

Identification of **Thyroid Cancer** by Using Machine Learning

Ertuğrul Sağdıç

150116061

Defne ÇİĞ

150816004

Kasım Anlatır

150816019

Cem Telliagaoğlu

150216031

Lukas Eckerle

199520021

TABLE OF CONTENTS



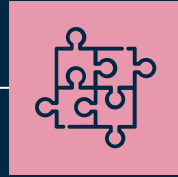
01

Proposal



02

Explore your
Data



03

Explore your
Data Part 2



04

Predictive
Analysis

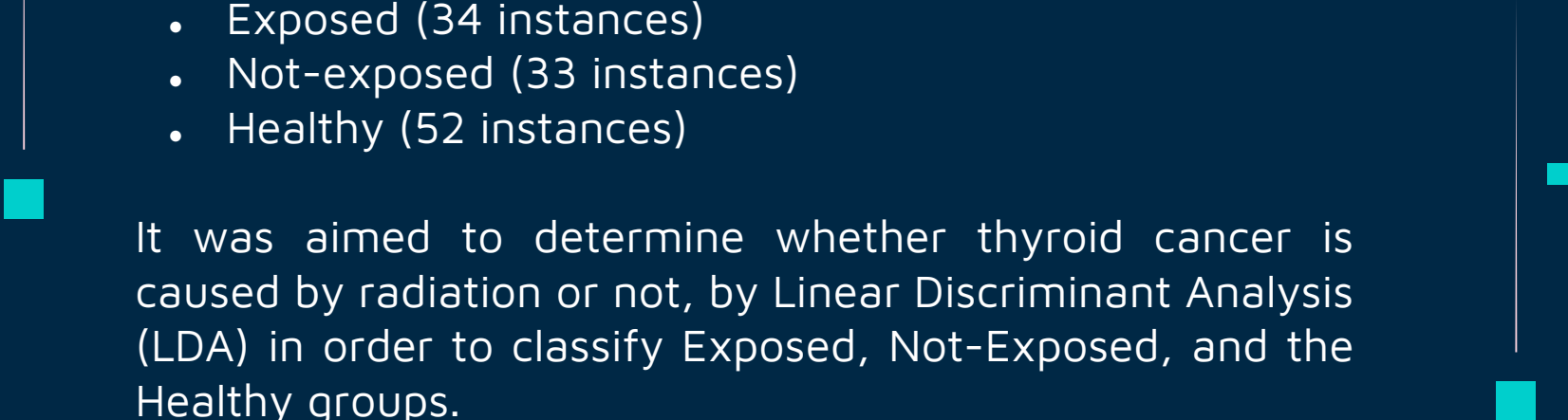
Proposal

01

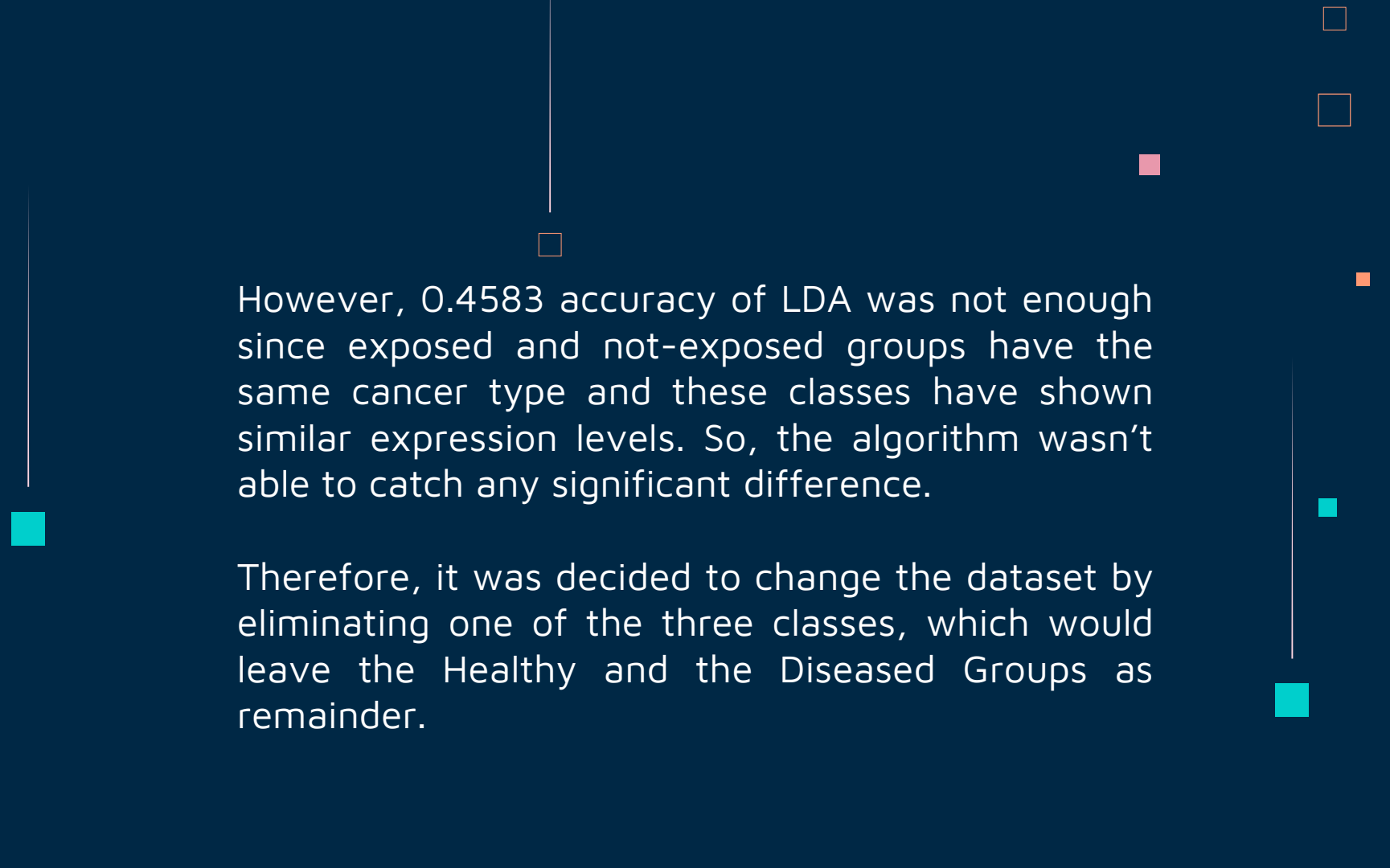


The dataset was made of three classes:

- Exposed (34 instances)
- Not-exposed (33 instances)
- Healthy (52 instances)



It was aimed to determine whether thyroid cancer is caused by radiation or not, by Linear Discriminant Analysis (LDA) in order to classify Exposed, Not-Exposed, and the Healthy groups.



However, 0.4583 accuracy of LDA was not enough since exposed and not-exposed groups have the same cancer type and these classes have shown similar expression levels. So, the algorithm wasn't able to catch any significant difference.

Therefore, it was decided to change the dataset by eliminating one of the three classes, which would leave the Healthy and the Diseased Groups as remainder.

A collection of small squares in various colors (cyan, pink, orange) arranged in a scattered pattern in the top right corner of the slide.

Description of the Project

The thyroid gland is an organ located under the thyroid cartilage that sits low on the front of the neck and has an endocrine function. A type of cancer caused by the transformation of cells in the thyroid gland into cancerous cells is called thyroid cancer and it has 4 subtypes: papillary, follicular, medullary, and anaplastic. Among these subtypes the most commonly seen is the papillary thyroid cancer (PTC)

The aim of this project is to classify samples that were taken from patients that are known to be diagnosed with the thyroid cancer and from the patients that are known to be healthy by using different classification algorithms such as k-NN and Neural Network.

The slide features a dark blue background with several decorative elements. A vertical line on the left has a teal square at its base. Another vertical line in the upper center has a small white square. A vertical line on the right has a pink square, a white square, a teal square, and a teal square at its base. A horizontal bar at the bottom is composed of an orange segment followed by a white segment.


Explore your Data

02

The data set contains gene expression levels of 54,675 genes as attributes.


The type of the attributes is numeric. There are two different class labels which are Diseased and Healthy.





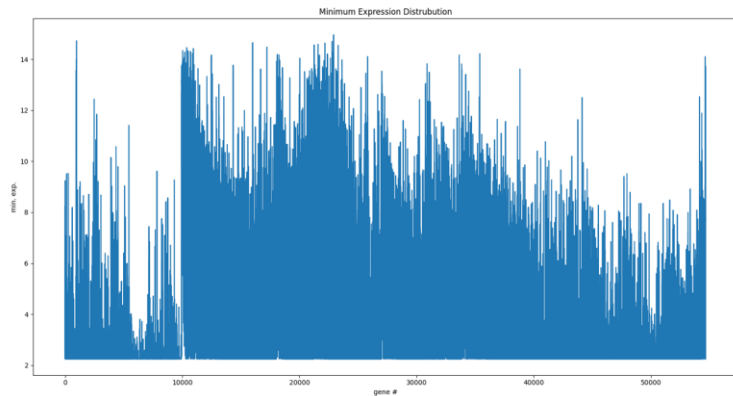
The minimum, maximum, average, standard deviation, and entropy of the attributes for each group's gene expression levels were investigated.

Only the first 10 attributes were investigated since there were too many attributes. Also, visualization of these attributes with plotting may provide better understanding of the data.

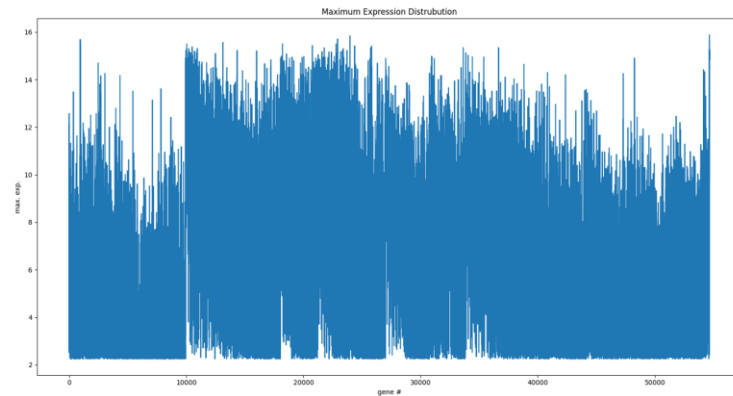


10 investigated attributes

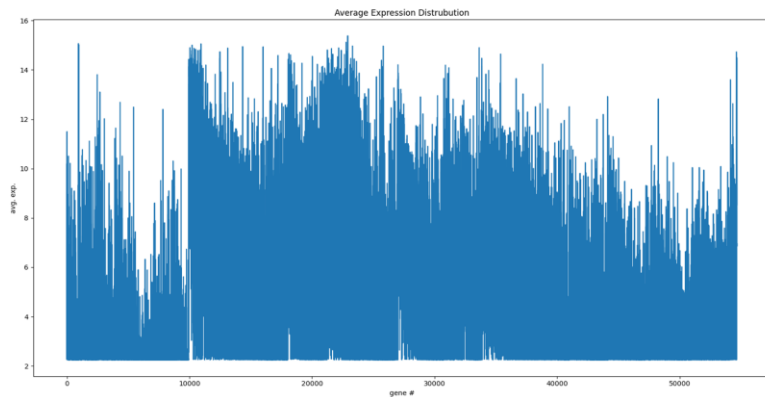
Gene ID	1007_s_a t	1053_at	117_at	121_at	1255_g_a t	1294_at	1316_at	1320_at	1405_i_at	1431_at
Min expression	8.216887	5.839823	2.505232	9.240184	2.249185	5.415159	2.386696	2.489	4.026917	2.249185
Max expression	11.59505	8.063517	8.714269	12.57491	2.508066	7.83027	4.154933	6.061259	12.14334	2.579561
Avg	10.17743	6.805636	3.323691	11.49767	2.253513	6.428938	2.868082	4.280523	7.376009	2.274973
Std.Dev.	0.460819	0.363002	1.188412	0.626009	0.03276	0.571834	0.305475	0.803372	2.012789	0.059632
Entropy	4.75359	4.705787	4.416789	4.75359	0.197798	4.75359	4.308748	4.75359	4.75359	2.05356



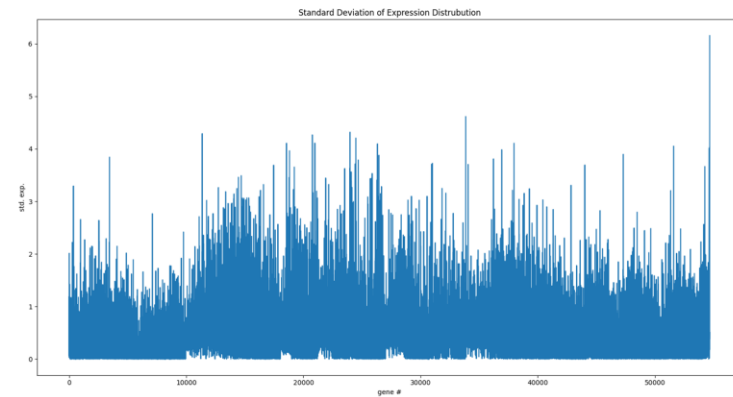
Minimum Expression Distribution



Maximum Expression Distribution



Average Expression Distribution



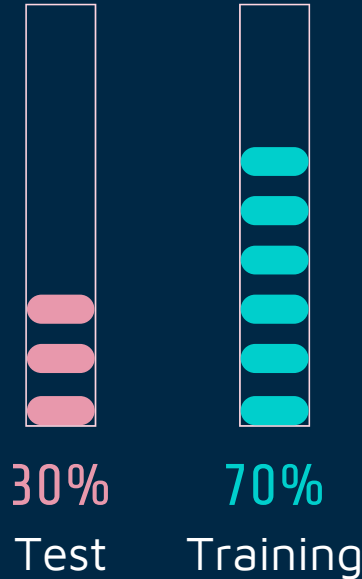
Standard Deviation Expression Distribution

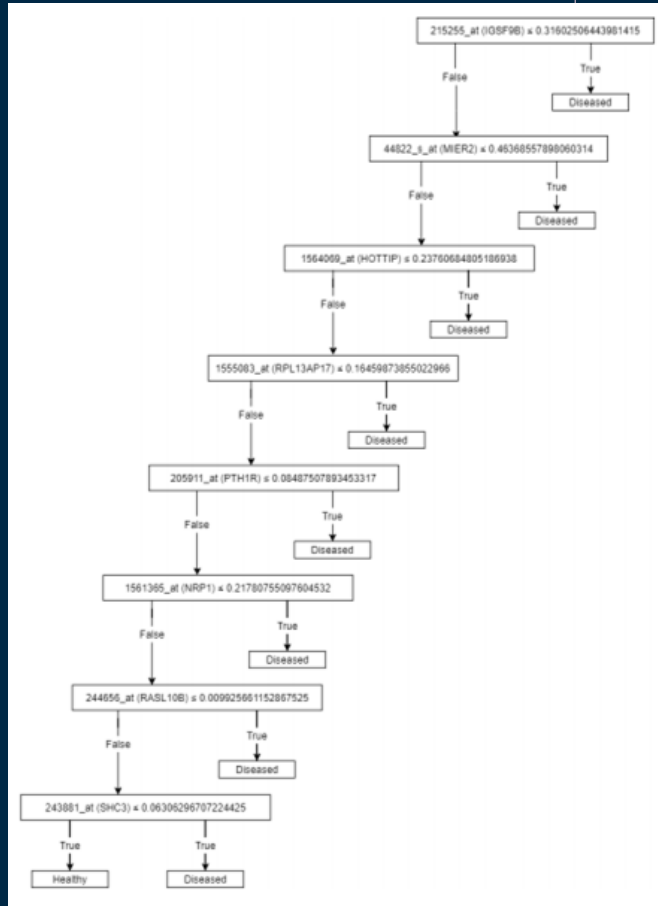
Explore your Data Part 2

03

The dataset was split into two sets for each investigation group. 70% of the data was split for the training set, 30% of the data was split for the test set.

The expression levels of the genes were checked by the decision tree algorithm and trees were constructed by checking the entropy values.





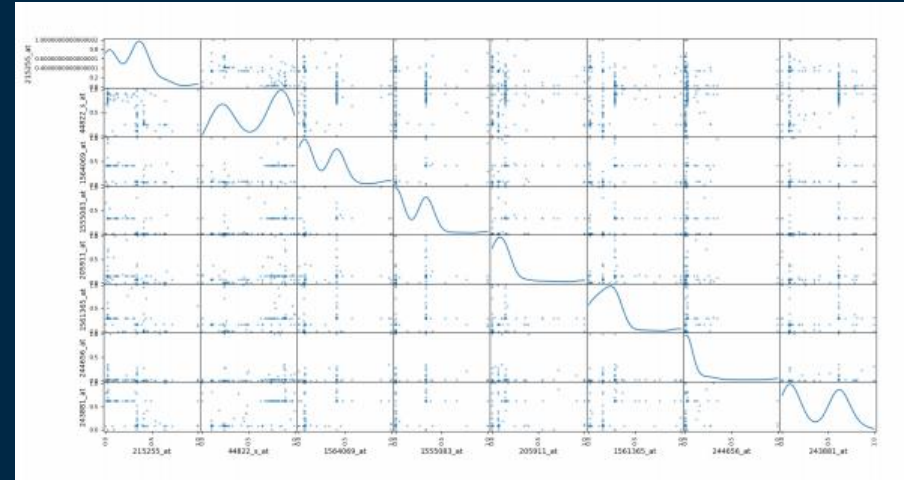
Then, the trees were added to the random forest with the accuracy level higher than 90% of the accuracy level. Then, the samples were classified based on the trees in the random forest.

	Gene ID	Gene Symbols	Cancer related findings
1	215255_at	IGSF9B	Favorable ovarian cancer biomarker
2	44822_s_at	MIER2	Possible biomarker in renal cancer, endometrial cancer and colorectal cancer
3	1564069_at	HOTTIP	Strong relation with the papillary thyroid carcinoma cells
4	1555083_at	RPL13AP17	No cancer related finding
5	205911_at	PTH1R	Enhanced cancer specificity in renal cancer
6	1561365_at	NRP1	Prognostic marker in stomach cancer, cervical cancer and renal cancer
7	244656_at	RASL10B	Unfavorable prognosis in endometrial cancer
8	243881_at	SHC3	Favorable hepatocellular carcinoma and unfavorable gastric cancer related findings

8 differentially expressed genes (DEGs) were discovered.

Differentially expressed genes were scattered by matrix plotting.

The points where the graph forms a positively inclined diagonal are positive correlations and the points where it forms a negatively inclined diagonal are negative correlations. However, in our case, no correlation was not observed



Predictive Analysis

04


Predictive analysis of our dataset is made by using classification algorithms, including K-Nearest Neighbors (k-NN) and Neural Network by using the differentially expressed genes found in the previous step.

We compared the performances of these classification algorithms trained with the training set which is 70% of our dataset and the classification capabilities on the test set by using the evaluation metrics. These metrics are accuracy, sensitivity, specification, precision, and F1 score

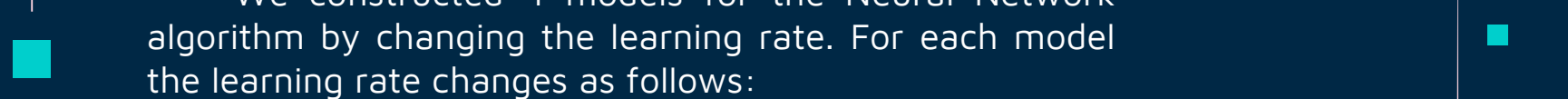


The evaluation metrics are calculated by using the Confusion Matrix.

Condition	Being Healthy	Having Disease
Being Healthy	TP	FN
Having Disease	FP	TN



We made experiments with models constructed consisting of 4 of the Neural Network and 4 of the k-NN algorithms.



We constructed 4 models for the Neural Network algorithm by changing the learning rate. For each model the learning rate changes as follows:

0.1, 0.01, 0.001, 0.0001.

We constructed 4 models for the k-NN algorithm by using different k values. For each model k value changes as follows:

3, 5, 9, 15.



Evaluation Metrics for each algorithm with different parameters

Algorithm	Experiment	Sensitivity	Specification	Precision	Precision	F1 Score
Neural Network	LR = 0.1	1.0	1.0	1.0	1.0	1.0
	LR = 0.01	1.0	1.0	1.0	1.0	1.0
	LR = 0.001	0.942	0.714	1.0	1.0	0.947
	LR = 0.0001	0.828	0.684	1.0	1.0	0.812
k-NN	k = 3	0.987	0.976	1.0	1.0	0.987
	k = 5	0.987	0.972	1.0	1.0	0.985
	k = 9	0.987	0.971	1.0	1.0	0.985
	k = 15	0.975	0.972	0.978	0.972	0.972

The background is a dark blue field decorated with a pattern of small, semi-transparent squares in teal, orange, and pink. Thin white vertical lines of varying lengths are scattered across the slide, some intersecting the colored squares.

Thank you for
listening
Any questions?