HPV preliminary diagnosis through complete blood sample test, a machine learning approach

Alejandro López Vázquez
Tecnológico de Monterrey
School of Engineering and Science

ABSTRACT

1. Cervical Cancer is a deadly desease which causes millions of deads yearly world wide. The main goal for this project was to develop a model which can pre-diagnose Human Papiloma Virus (HPV) with certain accuracy bases on a complete blood laboratory exam.

2. Key words:
    - Machine learning.
    - HPV.
    - Early detection.
    - PCR.
    - Neural Network

TABLE OF CONTENTS

01

INTRODUCTION

02

METHOD & DATA

03

RESULTS

TABLE OF CONTENTS

04

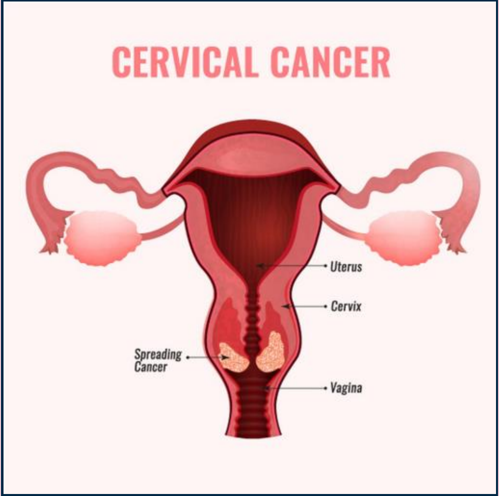
DISCUSSION

05

CONCLUSION

INTRODUCTION

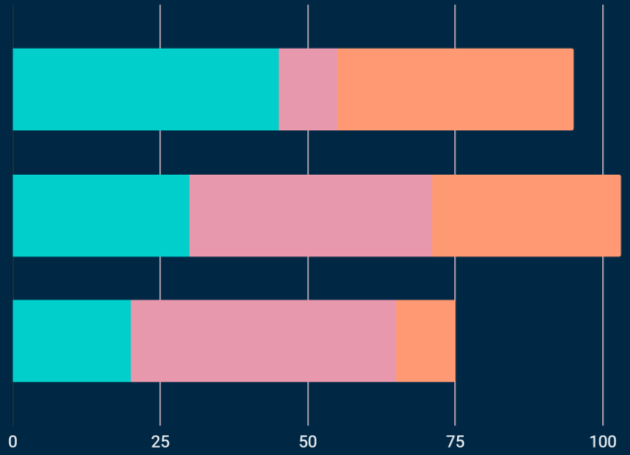
Cervical cancer is the third main cause of dead for women in Mexico. Cervical cancer is caused by some of the many HPV existing types and each can be more or less aggressive, types 16 and 18 being the most dangerous.



The research questions

1.- Can HPV be dectected through the use of machine learnig with data previously obtained of the patient?

The research questions



1.- Can HPV be pre-diagnosed based on a complete blood sample test?

The Data

Centers for Disease Control and Prevention

Source

The data was obtained from the UU.SS CDC from a the page of the National Health and Nutrition Examination Survey (NHANES). And we obtained the data specifically of labortory data from the Complete Blood Count with 5-part Differential - Whole Blood and Human Papillomavirus (HPV) DNA - Vaginal Swab from year 2003 to 2016

nhanes National Health and Nutrition Examination Survey

Complete Blood Count with 5-part Differential – Whole Blood

- SEQN - Respondent sequence number
- LBXWBCSI - White blood cell count (1000 cells/uL)
- LBXLYPCT - Lymphocyte percent (%)
- LBXMOPCT - Monocyte percent (%)
- LBXNEPCT - Segmented neutrophils percent (%)
- LBXEOPCT - Eosinophils percent (%)
- LBXBAPCT - Basophils percent (%)
- LBDLYMNO - Lymphocyte number (1000 cells/uL)
- LBDMONO - Monocyte number (1000 cells/uL)
- LBDNENO - Segmented neutrophils num (1000 cell/uL)

Complete Blood Count with 5-part Differential – Whole Blood

- LBDEONO - Eosinophils number (1000 cells/uL)
- LBDBANO - Basophils number (1000 cells/uL)
- LBXRBCSI - Red blood cell count (million cells/uL)
- LBXHGB - Hemoglobin (g/dL)
- LBXHCT - Hematocrit (%)
- LBXMCVSI - Mean cell volume (fL)
- LBXMCHSI - Mean cell hemoglobin (pg)
- LBXMC - Mean Cell Hgb Conc. (g/dL)
- LBXRDW - Red cell distribution width (%)
- LBXPLTSI - Platelet count (1000 cells/uL)
- LBXMPSI - Mean platelet volume (fL)
- LBXNRBC - Nucleated red blood cells

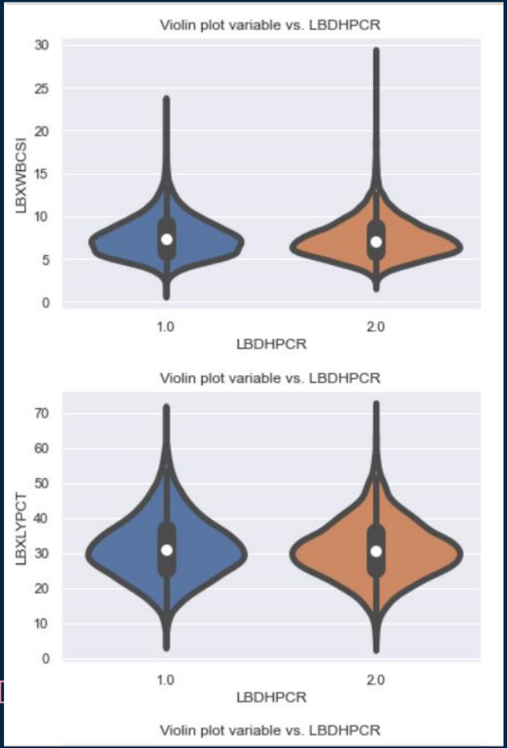
Human Papillomavirus (HPV) DNA - Vaginal Swab: **Roche Linear Array**

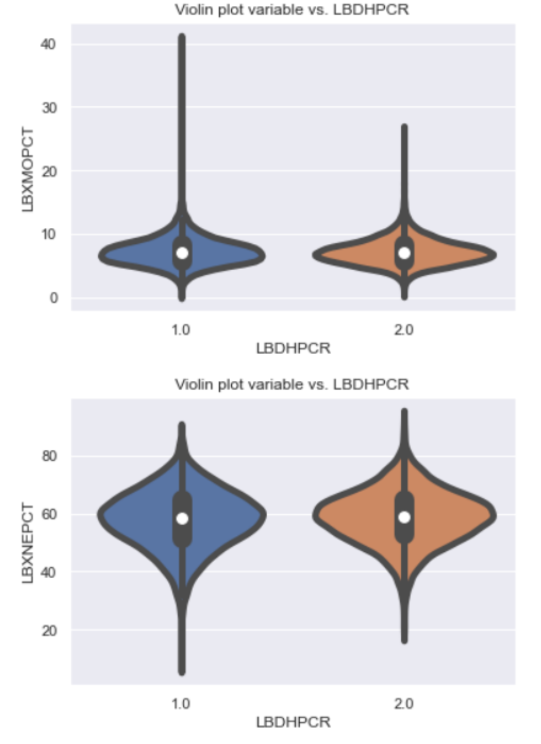
- SEQN - Respondent sequence number

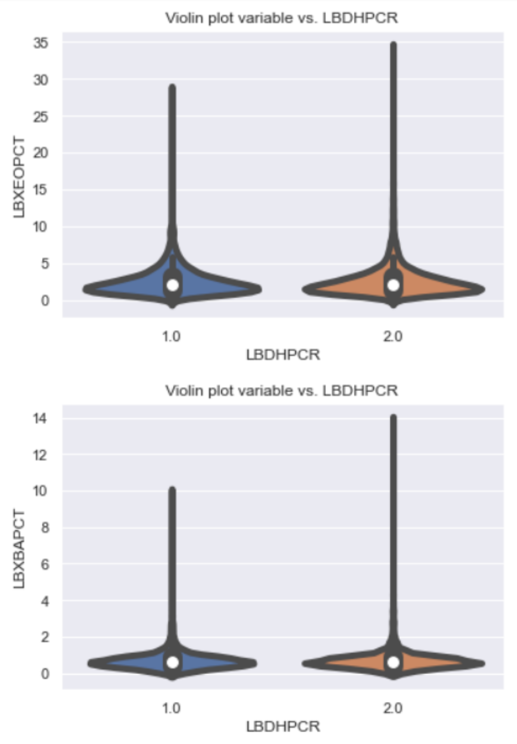
- LBDRPCR - Roche HPV linear array summary result

LBDRPCR - The HPV PCR Summary variable indicates if at least one type is positive (LBDRPCR=1); the sample is negative (LBDRPCR=2); the sample is inadequate (LBDRPCR=3); or the sample is missing (LBDRPCR=.)

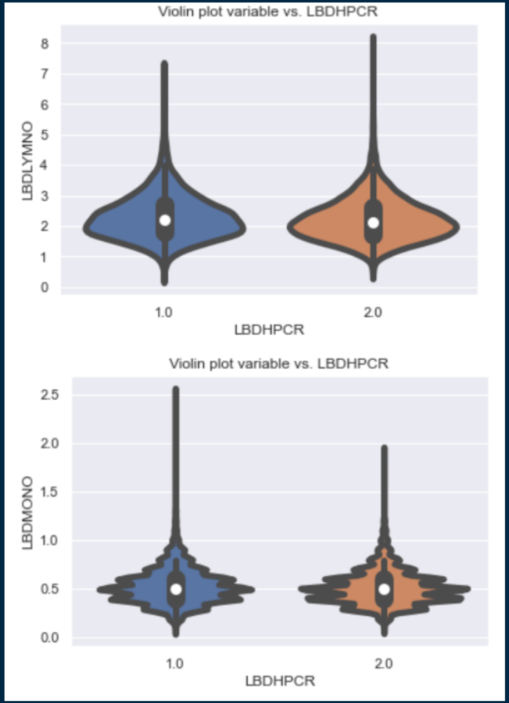
Data Understanding



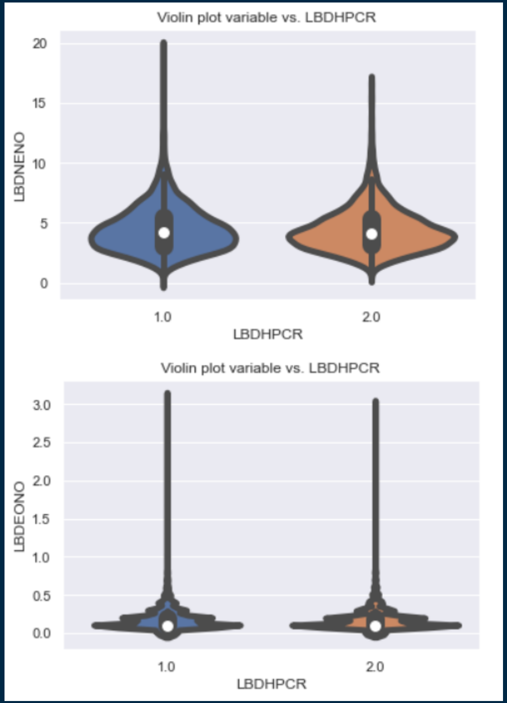




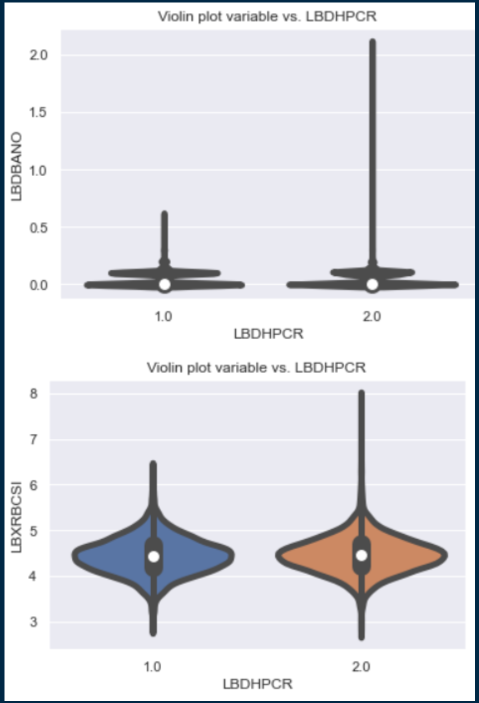
Data Understanding


Violin plot variable vs. LBDHPCR

Violin plot variable vs. LBDHPCR

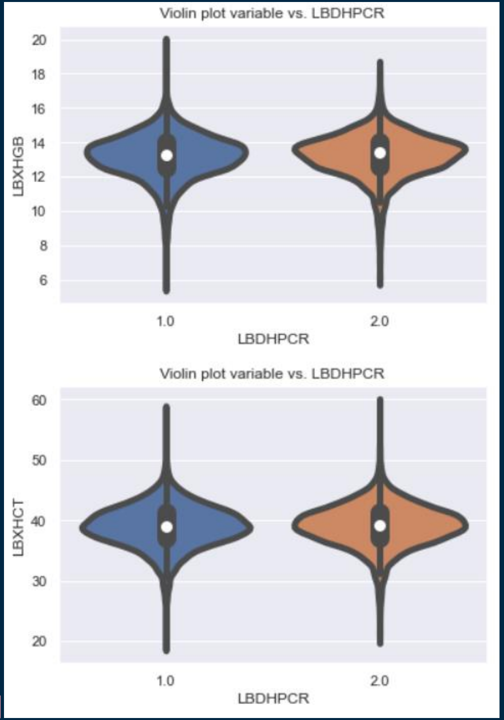

Violin plot variable vs. LBDHPCR

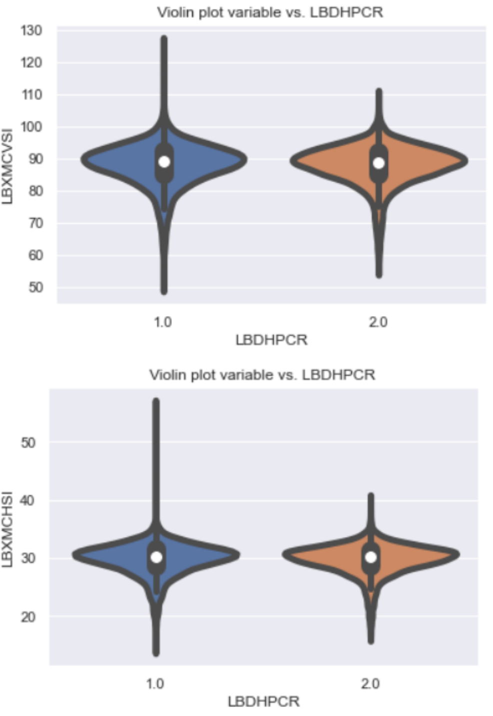
Violin plot variable vs. LBDHPCR

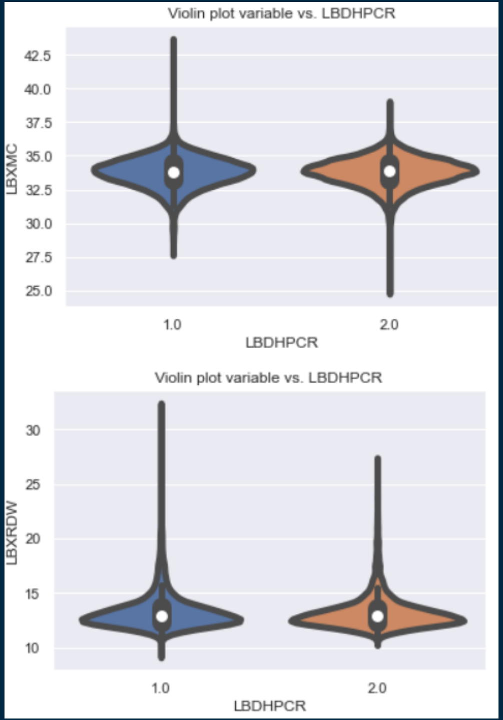

Violin plot variable vs. LBDHPCR

Violin plot variable vs. LBDHPCR

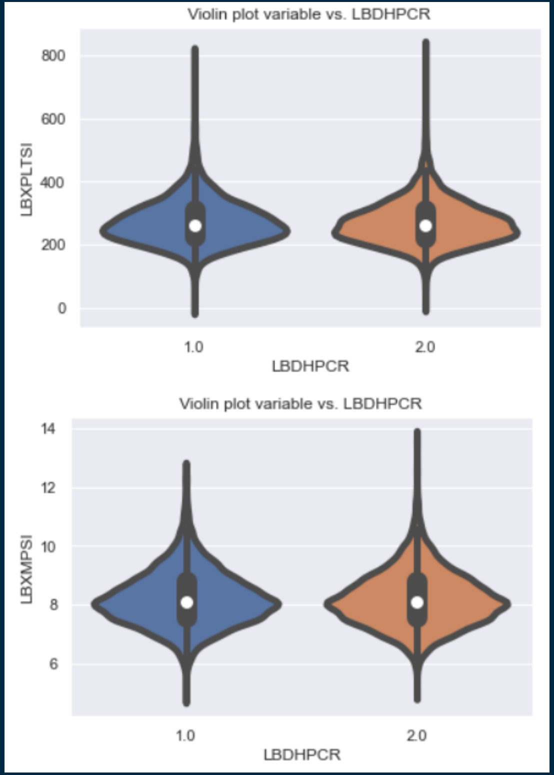
Data Understanding





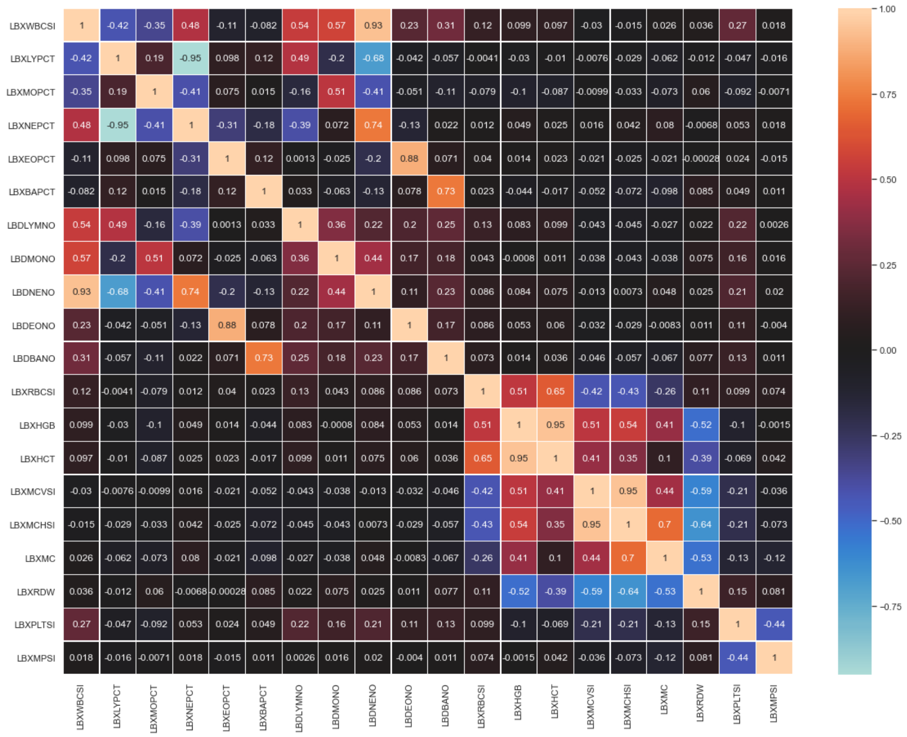


Data Understanding

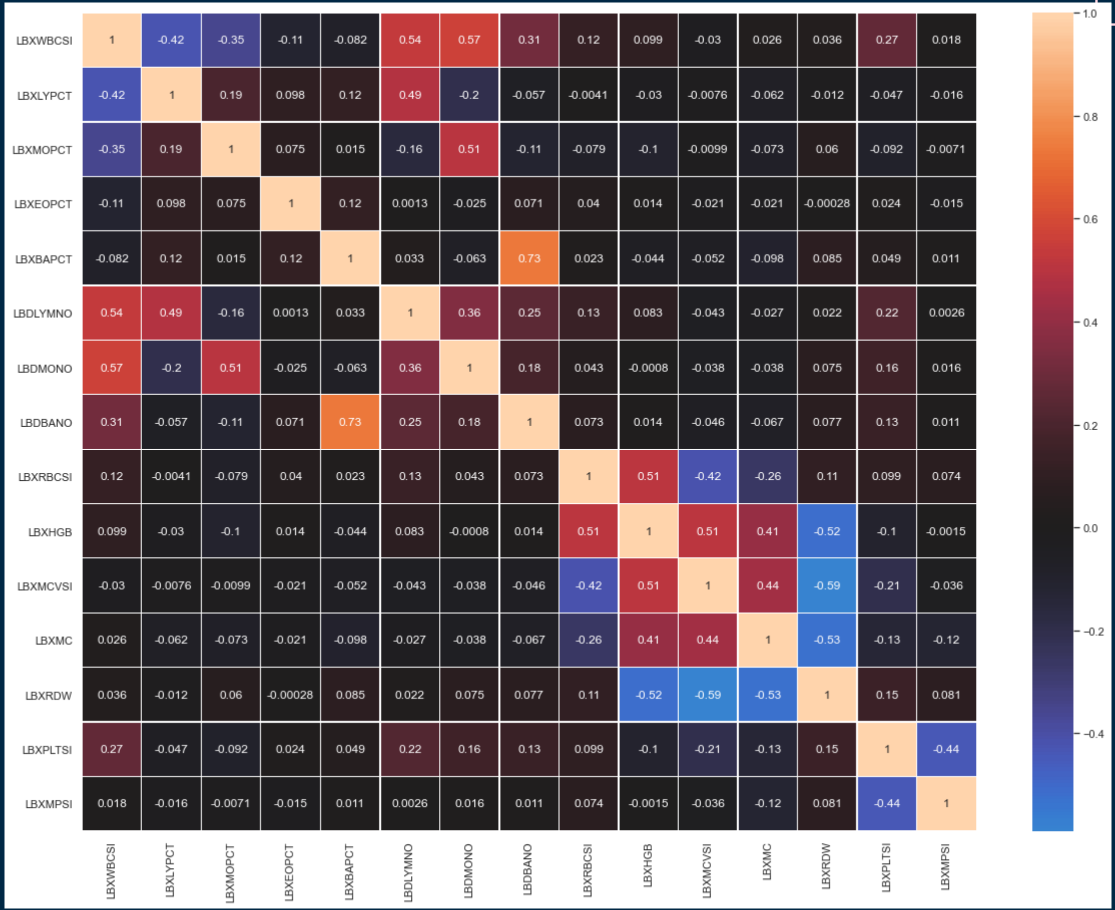


Correlation Matrix Removed Variables

- LBDEONO
- LBDNENO
- LBXNEPCT
- LBXHCT
- LBXMCHSI



	LBXWBCSI	LBXLYPCT	LBXMOPCT	LBXNEPCT	LBXEOPCT	LBXBAPCT	LBDLYMNO	LBDMONO	LBDNENO	LBDEONO	LBDBANO	LBXRBCSI	LBXHGB	LBXHCT	LBXMCVSI	LBXMCHSI	LBXMC	LBXRDW	LBXPLTSI	LBXMPSI
LBXWBCSI	1	-0.42	-0.35	0.48	-0.11	-0.082	0.54	0.57	0.93	0.23	0.31	0.12	0.099	0.097	-0.03	-0.015	0.026	0.036	0.27	0.018
LBXLYPCT	-0.42	1	0.19	-0.95	0.098	0.12	0.49	-0.2	-0.68	-0.042	-0.057	-0.0041	-0.03	-0.01	-0.0076	-0.029	-0.062	-0.012	-0.047	-0.016
LBXMOPCT	-0.35	0.19	1	-0.41	0.075	0.015	-0.16	0.51	-0.41	-0.051	-0.11	-0.079	-0.1	-0.087	-0.0099	-0.033	-0.073	0.06	-0.092	-0.0071
LBXNEPCT	0.48	-0.95	-0.41	1	-0.31	-0.18	-0.39	0.072	0.74	-0.13	0.022	0.012	0.049	0.025	0.016	0.042	0.08	-0.0068	0.053	0.018
LBXEOPCT	-0.11	0.098	0.075	-0.31	1	0.12	0.0013	-0.025	-0.2	0.88	0.071	0.04	0.014	0.023	-0.021	-0.025	-0.021	-0.00028	0.024	-0.015
LBXBAPCT	-0.082	0.12	0.015	-0.18	0.12	1	0.033	-0.063	-0.13	0.078	0.73	0.023	-0.044	-0.017	-0.052	-0.072	-0.098	0.085	0.049	0.011
LBDLYMNO	0.54	0.49	-0.16	-0.39	0.0013	0.033	1	0.36	0.22	0.2	0.25	0.13	0.083	0.099	-0.043	-0.045	-0.027	0.022	0.22	0.0026
LBDMONO	0.57	-0.2	0.51	0.072	-0.025	-0.063	0.36	1	0.44	0.17	0.18	0.043	-0.0008	0.011	-0.038	-0.043	-0.038	0.075	0.16	0.016
LBDNENO	0.93	-0.68	-0.41	0.74	-0.2	-0.13	0.22	0.44	1	0.11	0.23	0.086	0.084	0.075	-0.013	0.0073	0.048	0.025	0.21	0.02
LBDEONO	0.23	-0.042	-0.051	-0.13	0.88	0.078	0.2	0.17	0.11	1	0.17	0.086	0.053	0.06	-0.032	-0.029	-0.0083	0.011	0.11	-0.004
LBDBANO	0.31	-0.057	-0.11	0.022	0.071	0.73	0.25	0.18	0.23	0.17	1	0.073	0.014	0.036	-0.046	-0.057	-0.067	0.077	0.13	0.011
LBXRBCSI	0.12	-0.0041	-0.079	0.012	0.04	0.023	0.13	0.043	0.086	0.086	0.073	1	0.51	0.65	-0.42	-0.43	-0.26	0.11	0.099	0.074
LBXHGB	0.099	-0.03	-0.1	0.049	0.014	-0.044	0.083	-0.0008	0.084	0.053	0.014	0.51	1	0.95	0.51	0.54	0.41	-0.52	-0.1	-0.0015
LBXHCT	0.097	-0.01	-0.087	0.025	0.023	-0.017	0.099	0.011	0.075	0.06	0.036	0.65	0.95	1	0.41	0.35	0.1	-0.39	-0.069	0.042
LBXMCVSI	-0.03	-0.0076	-0.0099	0.016	-0.021	-0.052	-0.043	-0.038	-0.013	-0.032	-0.046	-0.42	0.51	0.41	1	0.95	0.44	-0.59	-0.21	-0.036
LBXMCHSI	-0.015	-0.029	-0.033	0.042	-0.025	-0.072	-0.045	-0.043	0.0073	-0.029	-0.057	-0.43	0.54	0.35	0.95	1	0.7	-0.64	-0.21	-0.073
LBXMC	0.026	-0.062	-0.073	0.08	-0.021	-0.098	-0.027	-0.038	0.048	-0.0083	-0.067	-0.26	0.41	0.1	0.44	0.7	1	-0.53	-0.13	-0.12
LBXRDW	0.036	-0.012	0.06	-0.0068	-0.00028	0.085	0.022	0.075	0.025	0.011	0.077	0.11	-0.52	-0.39	-0.59	-0.64	-0.53	1	0.15	0.081
LBXPLTSI	0.27	-0.047	-0.092	0.053	0.024	0.049	0.22	0.16	0.21	0.11	0.13	0.099	-0.1	-0.069	-0.21	-0.21	-0.13	0.15	1	-0.44
LBXMPSI	0.018	-0.016	-0.0071	0.018	-0.015	0.011	0.0026	0.016	0.02	-0.004	0.011	0.074	-0.0015	0.042	-0.036	-0.073	-0.12	0.081	-0.44	1



	LBXWBCSI	LBXLYPCT	LBXMOPCT	LBXEOPCT	LBXBAPCT	LBDLYMNO	LBDMONO	LBDBANO	LBXRBCSI	LBXHGB	LBXMCVSI	LBXMC	LBXRDW	LBXPLTSI	LBXMPSI
LBXWBCSI	1	-0.42	-0.35	-0.11	-0.082	0.54	0.57	0.31	0.12	0.099	-0.03	0.026	0.036	0.27	0.018
LBXLYPCT	-0.42	1	0.19	0.098	0.12	0.49	-0.2	-0.057	-0.0041	-0.03	-0.0076	-0.062	-0.012	-0.047	-0.016
LBXMOPCT	-0.35	0.19	1	0.075	0.015	-0.16	0.51	-0.11	-0.079	-0.1	-0.0099	-0.073	0.06	-0.092	-0.0071
LBXEOPCT	-0.11	0.098	0.075	1	0.12	0.0013	-0.025	0.071	0.04	0.014	-0.021	-0.021	-0.00028	0.024	-0.015
LBXBAPCT	-0.082	0.12	0.015	0.12	1	0.033	-0.063	0.73	0.023	-0.044	-0.052	-0.098	0.085	0.049	0.011
LBDLYMNO	0.54	0.49	-0.16	0.0013	0.033	1	0.36	0.25	0.13	0.083	-0.043	-0.027	0.022	0.22	0.0026
LBDMONO	0.57	-0.2	0.51	-0.025	-0.063	0.36	1	0.18	0.043	-0.0008	-0.038	-0.038	0.075	0.16	0.016
LBDBANO	0.31	-0.057	-0.11	0.071	0.73	0.25	0.18	1	0.073	0.014	-0.046	-0.067	0.077	0.13	0.011
LBXRBCSI	0.12	-0.0041	-0.079	0.04	0.023	0.13	0.043	0.073	1	0.51	-0.42	-0.26	0.11	0.099	0.074
LBXHGB	0.099	-0.03	-0.1	0.014	-0.044	0.083	-0.0008	0.014	0.51	1	0.51	0.41	-0.52	-0.1	-0.0015
LBXMCVSI	-0.03	-0.0076	-0.0099	-0.021	-0.052	-0.043	-0.038	-0.046	-0.42	0.51	1	0.44	-0.59	-0.21	-0.036
LBXMC	0.026	-0.062	-0.073	-0.021	-0.098	-0.027	-0.038	-0.067	-0.26	0.41	0.44	1	-0.53	-0.13	-0.12
LBXRDW	0.036	-0.012	0.06	-0.00028	0.085	0.022	0.075	0.077	0.11	-0.52	-0.59	-0.53	1	0.15	0.081
LBXPLTSI	0.27	-0.047	-0.092	0.024	0.049	0.22	0.16	0.13	0.099	-0.1	-0.21	-0.13	0.15	1	-0.44
LBXMPSI	0.018	-0.016	-0.0071	-0.015	0.011	0.0026	0.016	0.011	0.074	-0.0015	-0.036	-0.12	0.081	-0.44	1

Data preparation

Variable normalization

All the variables are continuos so in order to prepare the data a normalization conversion was performed in oorder to place them between the values of 0 and 1

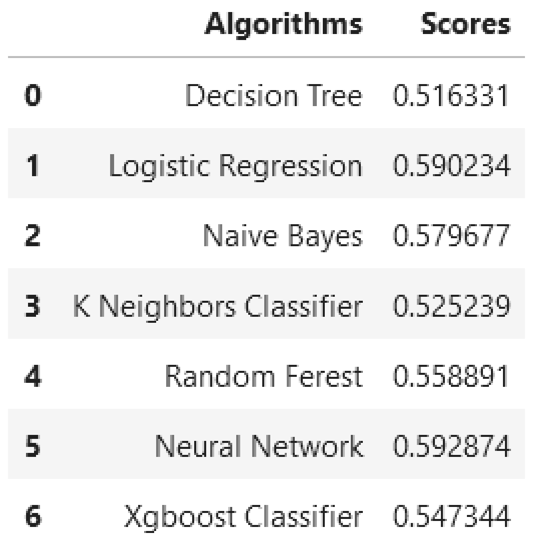
Train and test splitting

The data was divided into target and predictors, and for both of them a division for a training an a test set was performed with a 75% and a 25% of the data respectively

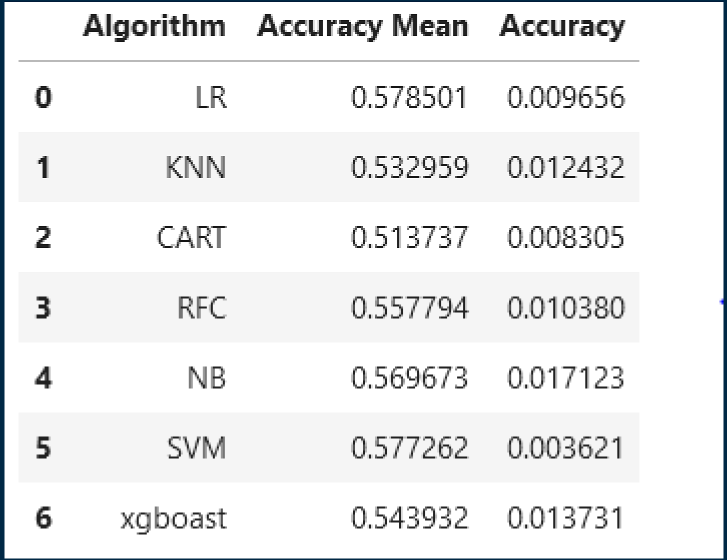
Models

- Decision Tree
- Logistic Regression
- Naive Bayes
- K neares Neighbors
- Random Forest
- Multi-Layer Perception

Results

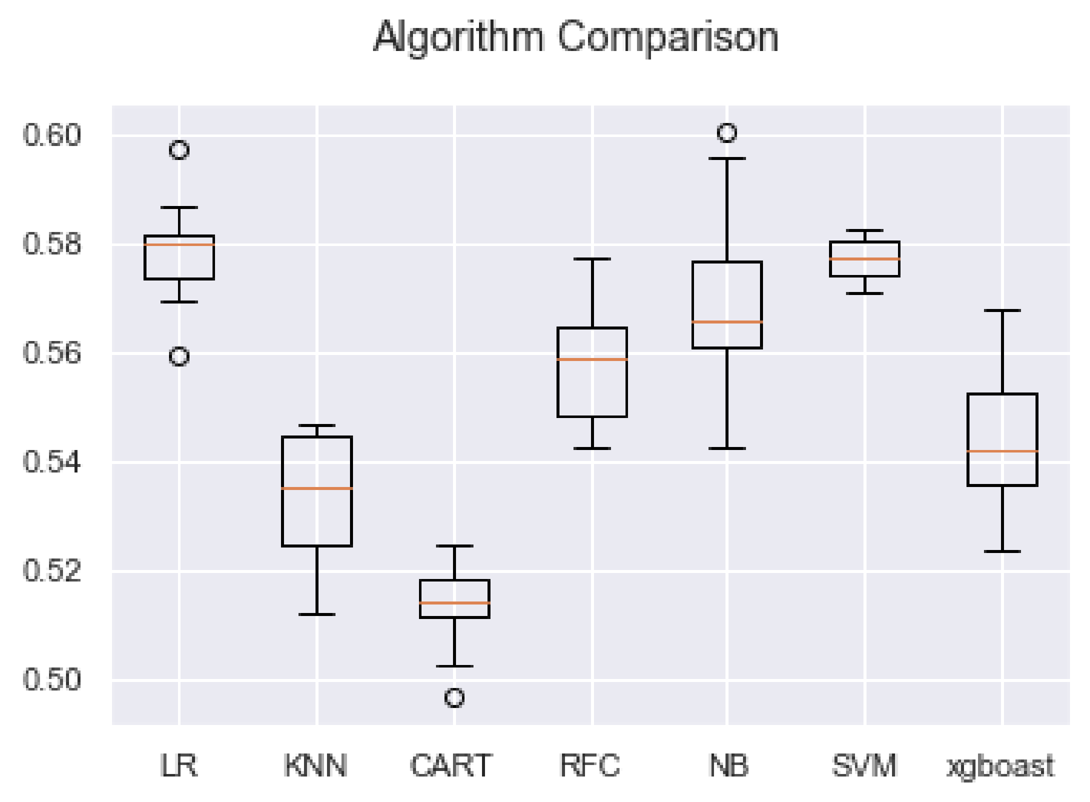


	Algorithms	Scores
0	Decision Tree	0.516331
1	Logistic Regression	0.590234
2	Naive Bayes	0.579677
3	K Neighbors Classifier	0.525239
4	Random Ferest	0.558891
5	Neural Network	0.592874
6	Xgboost Classifier	0.547344

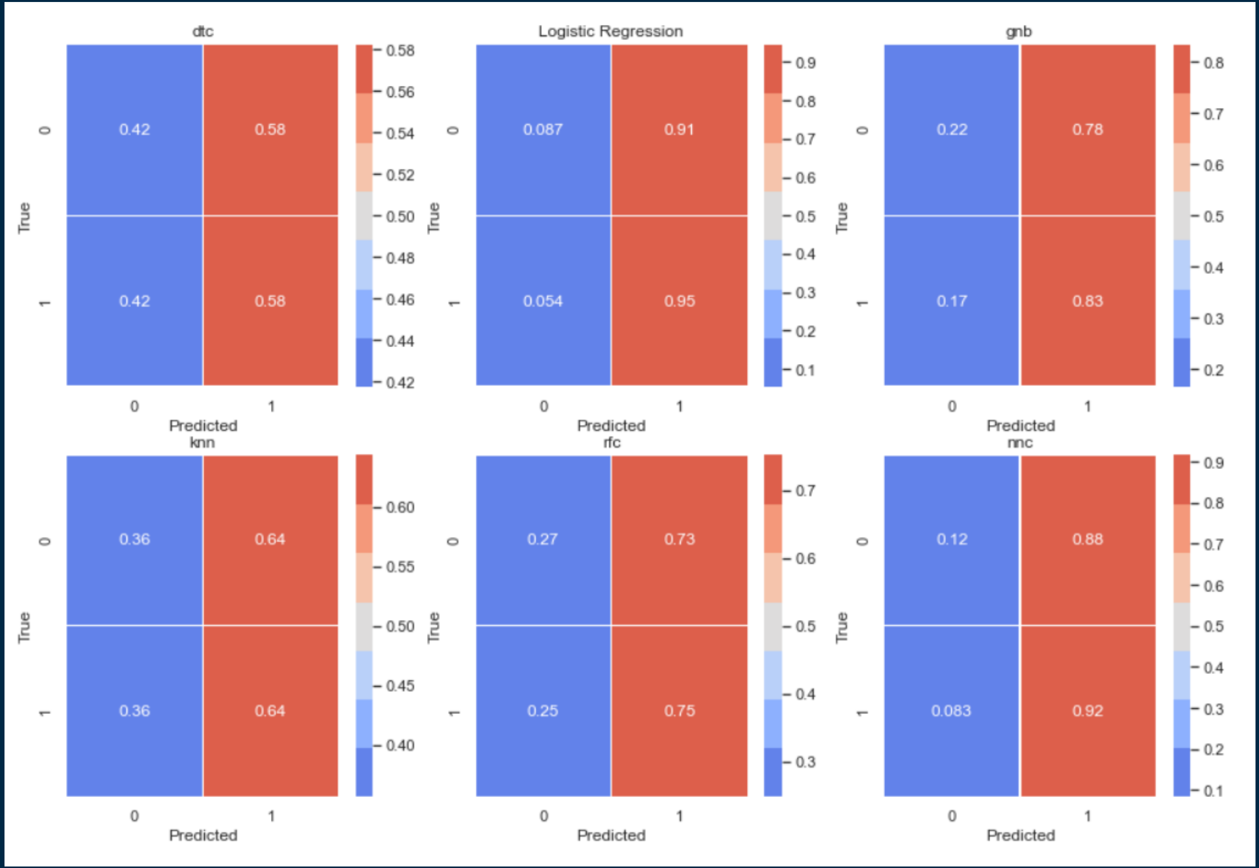


	Algorithm	Accuracy Mean	Accuracy
0	LR	0.578501	0.009656
1	KNN	0.532959	0.012432
2	CART	0.513737	0.008305
3	RFC	0.557794	0.010380
4	NB	0.569673	0.017123
5	SVM	0.577262	0.003621
6	xgboast	0.543932	0.013731

Comparisson Boxplot

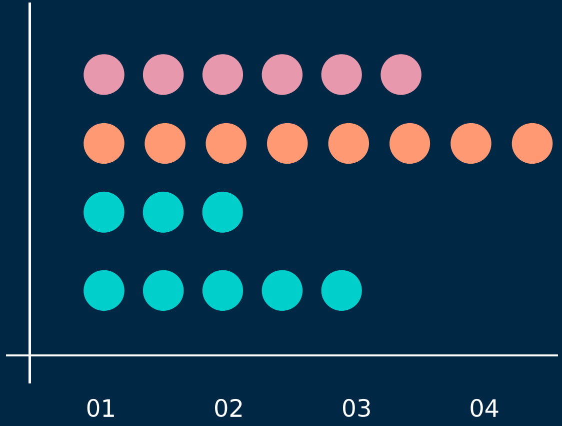


Correlation matrixes



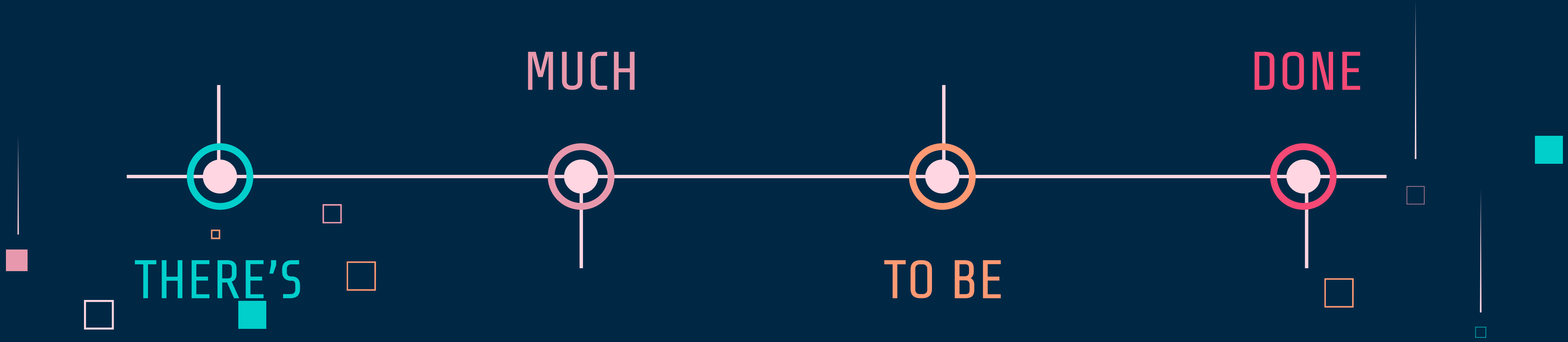
Discussion

- The propossed models cannot preddict with Good accuaricy the presence of HPV with the complete blood laboratory results.



Conclusion

- For future work:
  - There is still a lot of information in the NHANES page, including different laboratories or even quesionares filed by the patients
  - This model can be used also for other diseases uncluded in the same data set.



THERE'S MUCH TO BE DONE

ANY QUESTIONS?

CREDITS: This presentation template was created by Slidesgo,
including icons by Flaticon, and infographics & images by Freepik