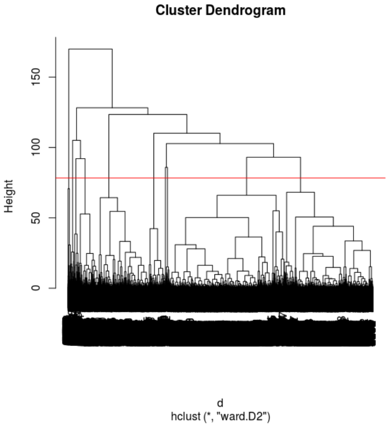


Figure 17: Hierarchical Clustering dendrogram

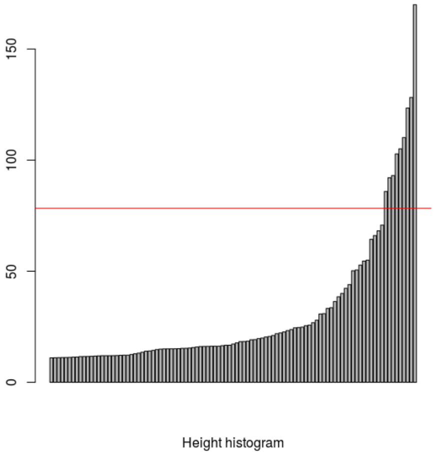


Figure 18: Hierarchical Clustering Heigth Histogram

5.1 Interpretation of the clusters

Using the catdes function, we can do the profiling of the clusters, in order to characterize the individuals they contain by their predictors and the modalities of the target variable. The following table summarizes very briefly the most relevant predictors and modalities of the target variable for the 10 clusters found previously.

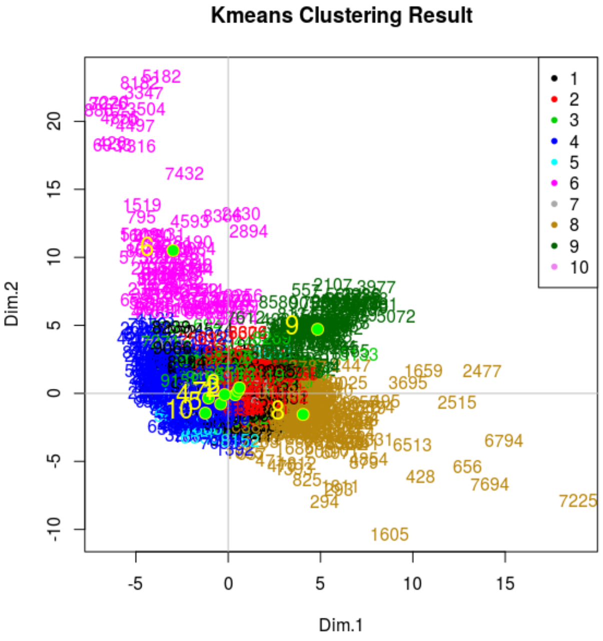


Figure 19: Cluster consolidation

Cluster	Predictors	Modalities (4 classes)	Modalities (7 classes)	Modalities (all classes)

1	on_thyrozine
	query_hypothyroid
	I131_treatment
		other(21%)
		hypothyroid(5%)
			replacement_t
			hypothyroid
				M,L, N,P

2	T4U T3 TT4	-(81%) other(13%)	binding - discordant	I,-, S, R,A
3	query_hyperthyroid on_antithyroid_medication T3	-(77%) hyperthyroid(4%)	antythyroid_t -	O,B
4	sex age sick	-(75%) hypothyroid(10%)	concurrent_t hypothyroid -	K,G,J,F,-,H
5	query_on_thyroxine sex		replacement_t -	L
6	TSH TBG FTI TT4	hypothyroid(76%)	hypothyroid antithyorid-T replacement_t	F,M,Q
7	psych lithium	-(88%)	-	-,R

8	FTI
	TT4
	T3
		hyperthyroid(49%)
		other(40%)
			hyperthiryoid
			replacemen_t
			discordant
				A,N,R,D,L

9	pregnant
	T4U
	TBG
		other(68%)
		-(22%)
		hyperthyroid(8%)
			binding
			hyperthyroid
				I,C

10	hypopituitary query_on_thyroxine	other(100%)	 -	L

Table 1: Profiling of the 10 clusters

As we can observe, the profiling of the clustering shows clearly that there is structure of the data well related to the classes (modalities of the target variable). The most important details are:

- The cluster 6 is in the majority populated with individuals of the class hypothyroid. The most relevant predictors of this cluster are TSH, TBG, FTI and TT4
- The cluster 8 is populated by individuals of the class hyperthyroid in an important proportion. Its most relevant predictors are FTI, TT4 and T3
- Individuals of the class other are prevalent in the clusters 10, 9, 8 and appear in the clusters 1 and 2. The predictors significant in the 10, 9 and 8 clusters are hypopituitary, query_on_thyroxine, pregnant, T4U, TBG, TT4, FTI and T3

- Individuals of the class negative ("-"), are prevalent in the clusters 7, 2, 3, 4 in which the most significant predictors are psych, lithium, sex, age, sick, query_hyperthyroid, on_antihyroid_medication, T3, T4U and TT4.

This results suggests that there is some relation between the classes and several variables that could have good results in predicting those classes(also called modalities of the target). This analysis also suggests that there is more information available to extract or separate clusters 6 and 8 (which are highly related to hypothyroid and hyperthyroid modalities respectively) than the others, and thus, it should be possible and even relatively easier to predict those two modalities than the other ones. Furthermore, given this results, it seems reasonable to limit our prediction study to a 4 classes problem instead of using 7 classes or the whole available classes.

6 Test and training data set

The data set is big enough to allow for a split of training and test data from the original data set. It is even possible and recommended that the training set is used for model selection by cross validation. The size of the data set suggests that the cross validation should be made using a fold not very small (for example in the hundreds), otherwise the resulting cross validation procedure could take too much time. The training and test data sets are drawn randomly from the preprocessed data set in a way to ensure that test set contains 30% of the total data set.

7 Prediction

Recalling the models that are explained during the course and the nature of the undertaken problem, we have 2 choices which are CART and Random Forest. They both support numerical and categorical variables and are suitable for multiclass classification. We choose one of them based on the mean validation error that comes from 10-fold cross-validation.

We obtained the first tree using Recursive Partitioning and Regression Trees function (*rpart* in R) with training data. The optimal tree was obtained by using 10-fold cross validation. To compute the cutoff value, we first extract a cutoff value that generates trees with the minimum cross-validation error. Next we add a standard deviation to the obtained value of the cross validation error, then all the cutoff values that have an equal or lower cross-validation error are candidates. From all of those, we select the cutoff value (complexity parameter) that generates a smaller decision tree. Then we pruned the tree using this complexity parameter. The pruned tree will be used later for model selection.

For Random Forest, there is a main parameter to train which is the number of trees. To do so, we have tried to train the model with different number of trees and observe the OOB (out of bag) error. Since OOB is a fair estimation of training error, based on that, we say that the optimal number of trees is 1000.

	Number of Trees	OOB
1	10	0.1004
2	16	0.0998
3	25	0.0855
4	40	0.0817
5	63	0.0826
6	100	0.0798
7	158	0.0759
8	251	0.0750
9	398	0.0786
10	631	0.0755
11	1000	0.0747
12	1585	0.0762

Table 2: OOB for different number of trees

After we got the best parameter and setting for both models, we apply 10-fold cross-validation to choose our final model.

	TR.error	VA.error
1	0.057	0.088
2	0.063	0.095
3	0.066	0.082
4	0.065	0.083
5	0.057	0.087
6	0.062	0.089
7	0.079	0.108
8	0.059	0.077
9	0.062	0.069
10	0.064	0.072
Average	0.063	0.085

Table 3: 10-fold cross-validation for pruned decision tree

	TR.error	VA.error
1	0.069	0.072
2	0.067	0.075
3	0.064	0.070
4	0.067	0.075
5	0.064	0.085
6	0.067	0.056
7	0.068	0.069
8	0.068	0.064
9	0.068	0.043
10	0.069	0.064
Average	0.067	0.0673

Table 4: 10-fold cross-validation for Random Forest

Since Random Forest results in a smaller average validation error, we chose it as our final model. We use it to fit the whole training data and to perform a prediction over the test set. As a result, the accuracy on the test set is **93.2%** and thus the generalization error is **6.8%**. The macro-averaged precision is **86.58%** and the macro-averaged recall is **89.28%**. The measures of goodness seem good, however, we have to take into account that this is a disease related classification problem. Thus, the precision and recall would be a crucial criteria for the goodness of a model and they have to be improved as much as possible.

	-	hyperthyroid	hypothyroid	other
-	2126	8	17	124
hyperthyroid	10	58	0	11
hypothyroid	0	0	192	7
other	20	5	6	468

Table 5: Confusion matrix of prediction over test set

8 Conclusion

Visualization of the data done in PCA showed that there exists some separability between the different classes of the individuals. Unsupervised analysis performed by PCA and clustering pointed out what are the variables most significant in relation with the latent concepts existing in the data and at the same time which are the classes (or modalities of the target variable) most well expressed by the data. Those hints suggest that prediction should be successful, and it fact, well performant prediction models have been trained sucessfully. However, as the prediction problem is aimed at detecting health conditions, the recall must be maximized. So we conclude that this prediction model must be further improved.

The fact that the decision tree performed better than the Random Forest, suggests that the target modalities we want to predict are not related to all of the predictor variables but a reduced subset. A refinement on the predictor variables could help improve the recall performance. To assure that our prediction model generalizes, we have avoided as much as possible to influence it with information extracted from the unsupervised analysis which is using all the data set. Furthermore, if we apply some refinement or transformation of the predictor variables, the mechanism used must not carry any information of the test data set. For example, if we decide to use the significant principal components to do prediction, those must be derived only from the train data set.