MARMARA UNIVERSITY FACULTY OF ENGINEERING 2020-2021 SPRING SEMESTER

CSE4062

Introduction to Data Science and Analytics

Group 6

Explore your data Part 2

Identification of radiation-induced papillary thyroid cancer after Chernobyl by using machine learning

150116061 Ertuğrul Sağdıç, Computer Engineering - ertugrulsagdic98@gmail.com
150816004 Defne ÇIĞ, Bioengineering - defnecig@marun.edu.tr
150816019 Kasım Anlatır, Bioengineering - kasimanlatirr@gmail.com
150216031 Cem Telliağaoğlu, Environmental Engineering - cem.telliagaoglu@gmail.com
199520021 Lukas Eckerle, Industrial Engineering - lukas.eckerle@gmail.com

EXPLORING THE DATA PART II

The gene expressions of 54,675 genes have been investigated to select the best attributes between exposed group and control group, and not-exposed group and control group. The way the genes investigated is using the Random Forest algorithm [1] with checking the entropy information of the genes [2] as a decision tree classifier [3]. This gives the genes genes that are differentially expressed.

First of all, the data set normalized between 0, 1. Then, the dataset preprocessed. The labels of the instances differentiated as diseased and healthy for exposed group and control group, and not-exposed group and control group. Then, the dataset splitted as in two sets for each investigation group. 70% of the data splitted for the train set, 30% of the data splitted for the test set.

The decision tree algorithm checks the expression levels of the genes and constructs trees by checking the entropy values. Then, the trees are added to random forest with the accuracy level higher than 90% of the accuracy level. Then, the samples are classified based on the trees in the random forest.

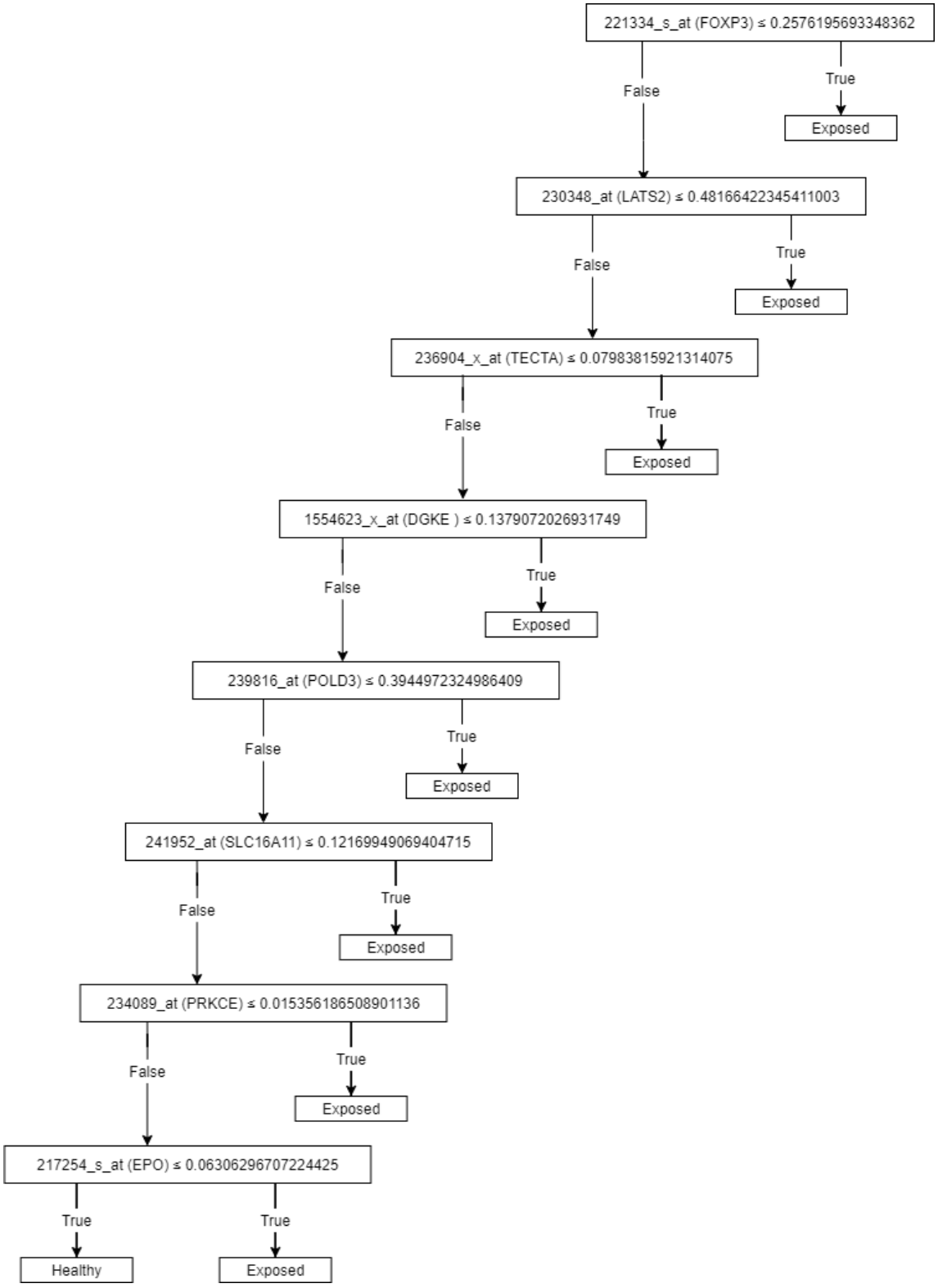

Figure 1: *Decision tree of the best attributes selected for exposed/control group.*

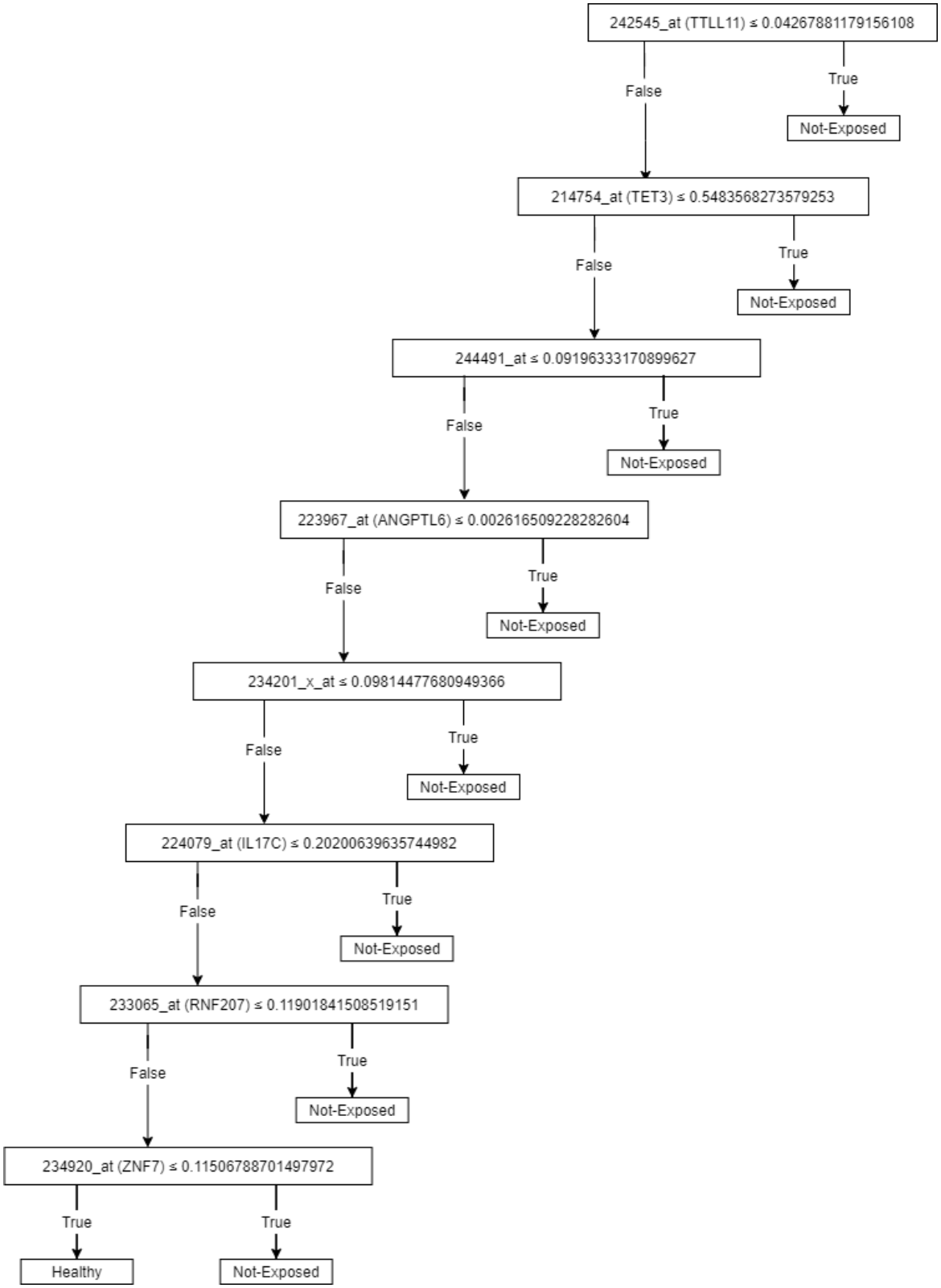

Figure 3: *Decision tree of the best attributes selected for the not-exposed/control group.*

A	Gene ID	Gene Symbols
1	221334_s_at	FOXP3
2	230348_at	LATS2
3	236904_x_at	TECTA
4	1554623_x_at	DGKE
5	239816_at	POLD3
6	241952_at	SLC16A11
7	234089_at	PRKCE
8	217254_s_at	EPO

B	Gene ID	Gene Symbols
1	242545_at	TTLL11
2	214754_at	TET3
3	244491_at	-
4	223967_at	ANGPTL6
5	234201_x_at	-
6	224079_at	IL17C
7	233065_at	RNF207
8	234920_at	ZNF7

Table 1: *A: Differentially expressed genes for exposed/control group. B: Differentially expressed genes for the not-exposed/control group.*

Differentially expressed genes found by the algorithm mentioned above are searched for their relevance to cancer by reviewing the literature and using servers such as Human Protein Atlas and GeneCard. Some of the genes have been shown to play a role in different types of cancer. The protein encoded by FOXP3 is a member of the forkhead/winged-helix family of transcriptional regulators [4]. Studies have shown that the altered expression of FOXP3 is found in autoimmune diseases, benign tumors and carcinomas [5]. SLC16A11, a member of the SLC16A family, is a protein coding gene [6]. Studies have shown that SLC16A family members play important roles in tumor formation and tumor progression [7]. Also, this gene is a prognostic marker in renal cancer, pancreatic cancer and liver cancer [8]. PRKCE, a member of the protein kinase C family, is a protein coding gene [9]. This gene is known to be associated with Ischemia and Anxiety. In addition, it is a prognostic marker for renal cancer, endometrial cancer and lung cancer [10].

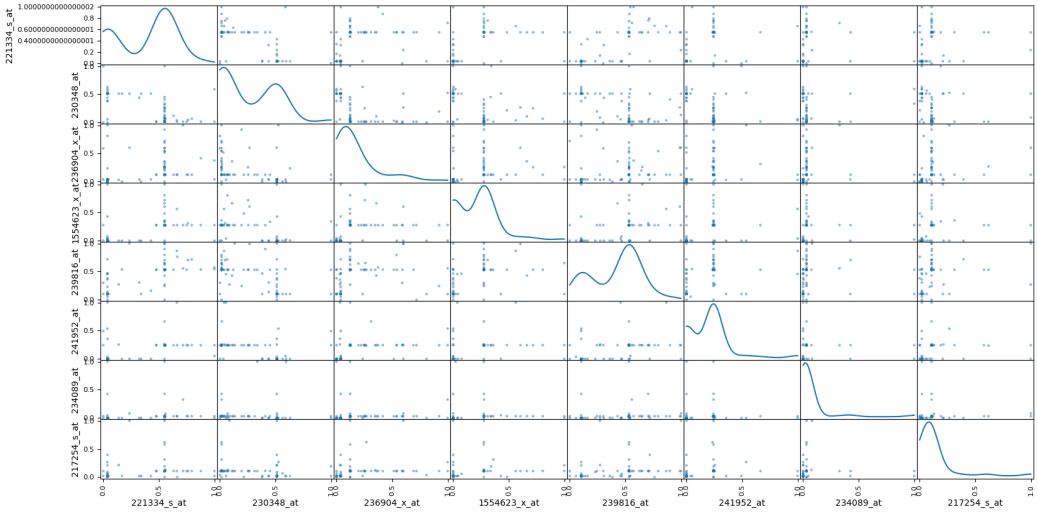

Figure 2: *Scatter matrix plotting of the genes for the exposed/control group.*

The relationships of the genes plotted shown in figure 2. X-axis and Y-axis shows the particular genes that have been found as the best attributes. Situations where the graph forms a positively inclined diagonal are positive correlations, and situations where it forms a negatively inclined diagonal are negative correlations [11]. However, in our case, the presence of correlation was not observed.

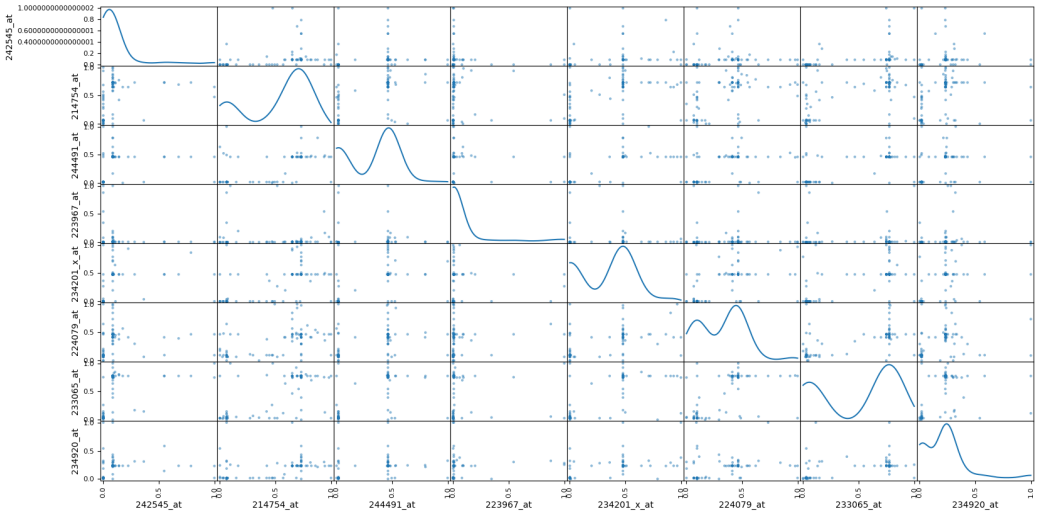

Figure 4: *Scatter matrix plotting of the genes for the not-exposed/control group.*

REFERENCES

[1] Hsueh, H. M., Zhou, D. W., & Tsai, C. A. (2013). Random forests-based differential analysis of gene sets for gene expression data. Gene, 518(1), 179-186.

[2] Butte, A. J., & Kohane, I. S. (1999). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In Biocomputing 2000 (pp. 418-429).

[3] Polaka, I., Tom, I., & Borisov, A. (2010). Decision tree classifiers in bioinformatics. Applied Computer Systems, 42(1), 118-123.

[4] FOXP3, The GeneCards. Retrieved from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=FOXP3&keywords=FOXP3

[5] Szylberg, Ł., Karbownik, D., & Marszałek, A. (2016). The role of FOXP3 in human cancers. Anticancer research, 36(8), 3789-3794.

[6] SLC16A11, The GeneCards. Retrieved from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=SLC16A11&keywords=SLC16A11

[7] Yu, S., Wu, Y., Li, C., Qu, Z., Lou, G., Guo, X., ... & Tai, S. (2020). Comprehensive analysis of the SLC16A gene family in pancreatic cancer via integrated bioinformatics. Scientific reports, 10(1), 1-12.

[8] SLC16A, The Human Protein Atlas. Retrieved from: https://www.proteinatlas.org/ENSG00000174326-SLC16A11/pathology

[9] PRKCE, The GeneCards. Retrieved from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=PRKCE&keywords=PRKCE

[10] PRKCE, The Human Protein Atlas. Retrieved from: https://www.proteinatlas.org/ENSG00000171132-PRKCE/pathology

[11] Swinscow, T. D. V., & Campbell, M. J. (2002). *Statistics at square one* (pp. 111-25). London: Bmj.