MARMARA UNIVERSITY FACULTY OF ENGINEERING 2020-2021 SPRING SEMESTER

CSE4062

Introduction to Data Science and Analytics

Group 1

Explore your data

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

The data set contains gene expressions of 54,675 genes which are the attributes. The type of the attributes are numeric attributes. The gene expressions are going to be used to explore if the patient has thyroid cancer or not.

3 different groups which are the exposed, not-exposed, and the control. Exposed group is the group that has thyroid cancer patients which are affected by the radiation. There are 33 numbers of instances in this group. Not-exposed group is the group that has thyroid cancer patients which are not affected by the radiation. There are 32 numbers of instances in this group. Control group is the group that has healthy patients.There are 51 number of instances in this group.

We investigated the minimum, maximum, average, standard deviation, and entropy of the attributes for each group's gene expressions. Since there are a very long number of attributes, we investigated the first 10 attributes for each group shown in Table 1, Table 2, Table 3. Also, visualization of these attributes with plotting may provide better understanding of the data shown in Figures 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12.

Exposed Group

Gene ID	1007_s_at	1053_at	117_at	121_at	1255_g_at	1294_at	1316_at	1320_at	1405_i_at	1431_at
Min expression	9.707504	5.839823	2.637636	10.59703	2.249185	5.522525	2.698692	2.925209	4.269485	2.249185
Max expression	11.59505	7.226503	5.248017	11.81842	2.249423	7.423284	4.154933	5.284941	11.32402	2.497929
Avg	10.49217	6.704626	3.050141	11.26153	2.249192	6.594616	2.966887	4.031642	7.334552	2.285518
Std.Dev.	0.376622	0.34069	0.5396	0.349783	4.15E-05	0.44691	0.355009	0.551517	1.789697	0.072161
Entropy	3.496508	3.496508	3.270607	3.496508	0.135794	3.496508	3.454499	3.496508	3.496508	1.435813

Table 1. *Attributes (Min, max, average, standard deviation, and entropy) of the Particular Exposed Genes*

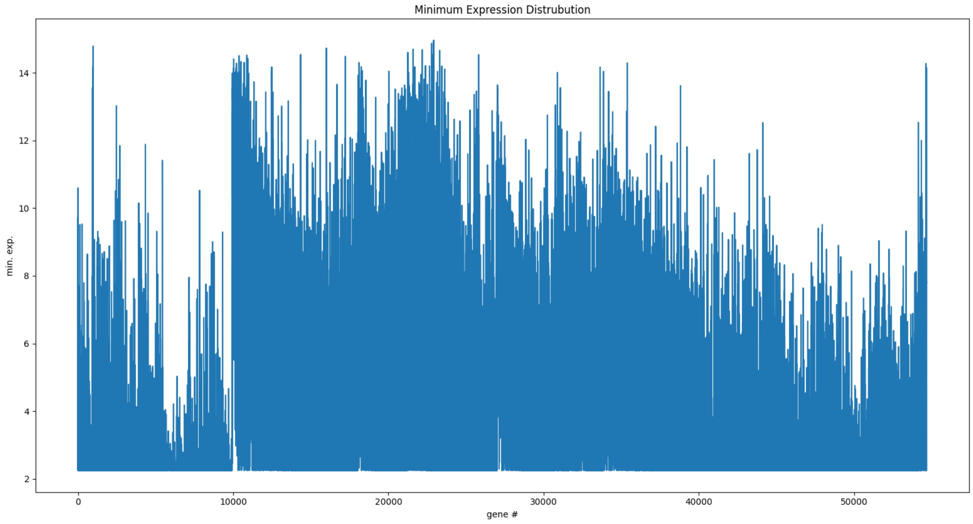

Figure 1. *Minimum expression distribution for the exposed genes.*

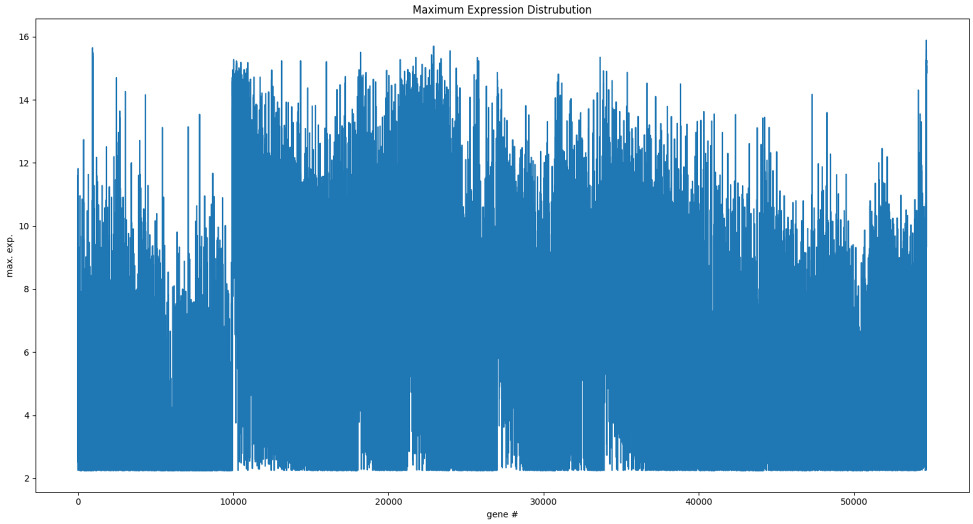

Figure 2. *Maximum expression distribution for the exposed genes.*

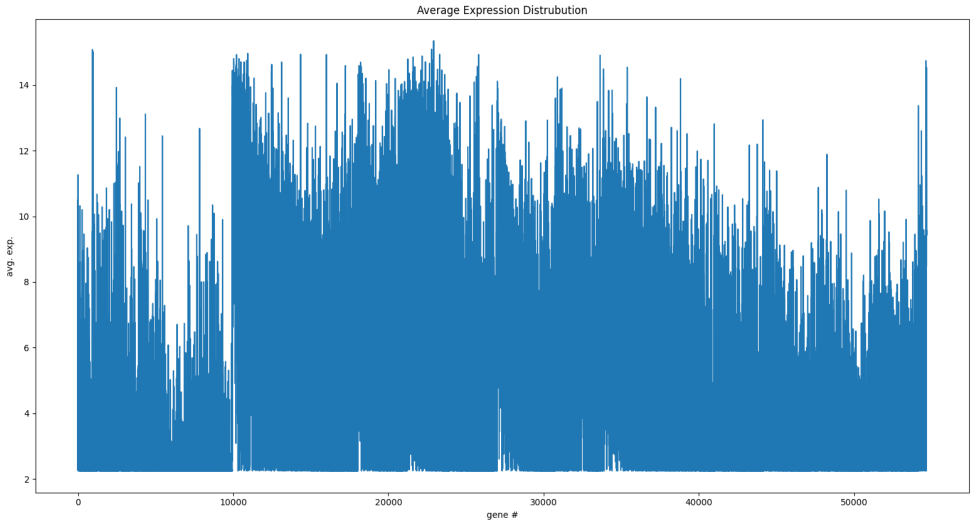

Figure 3. *Average expression distribution for the exposed genes.*

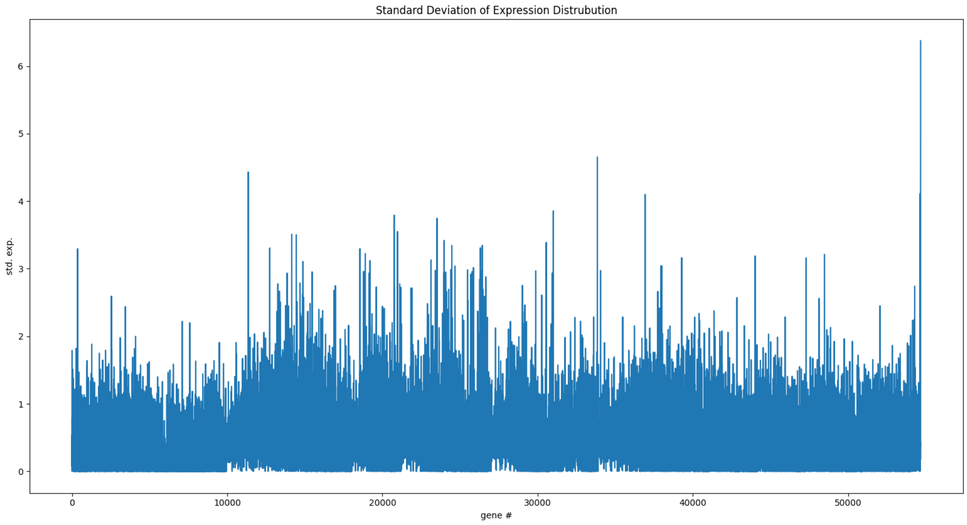

Figure 4. *Standard deviation of expression distribution for the exposed genes.*

Not-Exposed Group

Gene ID	1007_s_at	1053_at	117_at	121_at	1255_g_at	1294_at	1316_at	1320_at	1405_i_at	1431_at
Min expression	9.392436	6.252996	2.71127	9.240184	2.249185	5.422226	2.386696	2.621426	5.273909	2.249185
Max expression	10.79338	7.66106	8.000478	11.67132	2.49121	7.83027	3.678206	5.504938	12.14334	2.579561
Avg	10.13378	6.905059	3.646372	10.94597	2.256777	6.615797	2.840428	3.865763	8.168875	2.284327
Std.Dev.	0.314362	0.370982	1.375245	0.486917	0.042779	0.723979	0.281761	0.700195	1.794914	0.072756
Entropy	3.465736	3.465736	3.292449	3.465736	0.277113	3.465736	3.319419	3.465736	3.465736	1.960591

Table 2. *Attributes (Min, max, average, standard deviation, and entropy) of the Particular Not Exposed Genes*

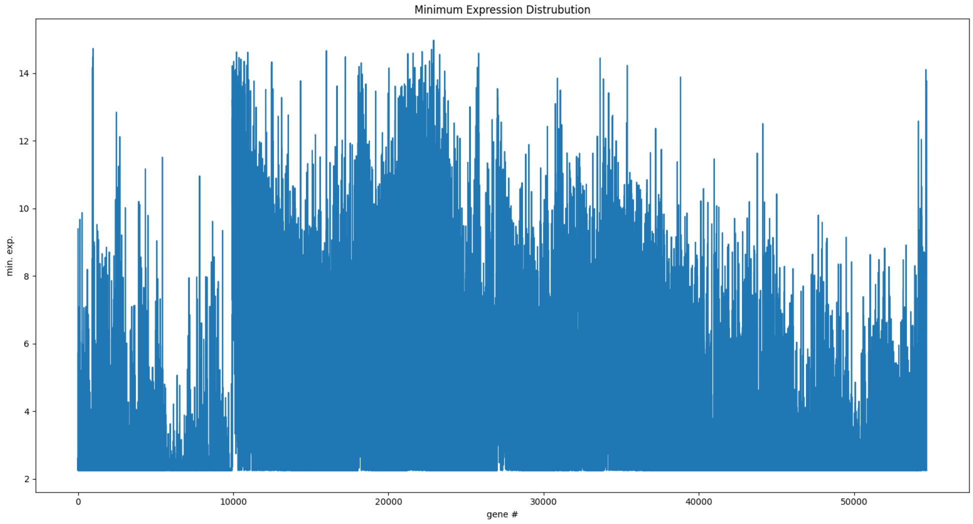

Figure 5. *Minimum expression distribution for the not exposed genes.*

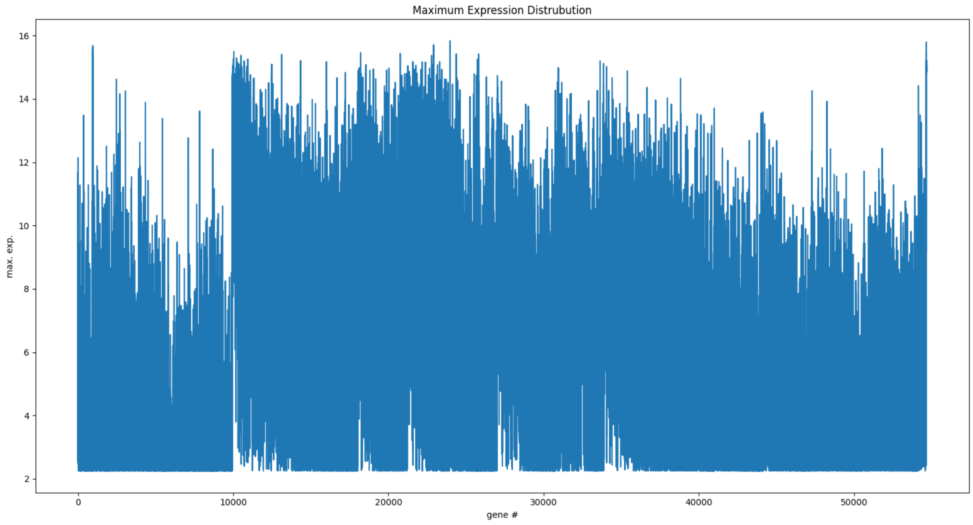

Figure 6. *Maximum expression distribution for the not exposed genes.*

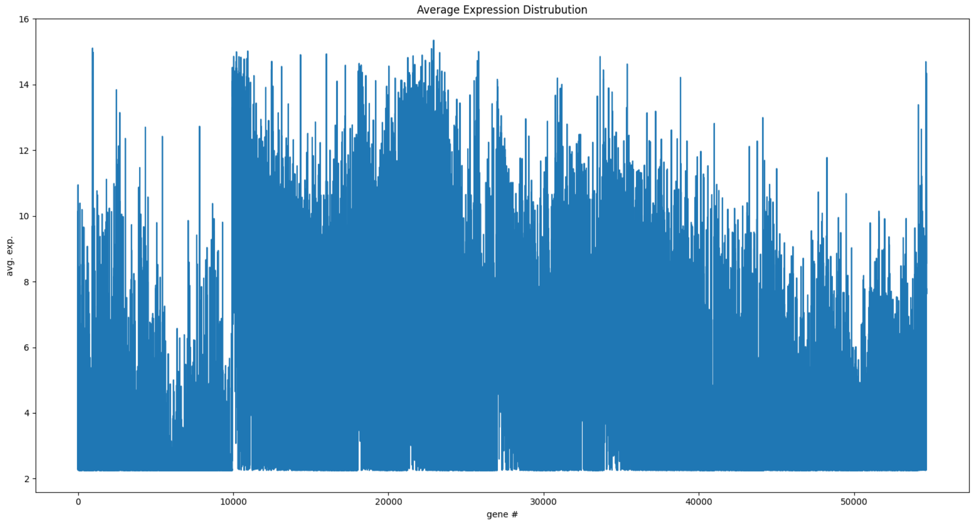

Figure 7. *Average expression distribution for the not exposed genes.*

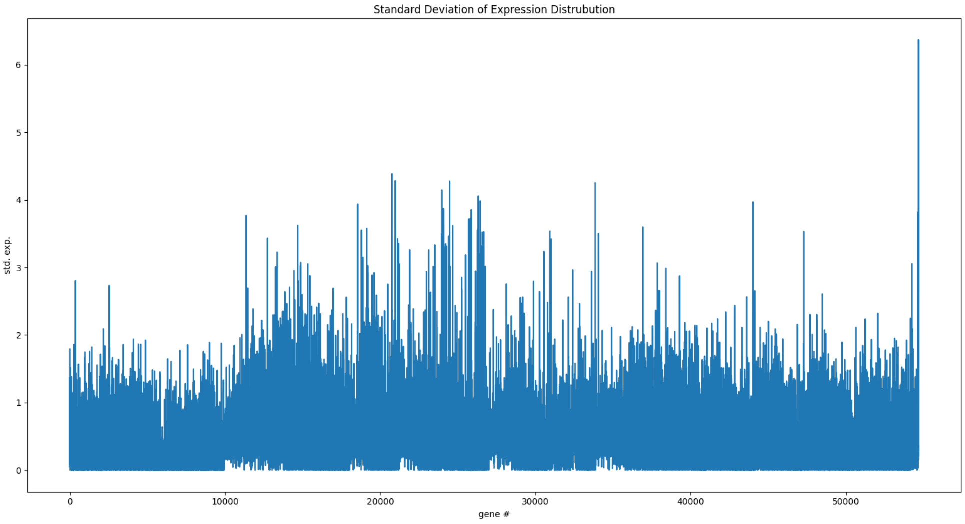

Figure 8. *Standard deviation of expression distribution for the not exposed genes.*

Control Group

Gene ID	1007_s_at	1053_at	117_at	121_at	1255_g_at	1294_at	1316_at	1320_at	1405_i_at	1431_at
Min_expression	8.216887	6.11793	2.505232	10.29233	2.249185	5.415159	2.530001	2.489	4.026917	2.249185
Max_expression	11.13376	8.063517	8.714269	12.57491	2.508066	7.402287	3.836213	6.061259	11.5126	2.438163
Avg	10.00116	6.808612	3.298227	11.99664	2.254261	6.204489	2.821502	4.701806	6.905349	2.26228
Std.Dev.	0.487925	0.361095	1.331429	0.436459	0.036251	0.455458	0.274949	0.808573	2.15435	0.035642
Entropy	3.931826	3.904643	3.785655	3.931826	0.096509	3.931826	3.622561	3.931826	3.931826	1.580919

Table 3. *Attributes (Min, max, average, standard deviation, and entropy) of the Particular Control Genes*

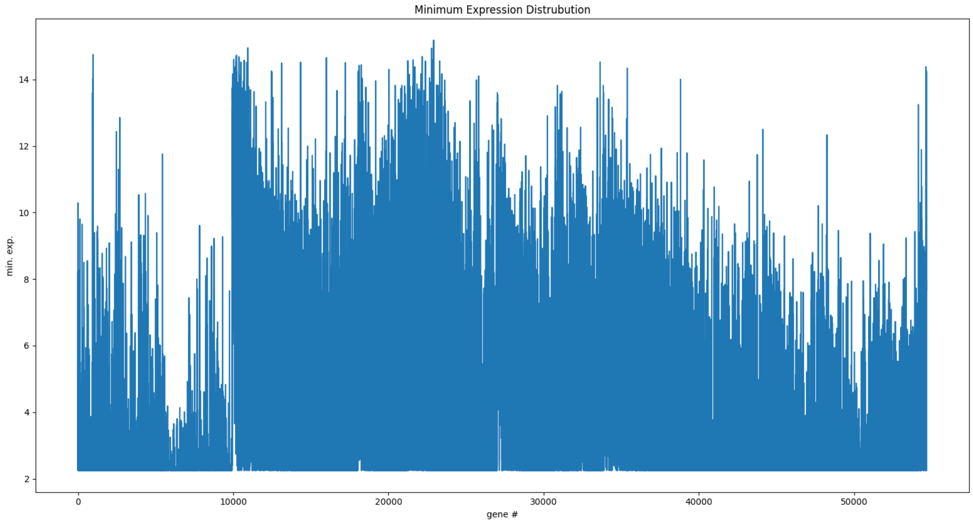

Figure 9. *Minimum expression distribution for the not exposed genes.*

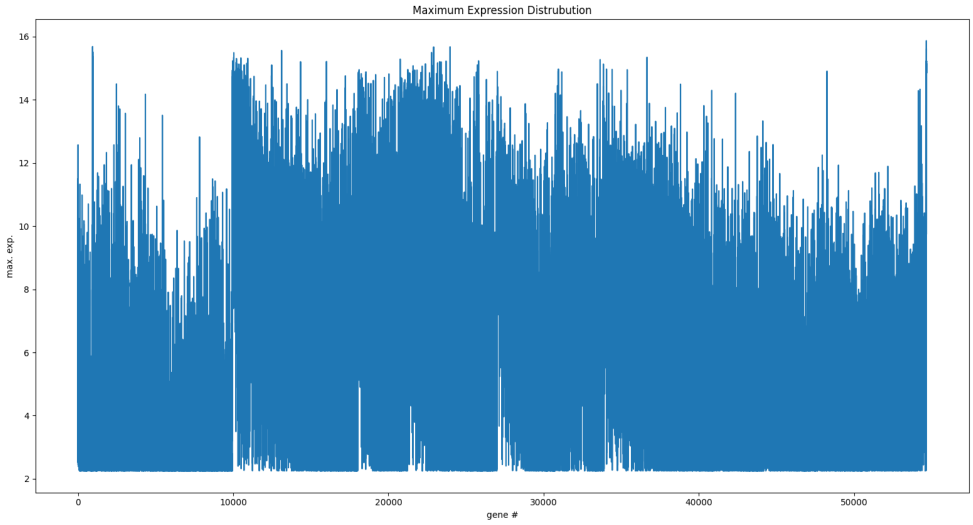

Figure 10. *Maximum expression distribution for the not exposed genes.*

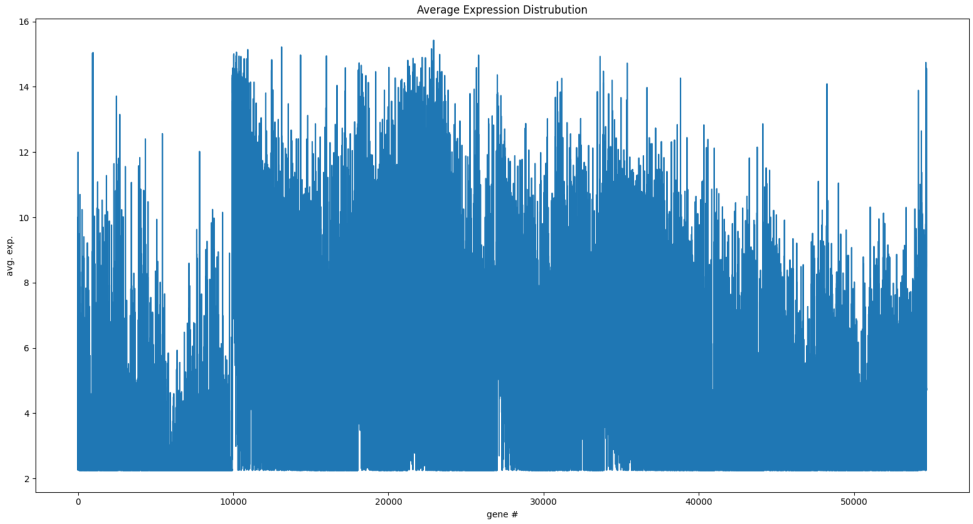

Figure 11. *Average expression distribution for the not exposed genes.*

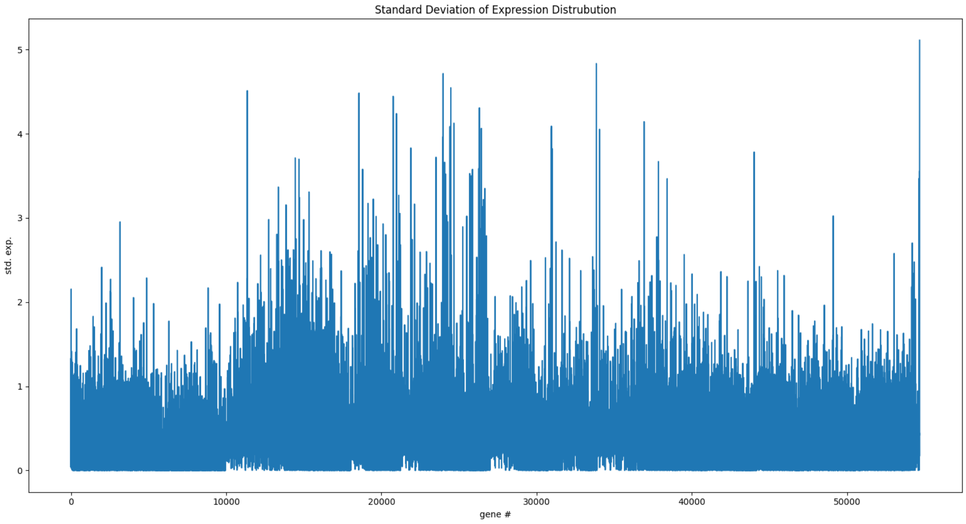

Figure 12. *Standard deviation of expression distribution for the not exposed genes.*