COVID-19 correlation with BCG vaccination policies a CRISP-DM approach

Diaz-Moraila J.E.
School of Sciences and Engineering
Tec de Monterrey
Monterrey, Mexico
A00828174@itesm.mx

***Abstract*—In December 2019, COVID-19 coronavirus was first identified in the Wuhan region of China. By March 11, 2020, the World Health Organization (WHO) categorized the COVID-19 outbreak as a pandemic. A lot has happened in the months in between with major outbreaks in Iran, South Korea, and Italy.**

We know that COVID-19 spreads through respiratory droplets, such as through coughing, sneezing, or speaking. But, how quickly did the virus spread across the globe? And whats the impact of the disease in different countries?.

Modeling the COVID-19 to answer to this questions is not something that's simple to accomplish, since there are different factors across countries that make this estimations fuzzy.

In this work its proposed a CRSIP-DM approach to explain one the factor that may lead to create a good model which explains national differences in COVID19 enriching data with different national policies respect to Bacillus Calmette-Guérin (BCG) childhood vaccination.

***Index Terms*—COVID-19, Modeling, BCG, vaccination, CRISP-DM**

I. Introduction

The COVID-19 pandemic originated in December 2019 in Wuhan China and it has quickly spread over all continents affecting most countries in the world. However, there are some striking differences on how COVID-19 is behaving in different countries.

Despite organizations around the world have been collecting data so that governments can monitor and learn from this pandemic. Notably, the Johns Hopkins University Center for Systems Science and Engineering created a publicly available data repository to consolidate this data from sources like the WHO, the Centers for Disease Control and Prevention (CDC), and the Ministry of Health from multiple countries, there exist gaps in several data entries that may lead to a good model that explain the COVID-19 behaviour.

A. Business understanding

As it is explained in **fivethirtyeigh** post [8] *"It's relatively easy to put together — the sort of thing people on our staff do while buzzed on a socially isolated conference call after work. The number of people who will die is a function of how many people could become infected, how the virus spreads and how many people the virus is capable of killing"*.

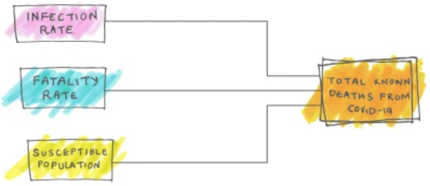


Fig. 1. COVID-19 model deconstruction archetype taken from [8]

"But then you start trying to fill in the blanks. Every variable is dependent on a number of choices and knowledge gaps. And if every individual piece of a model is wobbly, then the model is going to have as much trouble standing on its own as a data journalist who has spent too long on a conference call while socially isolated after work".

It is known that Different countries and regions collect data in different ways. There's no single spreadsheet everyone is filling out that can easily allow us to compare cases and deaths around the world. Even within the United States, doctors say we're under-reporting the total number of deaths due to COVID-19 [9].

The same inconsistencies apply to who gets tested. Some countries are giving tests to anyone who wants one. Others are not [7]. That affects how much we can know about how many people have actually contracted COVID-19, versus how many people have tested positive.

And the virus itself is an unpredictable contagion, hurting some groups more than others [3], meaning that local demographics and health care access are going to be big determinants when it comes to the virus' impact on communities.

The main objective of this work is to provide some insights and further exploration of the correlation between countries that have BCG vaccination policies with the morbidity and mortality for COVID-19, as well as proposing some predictive models which can be trained with the same features; The whole process is done under the light of the Cross-industry Standard Process for Data Mining (CRISP-DM).

II. Materials and Methods

The COVID-19 dataset is obtained by the R package *coronavirus* where the raw data is pulled and arranged by the Johns Hopkins University Center for Systems Science and Engineering (JHU CCSE) from the resources described in https://ramikrispin.github.io/coronavirus/#data-sources .
For the enriching part it is used the Immunization coverage dataset https://www.who.int/immunization/monitoring_surveillance/data/en/ and http://www.bcgatlas.org/.

In this work as it was mentioned the CRISP-DM methodology is going to be used which is the most widely used form of data-mining model and the case study is no exception. The popularity of CRISP-DM comes from its various advantages for solving the existing problems related to data mining even across different industries thanks to its robustness. CRISP-DM breaks the process of data mining into six major phases as is shown in table 2 :

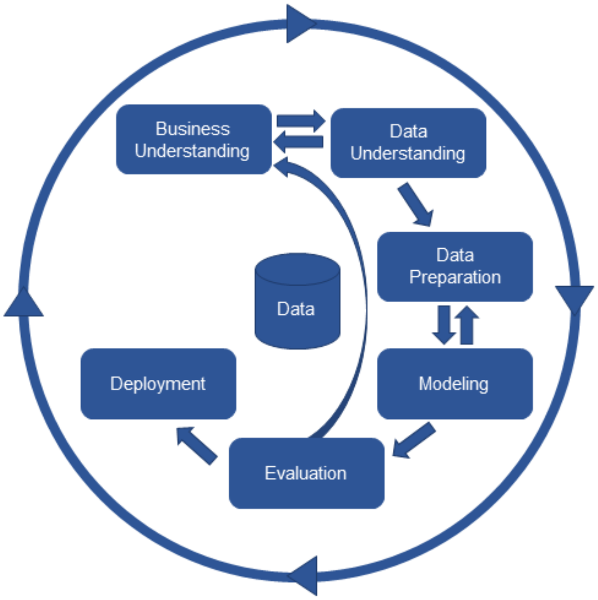


Fig. 2. CRISP-DM Methodology

A. Data understanding and preparation

To begin exploratory analysis, it could be informative to find out the mean and variance of the case types in the data, there are 3 possible case types : **confirmed, recovered, death** it is worth mention that for purposes of avoiding outliers and to much sparcity only the countries with 500+ confirmed cases where filtered, take in consideration that this filtering is performed on April 2nd of 2020.

Figure 3 shows a box plot for each type for all (500+ confirmed cases) countries as we can observe there still a wide range of values besides the filtering as shown in I

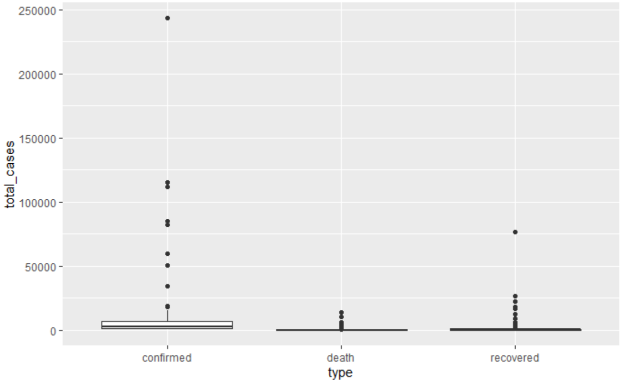


Fig. 3. Box plot of the different case types of COVID-19.

TABLE I
SUMMARY OF CONFIRMED CASES WORLDWIDE

Min:	505
1st Qu.:	1021
Median :	2454
Mean:	14706
3rd Qu.:	7154
Max.:	243453 .

Another exploratory question to ask about this dataset is if whether there is a significant difference in the means exhibited across different radiomic features or not. As an example, all 73 observations for 4 radiomic features (Figure 4) were taken and a one-way analysis of variance (one-way ANOVA) was performed taking each feature as a group. The summary statistics for the one-way ANOVA can be seen in Figure 5; it is observed that the F-value obtained in the test (162.3) is most likely not due to chance (p-value 2x10^{-16} for a random Fischer distribution with 3 degrees of freedom), and so it can be confirmed there is a significant difference between the mean values of these features. The latter can be clearly observed graphically in the box plots as the distributions of the *ClusterProm* and *HaralickCor* radiomics clearly look different. This is good as it implies that each radiomic feature probably would provide different kinds of information to a potential

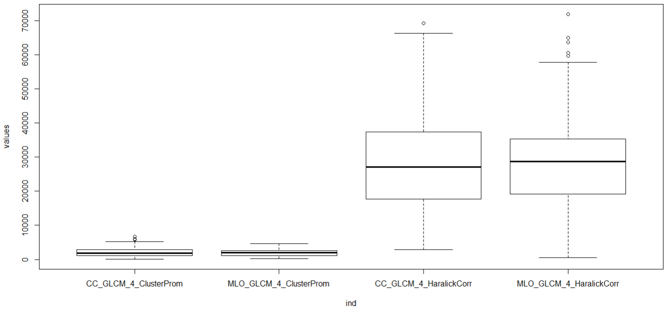
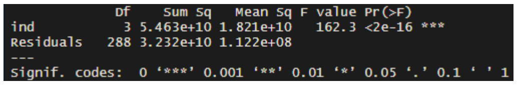


Fig. 4. Box plots for 4 different radiomic features in the San José dataset.



```
             Df    Sum Sq   Mean Sq F value Pr(>F)    
ind           3 5.463e+10 1.821e+10   162.3 <2e-16 ***
Residuals   288 3.232e+10 1.122e+08                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 5. One-way ANOVA summary statistics for 4 radiomic features of the San José dataset.

predictive model.

In addition to the radiomic features in the dataset there are 4 variables associated with risk scores. Although these risk features are only used to evaluate and compare against radiomic features we are interested to see how correlated this scores can be between themselves. Theoretically, each score assesses the following:

- *Oncotype*: recurrence risk score assesement, a continuous value based on the expression of selected breast cancer genes. It increases proportionally to the risk of recurrence.
- *Onco2*: in the dataset this is another version of the *Oncotype* score considering a different amount of genes. The idea is the same as with the *Oncotype* score though.
- *PAM50*: also known as the Prosigna breast cancer prognostic gene signature essay; it assigns a risk for distant recurrence of metastatic breast cancer derived lesions. The *Oncotype* score asseses primary site recurrence while this score focuses on metastasis.
- *AvGRISK*: a general risk for developing breast cancer. It is based off clinical attributes of the patients (not explicitly included in the dataset) as well as the age of the patients.

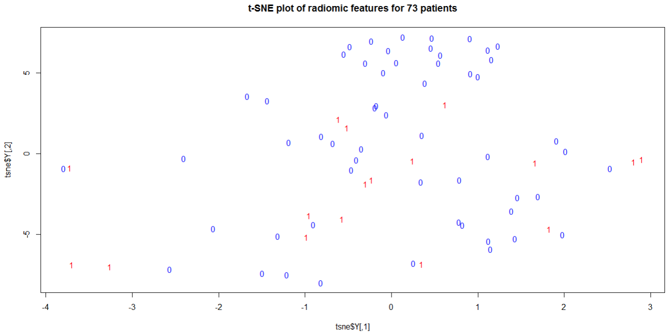


Fig. 6. Scatter plot of the radiomic features in the dataset reduced to 2 dimensions via t-SNE. Data points (patients) are depicted with their recurrence labels (0 = did not recurred in breast cancer, 1 = patient did recurred in breast cancer).

There is a risk variable (for breast cancer in general) associated with each one of the patients in the dataset; it could be interesting to see if there is a correlation between this variable and the PAM50 score of each patient (which denotes the likelihood that a particular breast cancer case will undergo metastasis).
Figure 7 presents a scatter plot between these two variables showing the values of Pearson and Spearman correlation at the top. Correlation is quite high (more than 0.85) with both methods indicating that maybe the average risk takes in its calculation elements from PAM50 or vice versa.

For exploring the dataset further in a graphical way, all the radiomic features were taken as an input for a t-distribution Stochastic Neighbor Embedding (t-SNE) dimensionality reduction process. The idea is to be able to visualize these radiomic features which are originally in a high dimensional space in a 2-d space via a scatter plot (Figure 6). The t-SNE method usually yields a visualization which keeps the general structure of the high dimensional data. From the plot we can identify that apparently those patients that didn't exhibit a breast cancer recurrence are more densely grouped by radiomic features as opposed to those that did exhibit recurrence. Although there can be variations of these patterns in higher dimensions of the same data, this intuition could be used later to for example, build a predictive model for no recurrence rather than for recurrence.

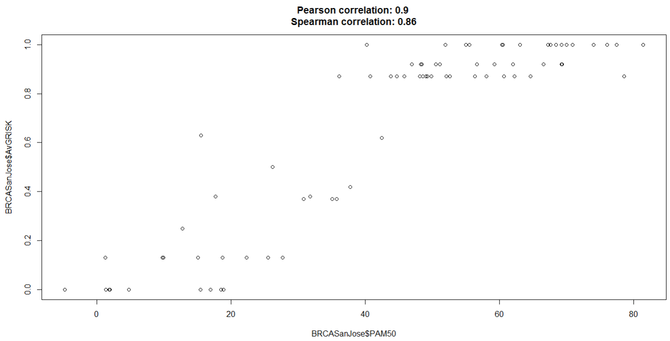


Fig. 7. Scatter plot of the average breast cancer risk vs. the PAM50 score for each patient in the San José dataset.

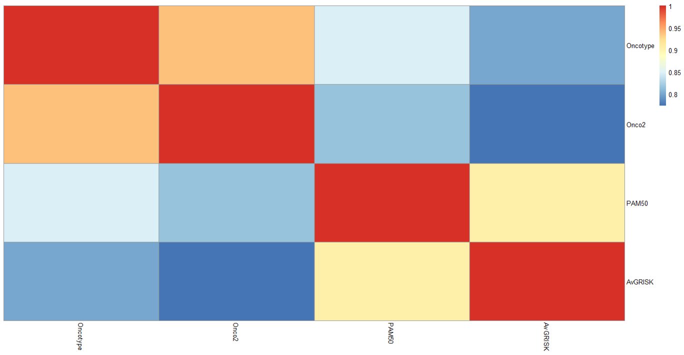


Fig. 8. Pearson correlations between the breast cancer risk score variables in the San José dataset.

In Figure 8, a Pearson correlation heatmap of these scores is presented. The correlation between all scores is in general strong (more than 0.8), however as expected the strongest correlations are observed between the *Oncotype* scores (as they probably consider overlapping gene panels in their calculations). *PAM50* and *AvGRISK* are also notably more correlated between them than to the *Oncotype* scores.

In order to fully understand the data better, a Shapiro-Wilk test of normality has been conducted for all numerical non-radiomic and radiomic features in the dataset (all radiomics comprise numeric variables). Before testing, the observations (rows) have been divided into 2 groups based on the *recurrence* variable (recurrence or no recurrence) and normality has been tested independently for each group and each variable.

As it can be seen in the previous table, for the non-radiomic features the only variable that has a p-value greater than 0.05 is the score *Onco2*. In the implementation of Shapiro-Wilk that was used, the null hypothesis corresponds to the

	Oncotype	Onco2	PAM50	AvGRISK	Age
rg.W.statistic	0.95	0.95	0.92	0.65	0.98
rg.p.value	0.39	0.40	0.16	0.00	0.99
nrg.W.statistic	0.94	0.96	0.96	0.79	0.96
nrg.p.value	0.00	0.08	0.04	0.00	0.04

input distribution being normal, therefore in this case for the significance level chosen, *Onco2* is the only feature that can't be rejected as non-normally distributed.

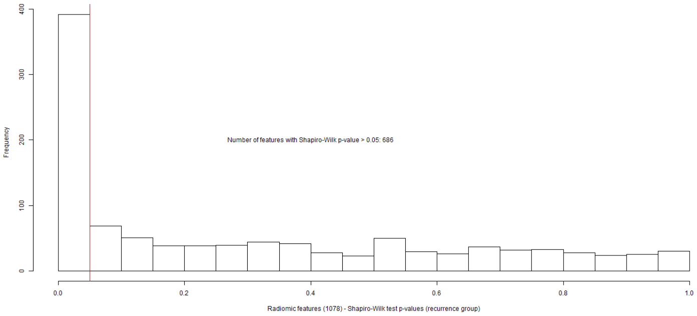


Fig. 9. Distribution of Shapiro-Wilk P-values obtained for all radiomic variables in the dataset for the recurrence group. Red line shows the significance level cutoff corresponding to a p-value of 0.05. All p-values lesser than the cutoff indicate rejection of the null hypothesis, the total number of variables that are normally distributed are indicated in text.

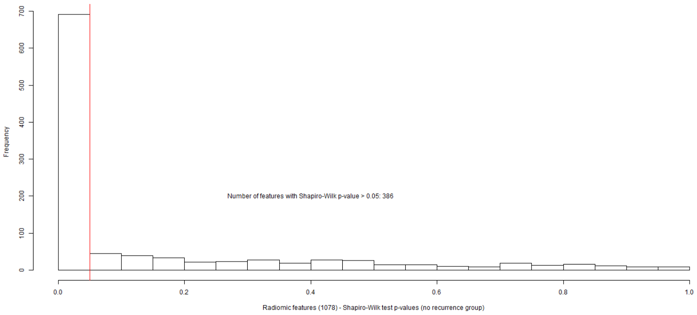


Fig. 10. Distribution of Shapiro-Wilk P-values obtained for all radiomic variables in the dataset for the no recurrence group. Red line shows the significance level cutoff corresponding to a p-value of 0.05. All p-values lesser than the cutoff indicate rejection of the null hypothesis, the total number of variables that are normally distributed are indicated in text.

In figures 15 and 10, the results of the Shapìro-Wilk test are presented for the radiomic features of each group. The total number of features that can be considered as normally distributed from the test results in both groups (i.e. The intersection set) is 334 (1 non-radiomic and 333 radiomics); this is shown graphically for the radiomic features in Figure 11.

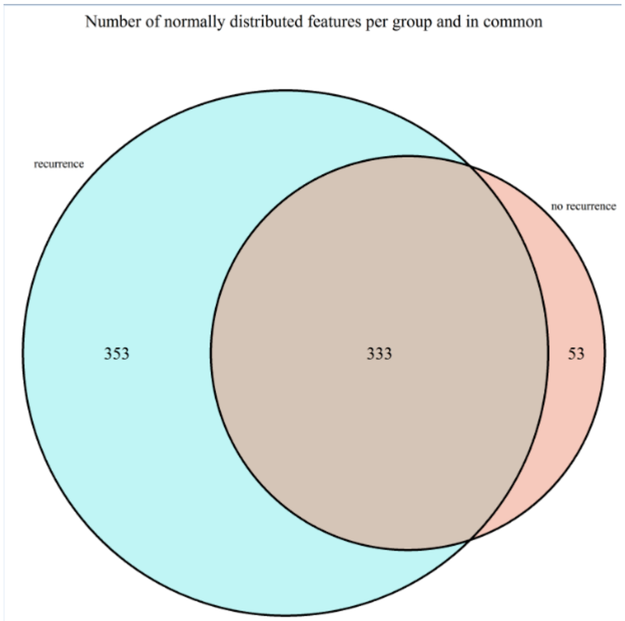

Number of normally distributed features per group and in common
recurrence
no recurrence
353
333
53

Fig. 11. Number of normally distributed radiomics per group and the intersection between them.

The latter results can be used to conduct a one-way ANOVA (O-WANOVA) test between the recurrence and no recurrence groups for each normally distributed variable; however, O-WANOVA has another assumption additional to normality of the groups to be testes, that is that the groups also have equal variance. Before performing O-WANOVA, each of the 334 normally distributed features were tested via a Levene test of variances. In this test, a groups vector (in this case a binary vector of recurrence and no recurrence) is given as an input as well as a numerical vector with the values for each observation addressing a variable of interest.The null hypothesis is that the variances of each group for their respective values in the numerical vector are not significantly different. Levene's test was performed for each of the 334 variables found to be normally distributed in both groups via the Shapiro-Wilk test.

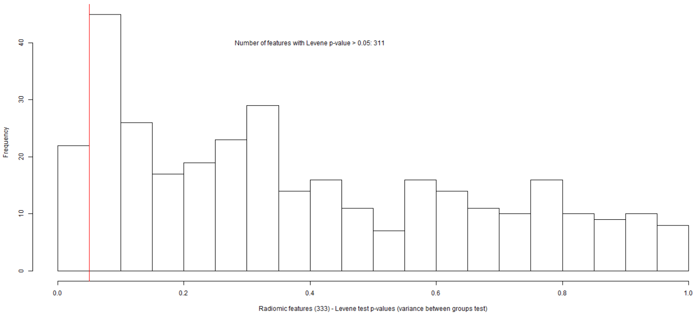


Fig. 12. Distribution of Levene P-values obtained for all radiomic variables in the dataset (both groups need to be tested at the same time). Red line shows the significance level cutoff corresponding to a p-value of 0.05. All p-values lesser than the cutoff indicate rejection of the null hypothesis, the total number of variables that exhibit equal variances are indicated in text.

In the case of the lone non-radiomic feature that was tested via Levene's test, an statistic of 1.68 with a p-value of 0.19 was obtained meaning that the null hypotheses may be accepted that the *Onco2* score has equal variances for both groups. In the case of the radiomic features, figure 12 shows the distribution of p-values obtained for all 333 features. At the end, 312 total features were identified with equal variances for both groups; these features are eligible for A-WANOVA analysis as they theoretically meet the assumptions of this test.

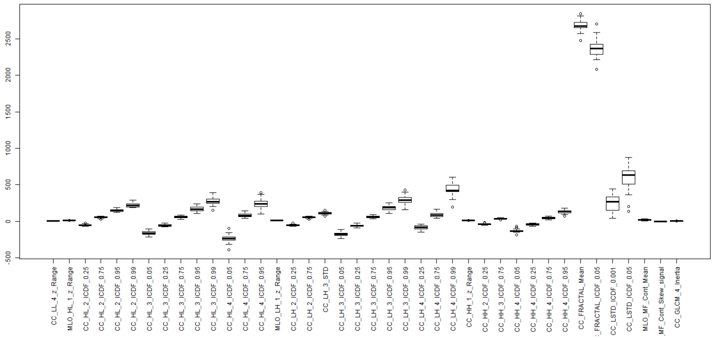


Fig. 13. Box plots of radiomics with different means between groups according to O-WANOVA test with significance level of 95% (recurrence group values shown).

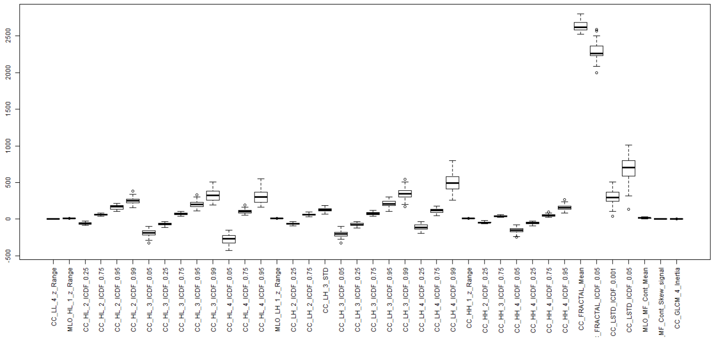


Fig. 14. Box plots of radiomics with different means between groups according to O-WANOVA test with significance level of 95% (no recurrence group values shown).

O-WANOVA for the recurrence and no-recurrence groups was performed for all variables (one test for each variable) found as eligible after the Shapiro-Wilk and Levene's test filters applied in the previous steps. From the O-WANOVA tests, 40 radiomic variable exhibited p-values lesser than or equal to 0.05. In the case of O-WANOVA, the null hypothesis is that the means of both groups for the given numerical variable are equal, and hence rejecting this hypothesis means recognizing the means as significantly different between groups. Figures 13 and 14 show box plots for the 40 features mentioned earlier (recurrence and no recurrence groups respectively); these are the features, which all happen to be radiomic, that have significantly different means between the groups of interest that were defined based on the *recurrence* variable of the dataset.

B. Modeling for the data understanding

One of the applications of statistics is to extract inferences in populations from the study of samples. This process is called Inferential Statistics and its studies aim to deduce (infer) properties or characteristics of a population from a representative sample [4].

One of this tools is correlation and determines whether causal or not, between two random variables or bivariate data are dependent. In the broadest sense correlation is any statistical association, though it commonly refers to the degree to which a pair of variables are linearly related. As we can see in figure 15 a scatter plot between these two variables showing the values of Pearson correlation and denoting a positive relationship (0.87) .

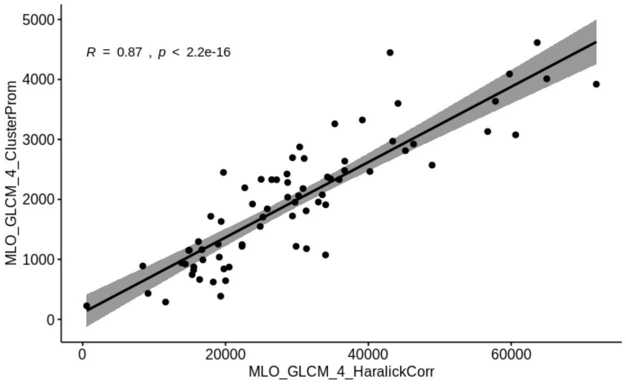


Fig. 15. Scatter plot and relationship by Pearson correlation.

The business goal, or the problem we are trying to solve through our data analysis is whether a patient is prone to have recurring health issues related to breast cancer through the analysis of medical images. Therefore, normally our response variable would be whether a patient presented a recurrence or not. Since this is a binary variable, it is not fit for the experiment we want to look at on this assignment.

Therefore, we made a Pearson correlation analysis in order to pick a couple variables that would indeed make a good linear regression model from which we can extract meaningful insights as seen on Figure 16. The variables we chose were:

Numeric response variable – S_CC_HH_4_RSxy

Numeric predictor variable - S_CC_LH_2_RSxy

Both are Sums transformation of the cranio-caudal (CC) view.

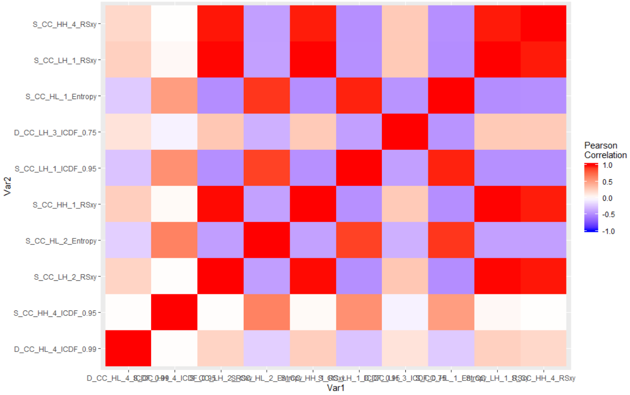


Fig. 16. Heatmap of random sample variables.

The strong correlation between these two variables can be confirmed visually by drawing a scatterplot. The positive relation is shown on Figure 16.

Once the Regression equation is performed over these two variables we obtain the following summary sown in table II.

TABLE II
SUMMARY OF LINEAR MODEL OF TWO CHOSEN VARIABLES

Min	1Q	Median	3Q	MAX
-104.245	*-19.673*	*-4.623*	*12.515*	*129.232*

By definition, the residuals have a mean zero. Hence, the median or second quartile should also not be far from zero. In our case, the median is -4.623, which is close enough to zero. The first quartile and third quartile have roughly the same absolute values, which means our data is pretty well balanced. This holds for the minimum and maximum as well, both absolute values are much larger than our interquartile values, but they have roughly the same absolute value.

Beta value. The estimate is the weight given to a variable. Our numerical variable has a coefficient close to one, which is another way to show the strong significance of that value on the model. Standard error. The standard error (SE) is related to the accuracy of the beta coefficients. It reflects how the coefficient varies under repeated sampling. It represents the average distance of the values from the regression line, which in the case of our variable is fairly small (0.02436), which means the values are fairly close to the line of the regression model.

T-Value. The t-statistic tests whether or not there is a statistically significant relationship between a given predictor and the outcome variable. The higher the t-statistic, the better. In our case, a value of 38.171 is considered high enough, much higher than the interceptor with -0.655.

P-Value. The p-value, like the t-value, also tests whether or not there is a statistically significant relationship between a given predictor and the outcome variable. The lower the p-value, the better. In our case, the p-value is very small at less than 2e-16 and given the code '***' which is the most significant awarded by the summary function on R.

From the beta coefficient section, we can conclude that the predictor and the outcome variables have a significant association, since both the p-values and the t-values are significant themselves. Therefore, we can reject the null hypothesis and accept the alternative hypothesis.

A brief of these statistics is displayed in the following table III:

TABLE III
SUMMARY OF BETA COEFFICIENTS: BETA VALUE, STANDARD ERROR, T-VALUE, P-VALUE STATISTICS

	Estimate Std.	Error	t value	Pr(>—t—)
(Intercept)	*-5.41386*	*8.26196*	*-0.655*	*0.515*
S_CC_LH_2_RSxy	0.92973	0.02436	38.171	<2e-16***

Now that we have defined S_CC_LH_2_RSxy as a variable heavily associated with the predicted value, we can take a look at how well the model fits the data. In our case, turns out the model fits the data pretty well, having a small residual standard error, a high value of r-squared and a high F-value. Residual Standard Error (RSE). We can use the RSE to get the prediction error rate by dividing it by the average value. This value should be as small as possible and for our variable it sits at 2.89 which is fairly small. Degrees of Freedom. For our model we have 62 degrees of freedom, which can be considered quite high. This supports the evidence of rejecting the null hypothesis as stated in the previous section. R-Squared. This value ranges from zero to one, where a high value means most of the information can be explained by the model, which is desirable. At 0.9592, most of out data can be explained by the model, therefore the linear regression model is a good fit for this data. Adjusted R-Squared. The adjusted R2 is very similar to the simple R2, which iat 0.9582 is also a desirable value for most of our data can be explained by the model. F-Statistics. At 1457, our F-Statistic is very high which, like the t-test, indicates the model is very significant. P-Value. Having a very small p-value like in this model is desirable, since the smaller the value the more significant the model is as shown in table IV.

TABLE IV
SUMMARY RSE RSS AND F-STATISTIC

Residual standard error:	*38.99 on 62 degrees of freedom*
Multiple R-squared:	*0.9592, Adjusted R-squared: 0.9585*
F-statistic:	*1457 on 1 and 62 DF, p-value: <2.2e-16*

Is there a relationship between the response and predictor variable?

Yes, there is a strong relationship between the response and the predictor variables as shown by the linear regression model 17 . Both the t-test and the p-value suggested that there's a strong association between the variables and that the model holds a high significance.

How does this relationship contribute to understanding your business case and the business problem understanding?

It is quite useful to see some behaviours, but in order to see the big picture over the real trends over the data, more sophisticated methods have to be performed, for example we can construct recurrence models using different ML approaches such as Linear Discriminant Analysis (LDA), L1 Penalized Logisit Models (LASSO 1se and LASSO min), Bootstrapped logistic models (BSWiMS), Random Forest(RF), Regression and Partition Trees (RPART), Support Vector Machines (SVM) or K-nearest neighbors (KNN).

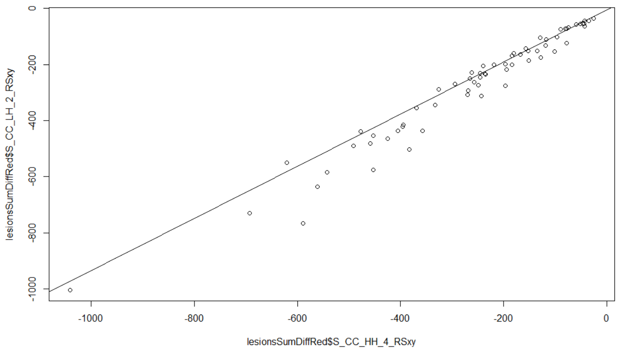


Fig. 17. Scatterplot of the correlation between the variables.

III. Results

A. Modeling and Evaluation

After having a clear understanding of the features, their relationships and characteristics, we proceeded to model and evaluate 8 different models, these models were trained using a 1000 cross-validation (cv) using 95% training set and 5% holdout employing FeatuRE Selection Algorithms for Computer Aided Diagnosis (FRESA.CAD) package's function *randomCV()* in R, the 8 models were trained as well its performance were analyzed and plotted using the *predictionStatsbinary()* function.

The cross-validation results imply that radiomic models can predict between 64% to 76% accuracy recurrence events, On the other hand it is observed that Linear Discriminant Analisis (LDA) have ROC AUC of $73\,\% \pm 10\,\%$, followed by K-nearest neighbors (KNN) with $70\,\% \pm 12\,\%$, for Bootstrapped logistic models (BSWiMS) $67\,\% \pm 12\,\%$, for L1 Penalized Logisit Models (LASSO 1se and LASSO min) $64\,\% \pm 10\,\%$, and $61\,\% \pm 12\,\%$ respectively, Support Vector Machines (SVM) with radial kernel $60\,\% \pm 12\,\%$, Random Forest(RF) $55\,\% \pm 11\,\%$, and finally, Regression and Partition Trees (RPART) $46\,\% \pm 11\,\%$

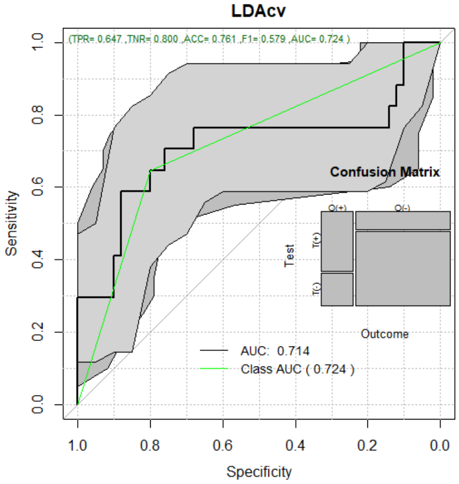


Fig. 18. Linear Discriminant Analysis ROC Curve. Model based on Radiomic Features

Figure 18 shows the ROC analysis of the LDA models generated using: radiomic features from the control This plot shows that the recurrence signature is present on the clinical features as well as the lesion features. It is also clear that the LDA radiomic based model was superior to all the clinical-data-based model.

Figure 19 and 20 compares the test accuracy and the balanced error results of the eighth ML models and their ensemble trained using only radiomic features. The Balanced error of the LDA model was statically better that the SVM, RF, and RPART models. The bottom plot of Figure 20 shows the Sensitivity.

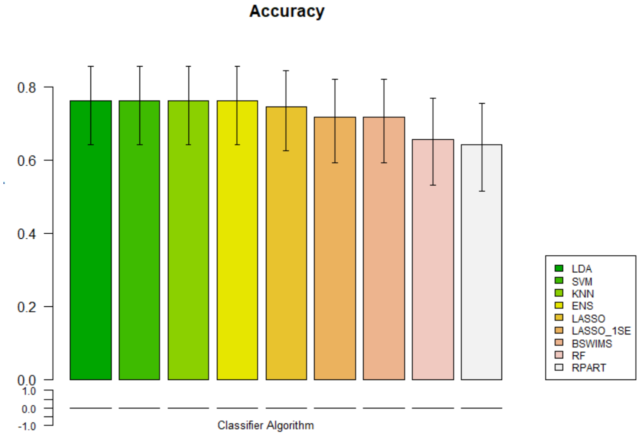


Fig. 19. Accuracy test results of the Lesion-Only Radiomic model

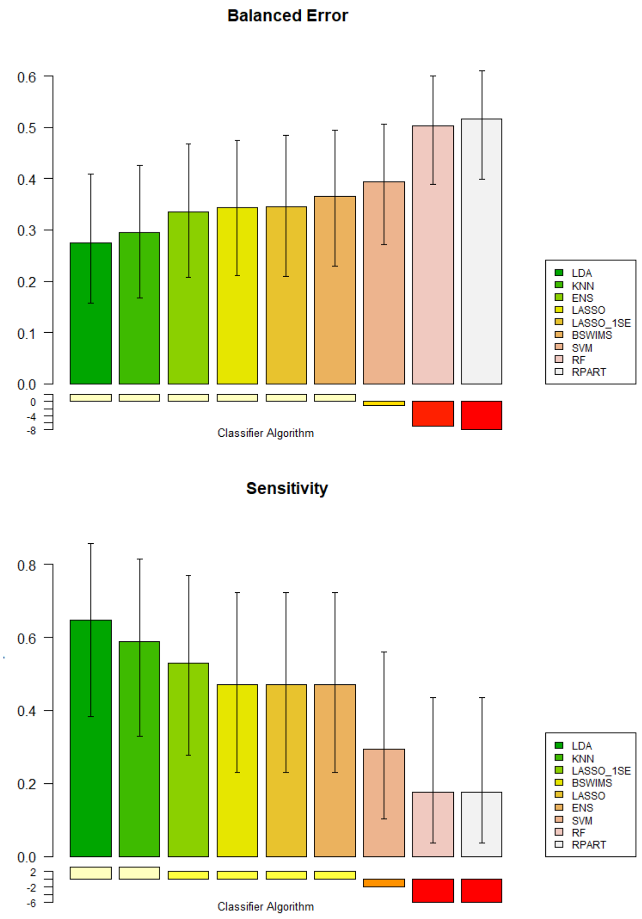


Fig. 20. Balanced Error and Sensitivity of Radiomic models

Figure 21 shows the LDA features required to get the observed LDA performance.The feature analysis indicated that the variation of the High-High (HH) sub-band of the Wavelets transform was present on almost 100% of the 1000 models. In the second place, the range of values present on the Fractal transform also was a strong indication of recurrence.

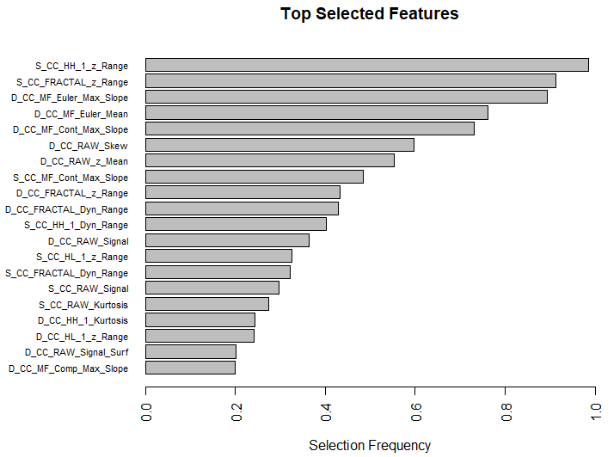


Fig. 21. Feature Selection frequency of top 20 Features of the LDA Radiomic Model

IV. Discussion

Discussion (How good is your model?)
Research efforts on BRCA have been directed toward better treatment, preventive strategies, and early detection, Although mammography has been the mainstay of early detection, its limitations are well recognized and the search for more effective technologies for this purpose has been receiving increased attention [1].

Molecular diagnosis is an approach that has an outstanding accuracy, but terms of affordability are unattainable for some population strata [2] and Prior studies such as [6] have noted the importance of radiomics signatures for predicting the risk of BRCA recurrence using logistic regression with results analogous to those presented in this work ,in general there exist a lack of explanation about the behavior of different algorithms when trying to model this recurrence data.

On the other hand if there are studies [5] [10] that show the behavior of different models to predict recurrence, but the scopes are lung and cervical cancer. Very little was found in the literature on the association between radiomic features extracted from mammograms to predict recurrence. To our knowledge this is the work that compares more metrics and more ML classification algorithms to explain the recurrence phenomena.

V. Conclusion

Conclusion (What you learned from this exercise)
we have applied the CRISP-DM methodology from top to bottom, starting from a simple Exploratory Data Analysis (EDA), in order to understand the characteristics of the features and its relations among them, using methods such ANOVA, correlation test (Pearson and Spearman), modeling data for understanding using t-SNE to visualize groups and trends contained in the dataset, and fitted and evaluating the models using different machine learning models using Cross-Validation and ROC curves.

The results presented in this paper are very encouraging.

First, it demonstrated that mammography alone has can capture cancer phenotype associated with the probability of recurrence with a sensitivity of 65% and specificity of 80%. Second, the radiomic model was better than the clinical only model: 0.76% vs. 65%. This second result implies that oncologists may have a better tool for predicting the cancer recurrence risk that outperforms current risk assessment tools. These results support our previous work that discovered that radiomic signatures are associated with molecular signatures of cancer recurrence namely Oncotype and PAM50.

References

[1] National Research Council et al. *Mammography and beyond: developing technologies for the early detection of breast cancer*. National Academies Press, 2001.

[2] RI Cutress, A McDowell, FG Gabriel, J Gill, MJ Jeffrey, A Agrawal, M Wise, J Raftery, IA Cree, and C Yiangou. Observational and cost analysis of the implementation of breast cancer sentinel node intraoperative molecular diagnosis. *Journal of clinical pathology*, 63(6):522–529, 2010.

[3] Scott Dylan. The covid-19 risks for different age groups, explained, 2020.

[4] María José Rubio Hurtado and Vanesa Berlanga Silvente. Cómo aplicar las pruebas paramétricas bivariadas t de student y anova en spss. caso práctico. *Reire*, 5(2):83–100, 2012.

[5] Elizabeth Huynh, Thibaud P Coroller, Vivek Narayan, Vishesh Agrawal, John Romano, Idalid Franco, Chintan Parmar, Ying Hou, Raymond H Mak, and Hugo JWL Aerts. Associations of radiomic data extracted from static and respiratory-gated ct scans with disease recurrence in lung cancer patients treated with sbrt. *PloS one*, 12(1), 2017.

[6] Hui Li, Yitan Zhu, Elizabeth S Burnside, Karen Drukker, Katherine A Hoadley, Cheng Fan, Suzanne D Conzen, Gary J Whitman, Elizabeth J Sutton, Jose M Net, et al. Mr imaging radiomics signatures for predicting the risk of breast cancer recurrence as given by research versions of mammaprint, oncotype dx, and pam50 gene assays. *Radiology*, 281(2):382–391, 2016.

[7] Koerth Maggie. How coronavirus tests actually work, 2020.

[8] Koerth Maggie, Bronner Laura, and Mithani Jasmine. Why it's so freaking hard to make a good covid-19 model, 2020.

[9] Nidhi Prakash. Doctors and nurses say more people are dying of covid-19 in the us than we know, 2020.

[10] Sylvain Reuzé, Fanny Orlhac, Cyrus Chargari, Christophe Nioche, Elaine Limkin, François Riet, Alexandre Escande, Christine Haie-Meder, Laurent Dercle, Sébastien Gouy, et al. Prediction of cervical cancer recurrence using textural features extracted from 18f-fdg pet images acquired with different scanners. *Oncotarget*, 8(26):43169, 2017.