“Keys to Survival and Recovery in Stage III or IV Lung Cancer with Pembrolizumab”

David Blasco Añó
José Manuel Esteso Morcillo
María Gómez Belenguer
Ainhoa Prado Alonso
María Verdú Gómez

Project III Final Report

Universitat Politécnica de València
Escuela Técnica Superior de Ingeniería Informática
Grado en Ciencia de Datos
València, Spain
2023-2024

INDEX

FIGURE INDEX

TABLE INDEX

INTRODUCTION

Cancer is the second most common cause of death worldwide according to the WHO (World Health Organization). It accounts for approximately 9.6 million deaths, or one out of six deaths, registered in 2018 [1]. More specifically, the leading type of cancer nowadays with 2.09 million diagnosed cases per year is the one this team is going to study: lung cancer. Moreover, this type is the one that causes the greatest number of deaths, with globally 1.76 deaths per year [2]. There are two main types of cancer: Non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) [3]. As the database has information of NSCLC's patients, the team has focused their research process on this type, which is also the most common one.

Typically, when diagnosing lung cancer, the physician orders tests to determine if there is a change in the patient's genes, a process known as genetic testing or molecular profiling. These tests help identify specific genetic mutations or alterations within the cancer cells, providing crucial information about the tumor's characteristics and behavior. By understanding the genetic makeup of the cancer, doctors can tailor treatment plans to target the specific vulnerabilities of the tumor, leading to more personalized and effective therapies. Additionally, genetic testing can also help predict the likelihood of treatment response and potential side effects, guiding decisions regarding the most suitable treatment options for each individual patient [4].

The research focuses on the application of the drug Pembrolizumab. Pembrolizumab is an immunotherapy drug, which in this case has been specifically administered to patients with non-small cell lung cancer already in advanced stages and with a PD-L1 protein expression greater than or equal to 50%. This medication binds to the PD-1 protein and helps immune cells destroy cancer cells. In 2023, a study was conducted demonstrating that Pembrolizumab (along with other treatments and surgery) improved survival by reducing the risk of events that may have a negative impact on the patient, compared to other treatments [5].

METHODOLOGY

To address this project, we have followed the CRISP-DM [6] methodology, which provides a structured approach to data mining. We organized our tasks using a Gantt chart Figure I and Figure II. in Appendix to ensure timely and efficient progress. Additionally, we familiarize ourselves with the Sustainable Development Goals (SDGs) [7] and determined that our work aligns with several key goals:

- SDG 3 (Good Health and Well-Being): Our project focuses on improving the personalization and effectiveness of lung cancer treatments, contributing to better health outcomes.
- SDG 9 (Industry, Innovation, and Infrastructure): We aim to drive medical innovation through the development of advanced predictive models, enhancing the infrastructure for cancer treatment.
- SDG 17 (Partnerships for the Goals): Our approach relies on collaboration between researchers, clinicians, and technologists, fostering partnerships to provide practical solutions in medical research.

By integrating these SDGs into our methodology, we aim to ensure that our work not only advances scientific knowledge but also contributes to broader societal goals.

JUSTIFICATION

It costs €4600 to diagnose lung cancer in a patient and treat it between €17,000 and €34,000. Additionally, 36.5% of lung cancer patients spend €3000 on pharmacy and parapharmacy, 61% between €50 and €750, and around 2.3% €5000. A blood test costs about €40. A chemotherapy session costs around €770, and the most common lung cancer operation (lobectomy) costs between €4000 and €5400, and lung transplant €70,000. A vial of Pembrolizumab costs €4225. All these data refer to how expensive it is to treat advanced-stage lung cancer. [8] [9] [10]

On the other hand, lung cancer treatments, and cancer treatments in general, have quite high toxicity. If drug or treatment toxicity can be prevented, good symptomatic control and a better quality of life can be achieved.

To sum up, the main value of this descriptive and predictive study is to help oncologists personalize treatments by administering the drug Pembrolizumab to those for whom it is truly effective, in order to anticipate and prevent treatment toxicities and tumor progressions, as well as to save costs.

OBJECTIVES

A. General objective

Predict the most important variables of the first disease assessment and treatment response, and conduct a survival analysis.

B. Specific objectives

Here are the objectives, concisely written for our project:

1. Analyze Missing Values: Assess and impute missing values to ensure data completeness.
2. Exploratory Data Analysis (EDA): Perform EDA to gain initial insights and identify patterns in the data and conduct a PCA to explore relationships between clinical variables and target responses.
3. Partial Least Squares (PLS) and PLS-Discriminant Analysis (PLS-DA): Implement PLS and PLS-DA to predict target variables and study data distributions.
4. Random Forest Model: Build a random forest model to enhance prediction accuracy.
5. XGBoost Model:Develop an XGBoost model for robust predictive performance.
6. Survival Analysis:Analyze patient survival rates and identify key influencing factors.
7. Model Comparison: Compare all models to determine the most effective one for predicting target variables and improving patient outcomes.

VARIABLE DESCRIPTION AND SELECTION

This study focuses on patients between 46 and 82 years diagnosed with stage III to IV NSCLC treated with Pembrolizumab as first-line therapy. With a sample of thirty-six patients treated from April 2018 to April 2021.

We initially faced a dataset containing 271 variables and 36 observations. This dataset posed a considerable challenge due to its dimensions, making work difficult and requiring careful manipulation. Due to the fact that 2 patients were missing the value in the predictable classes, we had to remove them from the database. Additionally, we also eliminated variables with more than 23% missing data and imputed the missing values for the remaining variables to achieve a better analysis. We can classificate the variables from the resulting database into 10 categories according to their meaning:

1. Physical Characteristics of the Patient:This includes all variables related to the physical state of the person, such as age, BMI, or ECOG.
2. Patient Habits:This category only includes the variable that reflects the patient's tobacco exposure.
3. Other Diseases of the Patient:This indicates the presence or absence of conditions such as heart disease, diabetes, or other neurodegenerative diseases.
4. PD-L1 Protein Expression: A single variable confirming that treatment with Pembrolizumab is recommended for the patients in the database.
5. Tumor Characteristics: This includes all characteristics related to the size, stage, or involvement of the tumor.
6. Cycles and Toxicity: This includes variables referring to the number of cycles or the toxicity presented by the patient.
7. Analytical Results: This includes various values of immune cells before starting treatment, during the first and second cycles of treatment, and after the first radiological evaluation.
8. Survival Indicators: This category includes variables related to patient survival.
9. Response Variables:This includes the variables "first radiological evaluation" and "best response," which are the variables to be predicted.
10. Dates: This includes important dates for each patient (date of birth, date of diagnosis, etc.).

The result of the first radiological evaluation (prim_eval_num_ok) reflects the patient's disease progression and is divided into four categories: RP (Partial Response), PS (Pseudo-progression), EE (Stable Disease), and PE (Disease Progression). These categories have been rearranged on an ordinal scale, where RP (1) is considered better than PS (2) and EE (2), and these two are superior to PE (3). Therefore, this result can be treated as an ordinal variable.

On the other hand, the outcome of the best treatment response (mejor_resp_num_ok) indicates the treatment's effectiveness in the patient and is also divided into four categories: RC (Complete Response), RP (Partial Response), EE (Stable Disease), and PE (Disease Progression). These categories have been similarly encoded on an ordinal scale, where 0 corresponds to RC, followed by 1 for RP, 2 for EE, and 3 for PE. Therefore, this variable can also be treated as ordinal.

DATA PROCESSING AND UNDERSTANDING

Missing value imputation was carried out using the *mice* [11] library. Categorical variables were transformed into factors to allow joint imputation with numeric variables using mean methods, and for factors with more than two categories, the cart method was employed.

Finally, to optimize the selection of variables to be included in the predictive model and to identify “individuos anómalos”, a principal component analysis (PCA) was conducted [12].

We performed a logarithmic transformation on the variables representing the SII indicator at any stage of the treatment, as it could help reduce the anomaly of the individuals. We used the T2-Hotelling, as we can see in Figure III. measure to identify which individuals are anomalous. After performing this transformation, no individual stands out above the 99% limit. Individual 25 is above the 95% limit, considered a severe anomaly, and therefore will not be excluded from subsequent models.

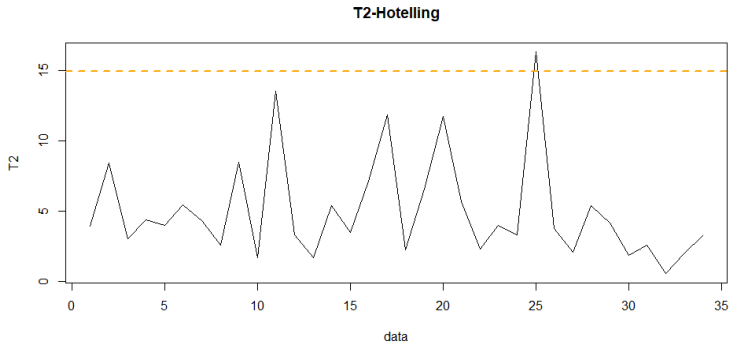

Figure III. T-2 Hotelling

This analysis identified the most relevant variables for predicting "first_eval" and "best_response." Summarizing the variables with the greatest contribution to each predictor variable, the results obtained are as follows::

- "first_eval": The variables like SII_1eval, NLR_1eval, and PLR_1eval show the strongest positive contributions, as can be seen in Figure IV. in Appendix , all above 0.55. This implies a strong direct correlation with "1st eval," suggesting these inflammation-related markers (Systemic Immune-Inflammation Index, Neutrophil to Lymphocyte Ratio, and Platelet to Lymphocyte Ratio from the first evaluation) are crucial in the assessment or outcome measured by "1st eval". Variables like IMC, Linf_2C, and more extreme cases such as Hb_1eval with contributions ranging from -0.21 to -0.45, represent the most significant inverse relationships in the dataset. The presence of these variables in significant amounts strongly correlates with lower "1st eval" scores, pointing to factors like nutritional status, overall health, and other metabolic indicators as critical inversely related components in the assessment.
- "best_response": SII_1eval (0.551), NLR_1eval (0.528), and SII_2C (0.523) are the top contributors Figure V. in Appendix. These are indices that measure systemic inflammation (Systemic Immune-Inflammation Index and Neutrophil to Lymphocyte Ratio), which could be critical markers in scenarios where immune response plays a significant role in the outcome, such as in certain diseases or treatments. The variables Hb_1C (Hemoglobin level at first cycle), Alb_2C (Albumin level at second cycle), down to N_ciclos (Number of cycles, -0.544) show the strongest negative impact. The inclusion of hemoglobin and albumin suggests that worse outcomes associated with "mejor respuesta" are significantly influenced by these variables, potentially indicating severe underlying conditions or responses to treatment.

RESULTS

A. First radiological evaluation prediction

To predict the variable "first_eval", which assigns a category to each patient according to the progression of the disease, we have conducted 3 predictive models (PLS and PLS-DA, Random Forest and XGboost). For this purpose, we have selected a subset of variables since it does not make sense to use variables from analytics after the first evaluation nor those related to survival.

Therefore, the input variables will be as follows: physical characteristics, habits, other diseases, expression of the PD-L1 protein, tumor characteristics, cycles and toxicity, results of analytics up to the first cycle of treatment. The expected output will be a numerical scale ranging from 1 to 3 indicating the level of the patient's disease. The closer the number is to 1, the disease will appear to be progressing positively, and the closer it is to 3, the disease will continue spreading.

PLS & PLS-DA:

The first model we propose is a PLS [13] and PLS-DA [14] model. These techniques are particularly useful for analyzing data with multiple predictor variables that may be highly correlated or for cases where the number of observations is less than the number of variables. The PLS model is a regression technique that not only seeks to maximize the covariance between predictor variables and response variables but also effectively handles multicollinearity among the variables. On the other hand, PLS-DA is a variant of traditional PLS adapted for classification. It utilizes PLS capabilities for feature extraction and constructs a model that can discriminate between predefined classes based on predictor variables.

The models are built using centered and scaled data to ensure that all variables contribute equally to the model, without the differences in feature scales negatively affecting model performance. The low number of observations requires us to train the PLS and PLS-DA models using the Leave One Out (LOO) technique [15].

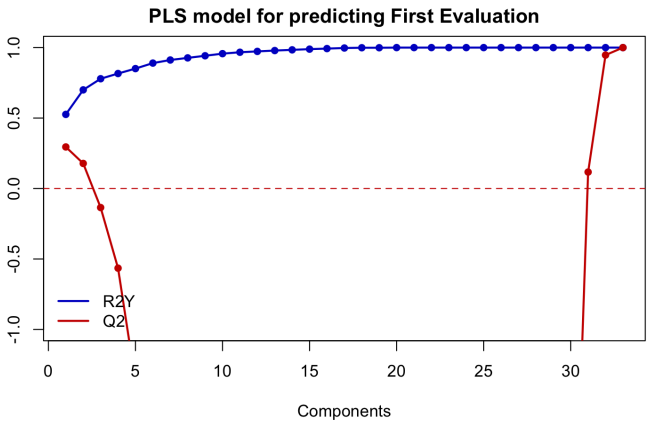

Figure VI. PLS First Eval

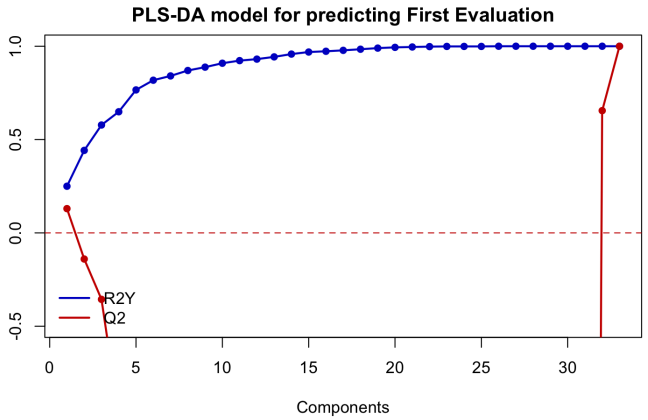

Figure VII. PLS-DA First Eval

When selecting the principal components, we encountered a poor scenario: most principal components have a negative Q2, as can be seen in Figure VI. and Figure VII. With the help of graphs representing the number of principal components along with their corresponding R2 and Q2 values, we ultimately decided on 2 principal components for the PLS model and 2 principal components for the PLS-DA model.

The reasons for choosing these numbers of components are to avoid model overfitting by selecting a higher number of components, even if they have higher Q2, and because the R2 value increases considerably when we choose this number of principal components. Given the few observations, we understand that this model will not generalize well to new observations, but the predictive capability of the model is somewhat important to us. In the case of PLS-DA, we opted for 2 principal components because we already assume model

overfitting with the 34 observations in our database. However, by choosing this number of principal components, we can explore the space of scores and loadings.

From the score plots of Figure VIII. and Figure IX., we can see that for both models, the first principal component separates individuals classified as class 1 and class 3 quite well, i.e., it differentiates between those for whom the disease progresses positively and those for whom it progresses negatively.

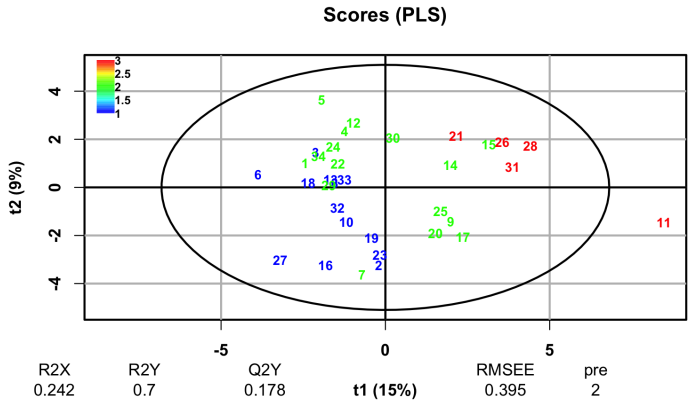

Figure VIII. Score PLS

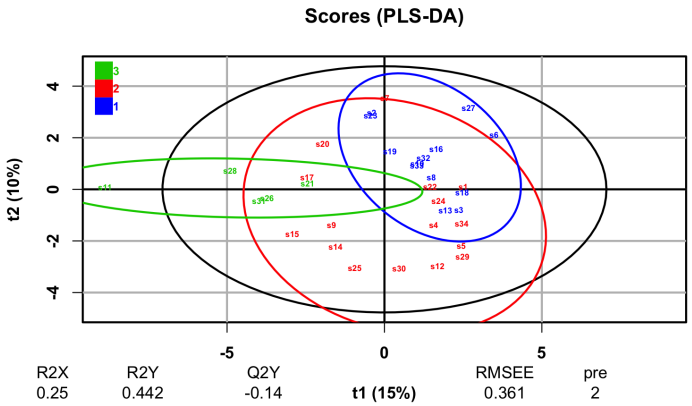

Figure IX. Score PLS-DA

In the loading plots of Figure X. and Figure XI., we can see how the regressor variables relate to the response variable. Patients in class 1 are associated with a lower number of treatment cycles, adenocarcinomas, and lower toxicity. Conversely, we can observe that individuals with worse disease progression (class 3) have higher SII (Systemic Immune-Inflammation Index) values, which is consistent with studies suggesting that patients with elevated SII may have a less effective response to therapy. Similarly, the NLR

(neutrophil/lymphocyte ratio) and PLR (platelet/lymphocyte ratio) variables are related: the higher the value, the worse the tumor progression.

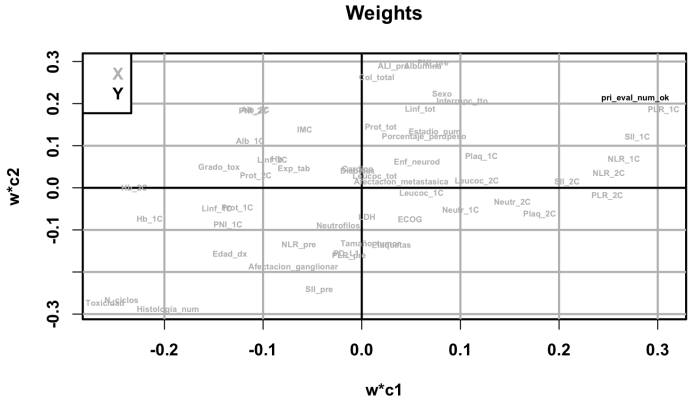

Figure X. Weights PLS

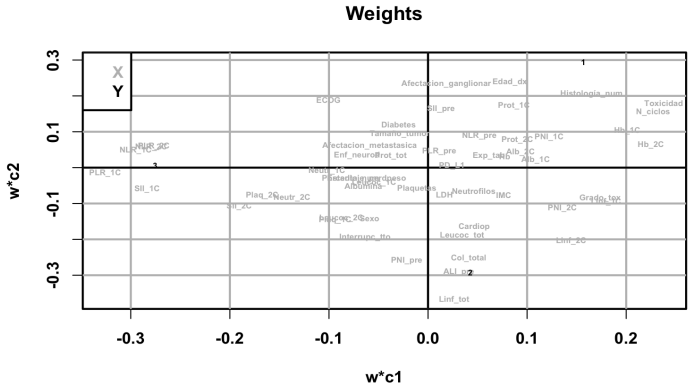

Figure XI. Weights PLS-DA

The plots of the variables with the highest VIP [16] reflect the same insights as the loading plots, as can be seen in Figure XII. to Figure XV. in Appendix.

Finally, we evaluated the PLS and PLS-DA models using appropriate evaluation measures for regression and classification models. Unfortunately, for the evaluation of the models, we cannot use data other than those with which we trained the model, as a train-test split would only worsen the models.

For the regression model, we used the RMSE [17], which expresses the difference between the values predicted by our model and the observed values, and the CVrmse. The RMSE we obtained is 0.3771, meaning that on average, the predicted value differs by +/-

0.3771 from the actual value. This is expected since the observed values are integers (1, 2, and 3) and the predicted values are real numbers between 1 and 3. The CVrmse we obtained is 0.2137, which means that the model's predictions have a 21% error relative to the scale of the original data. With this value, we can say that the model has a moderate to acceptable performance.

In the following graph, as we can see in Figure XVI., we can observe that class 1 and class 3 can be differentiated by the model, while class 2 is more difficult to predict due to overlap with both classes. This could suggest that the class assignment is somewhat ambiguous, which is understandable because, in medical contexts, there are many other factors that cannot be captured and that affect disease progression.

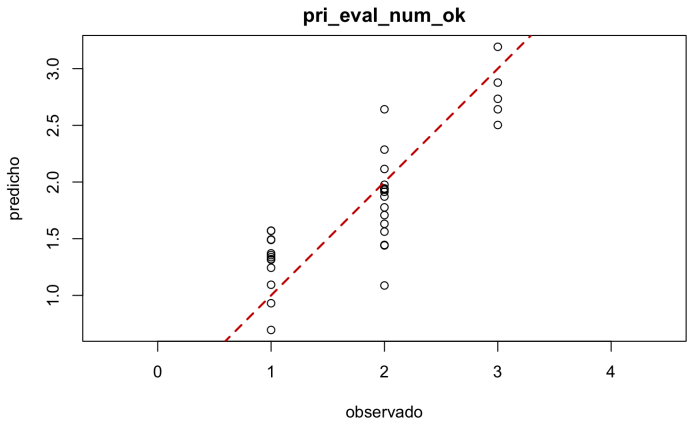

Figure XVI., Primera Eval Predict

For the classification model, we refer to the measures calculated from the confusion matrix, choosing class 3 as the positive class because we believe it is more important to identify patients who are significantly worsening, as it is the easiest class to explain. We will focus on metrics that account for imbalanced classes, as this is the case with our data, as we can see in Figure XVII in the Appendix.

The kappa index (0.67) indicates that there is substantial agreement between the PLS-DA classifications. The predictions are close to being made by chance.

The balanced accuracy values [17] (0.8278, 0.7882, 0.9828 for classes 1, 2, and 3 respectively) suggest that the model is making good predictions and, at the very least, is correctly predicting class 3. Similarly, it also predicts classes 1 and 2 quite well.

Random Forest:

The second model proposed for predicting "first_eval" is a Random Forest [18]. Random Forest is a machine learning algorithm used for both classification and regression tasks. It is an ensemble technique that combines multiple decision trees during training and returns the averaged prediction of the individual trees. The central idea behind Random Forest is to create a "forest" of independent decision trees, where each tree is trained with a random sample of the data and a random selection of features. Then, the predictions from each tree are combined by voting (for classification) or averaging (for regression) to produce a final prediction. Random Forest is known for its robustness and ability to handle large datasets with many features, as well as its ability to handle missing data and avoid overfitting. In addition, it can provide measures of feature importance, which helps to understand which features are most influential in the model. We will compare a Random Forest for classification with one for regression.

The training of the random forest regression and classification models was carried out using the ranger function from the ranger package in the R statistical software [19].

For both Random Forest models, classification and regression, we compared a model without tuning the hyperparameters, simply using the default settings, with a model with tuned hyperparameters, and a model with optimized parameters.

The regression model without hyperparameter tuning obtained the following metrics:

MSE	RMSE	MAE	R-squared
0.0721	0.2686	0.2374	0.8477

Table I. Metrics Regression Model I

For the model with hyperparameter tuning, we used the grid search technique based on cross-validation, optimizing the parameters ntrees (number of trees included in the model), mtry (number of predictors considered at each split), and max.depth (maximum depth the trees can reach). In Random Forest, the number of trees is not a critical hyperparameter since adding trees can only improve the result. Random Forest does not experience overfitting due to an excess of trees; however, adding trees beyond the point where improvement stabilizes is a waste of computational resources. The mtry value is one of the most important

hyperparameters in random forest, as it controls how much the trees are decorrelated from each other. The ranger package does not have any built-in functionality for cross-validation, so we used tidymodels.

The model that minimizes the RMSE value is the one with 500 ntrees, 7 mtry, and 20 max.depth. Although the RMSE is reduced, as we now obtain an RMSE of 0.264 compared to the initial 0.2684, the improvement in the data is minimal. The class membership accuracy of a patient is now 0.0044 more precise. As we mentioned earlier, this variation is entirely normal because, given that we are approaching a classification problem as a regression problem, we were not expecting exact class predictions but rather an integer value between the classes.

The classification model without hyperparameter tuning obtained the following metrics:

	CLASS 1	CLASS 2	CLASS 3
SENSITIVITY	0.9231	1.00	1.00
SPECIFICITY	1.00	0.9444	1.00
POS PRED VALUE	1.00	0.9412	1.00
NEG PRED VALUE	0.9545	1.00	1.00
PREVALENCE	0.3824	0.4706	0.1471
DETECTION RATE	0.3529	0.4706	0.1471
DETECTION PREVALENCE	0.3529	0.500	0.1471
BALANCED ACCURACY	0.9615	0.9722	1.00

Table II. Classification Model I

The hyperparameter tuning was performed in the same manner as in the regression model. The model that minimizes the F1-score value [17] is the one with 500 ntrees, 7 mtry, and 20 max.depth.

With the classification model, we were also able to assess the probability of classifying an observation into each class. This is very useful for a physician using this model in the future, as it can help them make better decisions.

Finally, we compared the variables that contribute the most, as we can see in Figure XVIII. and Figure XIX. in the Appendix, to each Random Forest model according to permutation and node purity. There is nothing noteworthy compared to what was observed in the PLS model.

XGBoost:

The third model we propose is an XGBoost or Extreme Gradient Boosting, one of the most widely used machine learning algorithms today. This algorithm is known for achieving good predictive results, particularly for problems with heterogeneous data. XGBoost is a supervised predictive algorithm that uses the principle of boosting. The idea behind boosting is to sequentially generate multiple "weak" predictive models, with each model taking the results of the previous one to create a "stronger" model with better predictive power and greater stability in its results. To achieve a stronger model, an optimization algorithm is employed, in this case, Gradient Descent. During training, the parameters of each weak model are iteratively adjusted in an attempt to find the minimum of an objective function. Each model is compared with the previous one. If a new model has better results, it is used as the basis for further modifications. If it has worse results, it reverts to the best previous model and modifies it in a different way.

For the regression model, we used the xgboost function from the xgboost package [20] in the R statistical software. In the parameter objective, we set "reg:linear" since we are addressing a regression problem.

In the tenth iteration of the model, we achieved an RMSE of 0.18, the best value among the regression models applied for predicting the "first_eval" variable.

B. Best response prediction

To predict the variable "best_response", which indicates the treatment's effectiveness in the patient and is also divided into four categories, we have conducted the same 3 predictive models as to predict “first_eval”. Again, for this purpose, we have selected a subset of variables since it does not make sense to use variables related to survival.

Therefore, the input variables will be as follows: physical characteristics, habits, other diseases, expression of the PD-L1 protein, tumor characteristics, cycles and toxicity, results of analytics referring to the before, first and second cycle of treatment. The expected output will be a numerical scale ranging from 0 to 3 will indicate the level of the patient's response to treatment. The closer the number is to 0, the more positively the patient is responding to treatment, and the closer it is to 3, the disease will continue spreading.

PLS & PLS-DA:

The first model we propose is a PLS and PLS-DA model. The models are built using centered and scaled data, and the training is carried out using the Leave One Out (LOO) technique.

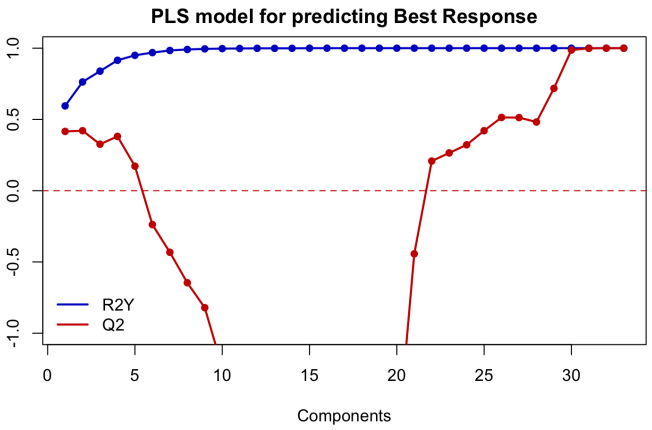

Figure XX. PLS Best Response

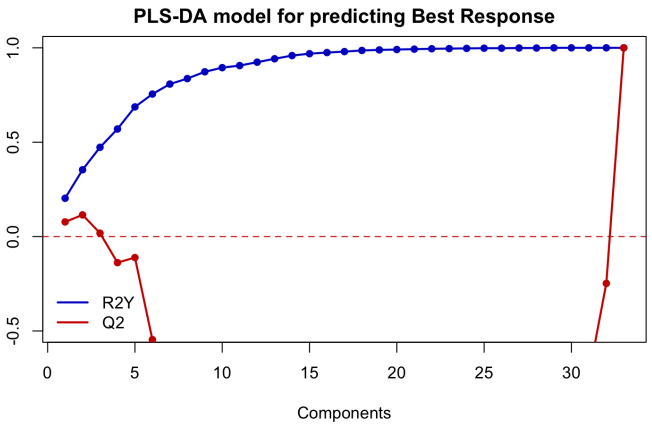

Figure XXI. PLS Best Response

Given the same scenario as when predicting the "first_eval" variable, we will follow the same criteria for choosing the number of components as in the previous section. This time, for the PLS model, we choose 4 principal components and for the PLS-DA model, we choose 2 principal components, as we can see in Figure XX. and Figure XXI.

Based on the score plots in Figure XXII and Figure XXIII., we can observe that for the PLS model, component 1 separates classes 0 and 1 from classes 2 and 3 very well, i.e., the extremes. For the PLS-DA model, component 2 differentiates classes 2 and 0, and component 1 differentiates classes 1 and 3 quite well.

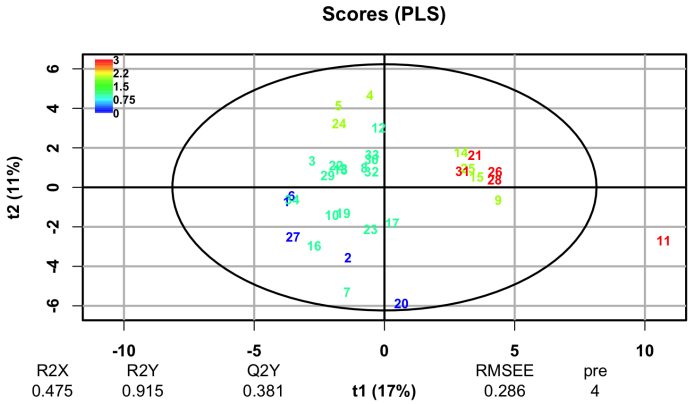

Figure XXII. PLS Score Best response

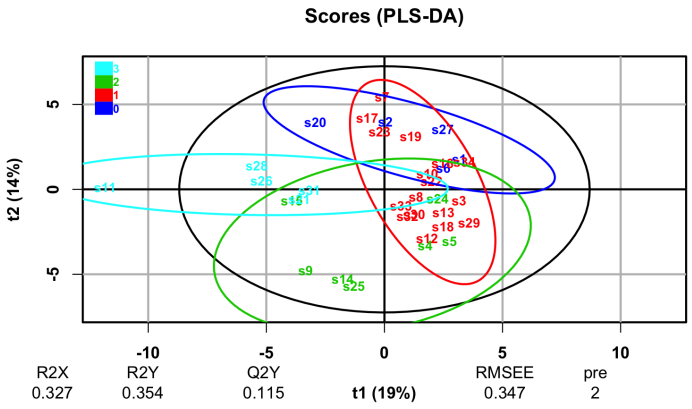

Figure XXIII PLS -DA Score Best response

In the loading plots in Figure XXIV and Figure XXV., we can see how the regressor variables relate to the response variable. Patients with a better response to treatment (values 0 or 1) are associated with a lower number of treatment cycles and lower or no toxicity. We can also observe that the initial values of SII and PLR are lower, indicating that these patients are in better condition from the outset compared to the rest. Conversely, we see that individuals who have a poorer response to treatment (class 2 or 3) have higher values of SII, NLR, PLR, or platelets. As mentioned earlier, these high values are completely logical for people who do not recover from cancer.

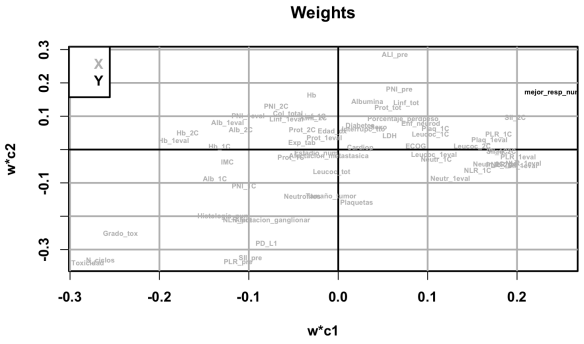

Figure XXIV. Loading Plots Best Response PLS

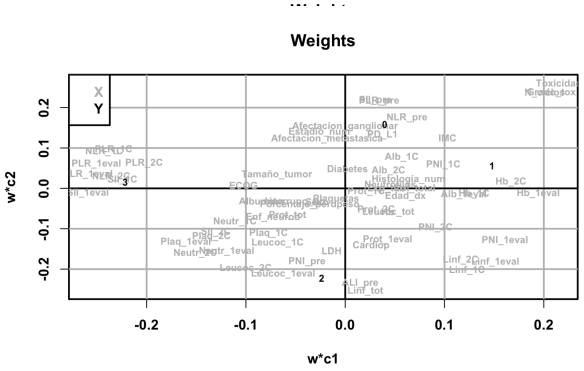

Figure XXV. Loading Plots Best Response PLS-DA

The evaluation measures proposed are the same as in the previous section for each regression or classification model. Similarly, it was impossible to reserve a test dataset to evaluate the model.

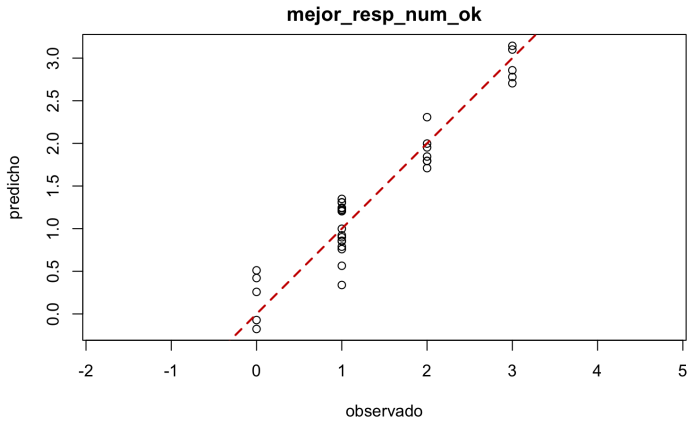

Figure XXVI. Loading Plots Best Response PLS-DA

The RMSE we obtained is 0.2638, meaning that on average, the predicted value differs by +/- 0.2638 from the actual value. This is expected since the observed values are integers (0, 1, 2, and 3) and the predicted values are real numbers between 0 and 3. The CVrmse we obtained is 0.195, which means that the model's predictions have an error of 19.5% relative to the scale of the original data. With this value, we can say that the model has a moderate to acceptable performance, as we can see in Figure XXVI.

The evaluation measures, as we can see in Figure XXVII. in the Appendix, of the PLS-DA model suggest that the predictions are not entirely incorrect. The kappa index (0.55) indicates decent agreement among evaluators, but there is still room for improvement. Looking at the confusion matrix, we realize that the model does not classify any observation in class 0 correctly and even misclassifies one as class 3, exactly the opposite.

The balanced accuracy values (0.5000, 0.7941, 0.71429, 0.9655 for classes 1, 2, and 3 respectively) suggest that the model is making decent predictions for classes 1, 2, and 3, but is predicting class 0 randomly. This is somewhat alarming, as it would be particularly important to identify the individuals who respond positively to the treatment, which this

model fails to detect. This is due to the highly imbalanced classes and the small number of observations available.

Random Forest:

The second model we propose is a Random Forest for classification and regression. As with the prediction of "first_eval," for both Random Forest models, classification and regression, we compared a model without hyperparameter tuning, simply using the default settings, with a model with tuned hyperparameters, and a model with optimized parameters.

The regression model without hyperparameter tuning obtained the following metrics:

MSE	RMSE	MAE	R-squared
0.0811	0.2848	0.2040	0.9

Table III. Metrics Regression Model II

For the model with hyperparameter tuning, we used the grid search technique based on cross-validation, optimizing the parameters ntrees, mtry, and max.depth. The model that minimizes the RMSE value is the one with 50 ntrees, 3 mtry, and 10 max.depth. Now, the situation is the opposite; the model increases the RMSE value to 0.32, meaning the model's performance is 3.5% worse. Despite considering all possible scenarios that could affect the grid search, such as using cross-validation, expanding the hyperparameter space, or implementing regularization and simplification techniques, the best RMSE achieved by tuning the hyperparameters is this one. Therefore, the best model obtained is the one using the default hyperparameters.

The classification model without hyperparameter tuning obtained the following metrics:

	CLASS 0	CLASS 1	CLASS 2	CLASS 3
SENSITIVITY	1.00	1.00	0.8571	1.00
SPECIFICITY	1.00	0.9412	1.00	1.00
POS PRED VALUE	1.00	0.9444	1.00	1.00

NEG PRED VALUE	1.00	1.00	0.9643	1.00
PREVALENCE	0.1471	0.500	0.2059	0.1471
DETECTION RATE	0.1471	0.500	0.1765	0.1471
DETECTION PREVALENCE	0.1471	0.5294	0.1765	0.1471
BALANCED ACCURACY	1.00	0.9706	0.9286	1.00

Table IV. Metrics Classification Model II

The hyperparameter tuning was performed in the same manner as in the regression model. The model that minimizes the F1-score value is the one with 500 ntrees, 7 mtry, and 20 max.depth.

Finally, we compared the variables that contribute the most (see appendix) to each Random Forest model according to permutation and node purity. There is nothing noteworthy compared to what was observed in the PLS model.

XGBoost:

The third model we propose is an XGBoost. For the regression model, we used the xgboost function from the xgboost package in the R statistical software. In the objective parameter, we set "reg:linear" since we are addressing a regression problem.

In the tenth iteration of the model, we achieved an RMSE of 0.16, the best value among the regression models applied for predicting the "best_response" variable.

C. Survival analysis

For survival analysis, we used the packages in the R statistical software that CRANS offers for this type of analysis [21].

In many cancer studies, the main outcome under assessment is the time to an event of interest. The generic name for the time is *survival time*, although it may be applied to the time 'survived' from complete remission to relapse or progression as equally as to the time from diagnosis to death. If the event occurred in all individuals, many methods of analysis would be applicable. However, it is usual that at the end of follow-up some of the individuals have not had the event of interest, and thus their true time to event is unknown. Further, survival data are rarely Normally distributed, but are skewed and comprise typically of many early events and relatively few late ones. It is these features of the data that make the special methods called *survival analysis* necessary [22].

For this task, a new dataframe has been created, including the following variables (explained in the appendix) of interest that are crucial for understanding the patients' profiles and the outcomes of the applied treatment: Age at diagnosis, Body Mass Index, Percentage of Weight Loss, Smoking Exposure (Packs/year), PD-L1 expression level, LDH, Total Protein, Albumin, Date of Birth, Date of Diagnosis, Date of Treatment Initiation, Date of SLP Event, Date of Death or Last Follow-up, and Censory Indicators.

For the graphs, we needed to have the survival time in days. This was calculated as the difference between the Date of Diagnosis and the Date of SLP.

Exploratory Data Analysis:

In the exploratory analysis, we compare the occurrence of death or treatment interruption in the sample of individuals with different attributes. We will say that the data is censored if death or treatment interruption has not been observed during the follow-up period.

The death or treatment interruption has only been seen in 67.65% of the patients, meaning that 32.35% of the data is censored.

The proportion between sex and censorship is similar in both sexes. Smokers and ex-smokers have a lower proportion of censored data, indicating that smoking at some point may contribute to the occurrence of death or treatment interruption. The absence of toxicity is also associated with higher censorship, which is not surprising because patients with higher toxicity are more likely to have their treatment interrupted or die. Patients with negative progression are not censored, while patients with positive progression are censored, suggesting that the negative evolution of the disease is strongly associated with death or treatment interruption.

We can separate the age in 3 groups:

- Age Group (40, 55): Mortality and censorship are balanced, with an equal number of deceased and non-deceased patients.
- Age Group (55, 70): The majority of patients have not died, indicating higher survival in this age group.
- Age group (70,85): Mortality and censorship are almost balanced, though there is a slight majority of patients whose death has not occurred.

Non-parametric estimation:

Non-parametric models do not assume any specific distribution for survival times and are directly based on the observed data.

After conducting a survival analysis combining the attributes of Gender, Age, Death, Toxicity, Smoking Habit, and Tumor Progression, the most relevant data are from the groups: Gender & Age, Gender & Smoking Habit, Gender & Progression, Smoking Habit & Progression, Toxicity & Progression, and Age.

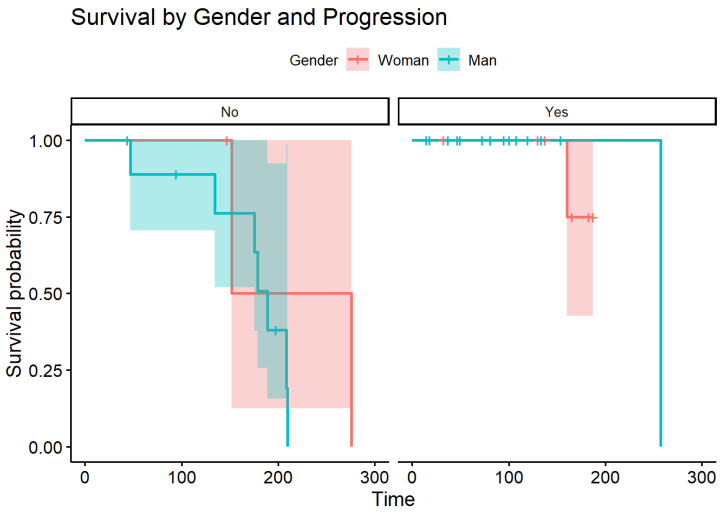

Figure XXVII. Gender and Progression

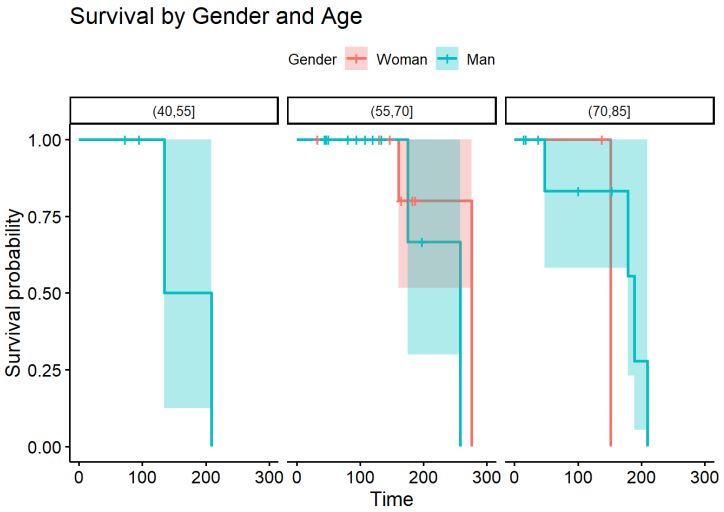

Figure XXVIII. Gender and Age

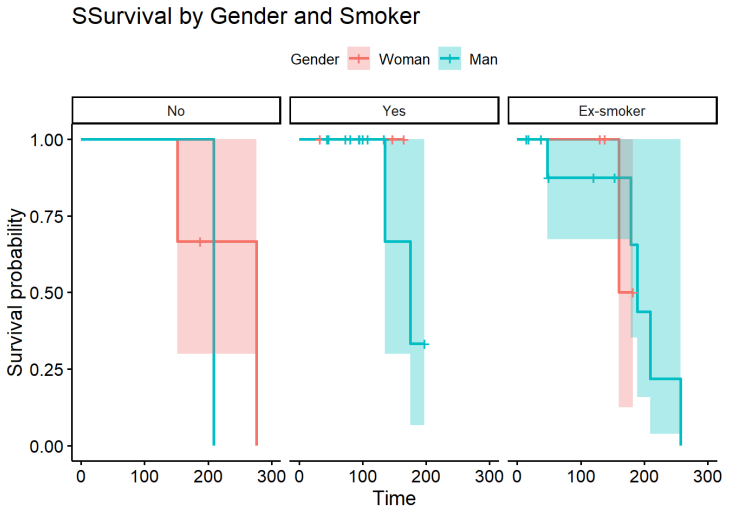

Figure XXVIV. Gender and Smoker

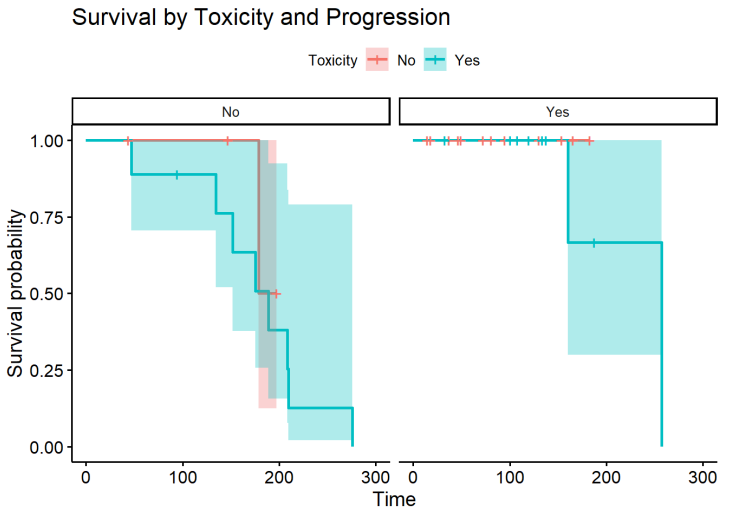

Figure XXX. Toxicity and Progression

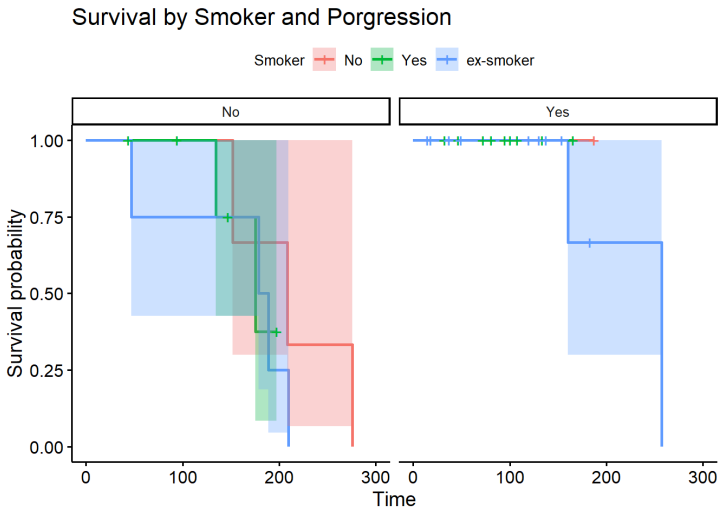

Figure XXXI. Smoker and Progression

To compare the survival functions between the following groups and to check that the survival curves represented in the graphs in Figure XXVII., Figure XXVIII., Figure XXVIV., Figure XXX. and Figure XXXI., are not due to chance, various Tarone-Ware tests (hypotheses) are conducted.

The null and alternative hypotheses are defined as follows:

- Null hypothesis (H0): There are no significant differences in the survival functions between the defined groups, meaning the survival curves are the same for all groups. p-value > 0.05
- Alternative hypothesis (H1): At least one group has a different survival function, meaning there is at least one significant difference in the survival curves between the groups. p-value < 0.05

The following table summarizes the groups, the p-value obtained for the test, and whether the null or alternative hypothesis is considered:

GROUPS	P-VALUE	HYPOTHESIS
Gender & Age	0.3	H0
Gender & Smoking habit	1	H0

Gender & Progression	0.4	H0
Smoking & Progression	0.7	H0
Toxicity & Progression	0.4	H0
Age	0.3	H0
Trends in age groups	0.1	H0

Table V. P-value Survival

The tests performed all show a p-value greater than 0.05, suggesting that there are no significant differences between the different groups studied. This implies that the observed differences in the survival curves could be due to chance. Therefore, it would not make sense to continue exploring survival, as the groups we have are not significant.

FUTURE WORK

Future research should focus on several areas to build upon the findings of this study. Expanding the sample size to include more patients will enhance the robustness and generalizability of the predictive models. Conducting prospective studies will help validate the findings and reduce potential biases associated with retrospective data. Continuously refining the predictive models by incorporating additional relevant variables and improving computational techniques will further enhance their accuracy. Developing user-friendly tools that integrate these predictive models into clinical workflows will facilitate their practical application in personalized treatment planning. In order to be able to reproduce this work and improve it, all the data and analysis performed can be accessed through the following link to the GitHub repository that hosts them: https://github.com/ainhoaprado/ProyIII-Group07-3CD

VALUE ASSESSMENT

The project's value and novelty lie in its development of advanced predictive models for assessing the response of advanced stage III or IV NSCLC patients to Pembrolizumab. These models enhance the personalization of cancer treatments, improving patient outcomes and optimizing healthcare resources. The primary beneficiaries include patients, oncologists, and healthcare systems, as the models provide a data-driven foundation for treatment decisions, potentially reducing costs and improving the quality of care.

The models are designed to be integrated into clinical workflows, aiding oncologists in making informed decisions by predicting patient responses in real-time. The project's impact is significant, offering better patient outcomes and more efficient resource use. However, the small sample size and retrospective data present limitations, emphasizing the need for future validation with larger cohorts.

In summary, the project's big picture addresses the need for personalized cancer treatment, highlighting the integration of predictive analytics into clinical practice and its broader implications for patient care and healthcare efficiency.

CONCLUSIONS

This study aimed to develop predictive models to assess the first radiological evaluation, the best response to treatment, and the overall survival of patients with advanced stage III or IV non-small cell lung cancer (NSCLC) treated with Pembrolizumab. Our findings highlight several key points.

Predictive modeling for the first radiological evaluation was conducted using PLS and PLS-DA models, as well as Random Forest and XGBoost models. Among these, XGBoost provided the best performance with an RMSE of 0.18, indicating its superior predictive accuracy. Similarly, for predicting the best response to treatment, XGBoost outperformed other models with an RMSE of 0.16, demonstrating its robustness in handling complex data and providing reliable predictions.

In the survival analysis, Kaplan-Meier plots and Cox proportional hazards regression were utilized. The exploratory analysis revealed significant associations between survival and factors such as smoking history, toxicity levels, and tumor progression. However, no significant differences in survival functions were found across different age and gender groups, suggesting these factors might not be as critical in influencing survival outcomes in our patient cohort.

Key predictors identified include the Systemic Immune-Inflammation Index (SII), Neutrophil to Lymphocyte Ratio (NLR), and Platelet to Lymphocyte Ratio (PLR), which were strong predictors for both the first radiological evaluation and the best response to treatment. These markers of systemic inflammation are crucial in understanding the patient's immune response to Pembrolizumab. Other important predictors included physical characteristics such as BMI, patient habits like smoking history, and tumor characteristics including PD-L1 expression.

Clinically, the models developed in this study provide valuable tools for clinicians to predict the outcomes of Pembrolizumab treatment in NSCLC patients. By identifying patients who are more likely to respond positively to the treatment, personalized treatment plans can be developed, potentially improving patient outcomes and optimizing resource allocation. Additionally, the ability to predict treatment response and survival can aid in clinical decision-making, allowing for more informed discussions with patients regarding their prognosis and treatment options.

Despite the promising results, the small sample size of 34 patients limits the generalizability of our findings. Future studies with larger cohorts are necessary to validate and refine our models. The retrospective nature of the study and potential biases in data collection could affect the results, highlighting the need for prospective studies to confirm the predictive power of the identified variables.

In conclusion, this study highlights the potential of using advanced predictive modeling techniques to improve the management of NSCLC patients treated with Pembrolizumab. By leveraging these models, clinicians can better tailor treatments to individual patients, ultimately improving outcomes and optimizing healthcare resources.

APPENDIX

GANTT DIAGRAM:

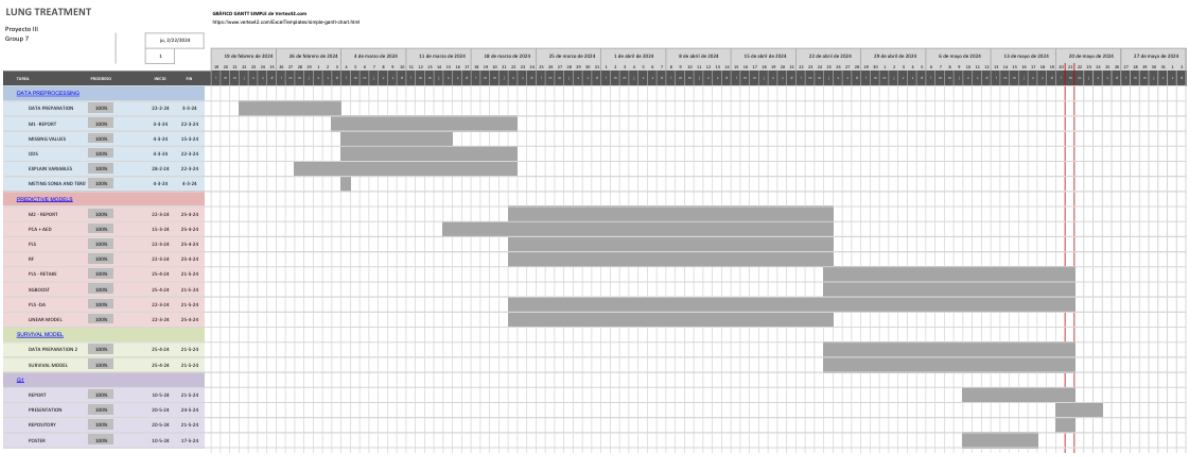

Figure I. Gantt diagram

DATA PREPROCESSING	María Gómez	María Verdú	Ainhoa	Blasco	Josema	Entrega
DATA PREPARATION	X		X			25/02/2024
M1 -REPORT		X		X		22/03/2024
MISSING VALUES	X		X			26/03/2024
ODS				X		10/04/2024
EXPLAIN VARIABLES					X	08/03/2024
METING SONIA AND TERESA	X	X	X	X	X	04/03/2024
PREDICTIVE MODELS						
M2 - REPORT		X				25/04/2024
PCA + AED	X					15/04/2024
PLS			X			25/04/2024
RF	X		X			25/04/2024
PLS - RETAKE		X				10/05/2024
XGBOOST		X				20/05/2024
PLS -DA		X			X	10/05/2024
LINEAR MODEL					X	25/04/2024
SURVIVAL MODEL						
DATA PREPARATION 2	X					10/05/2024
SURVIVAL MODEL	X					20/05/2024
G1						
REPORT	X	X	X			21/05/2024
PRESENTATION	X	X				24/05/2024
REPOSITORY			X			21/05/2024
POSTER	X	X		X		17/05/2024

Figure II. Gantt diagram (tasks)

PCA: Variable Contribution Graphics

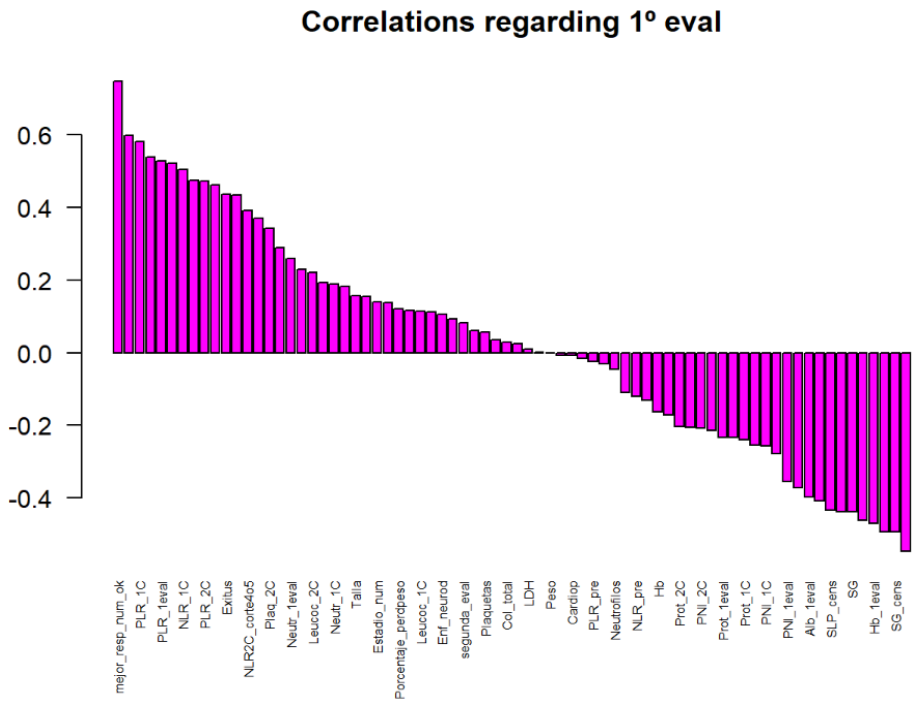

Figure IV. PCA-Correlation First Eval

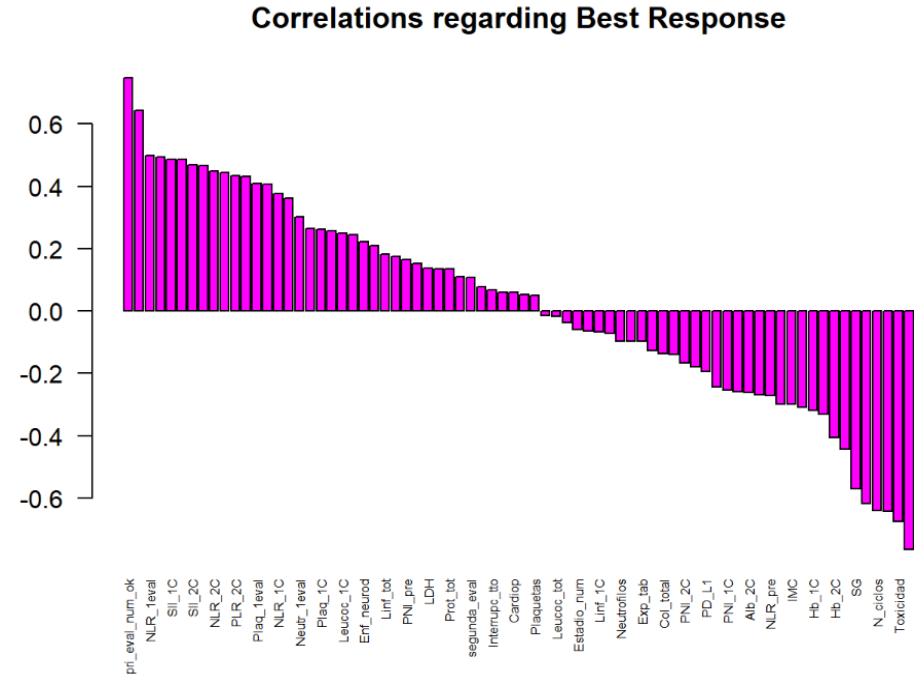

Figure V. PCA-Correlation Best Response

PLS: Variable VIP

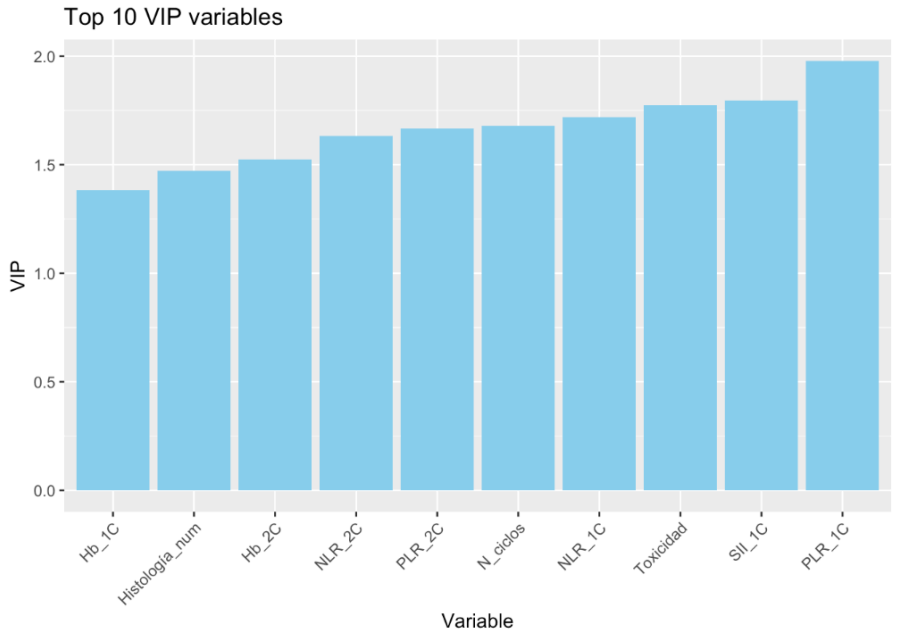

Figure XII. Primera Eval PLS

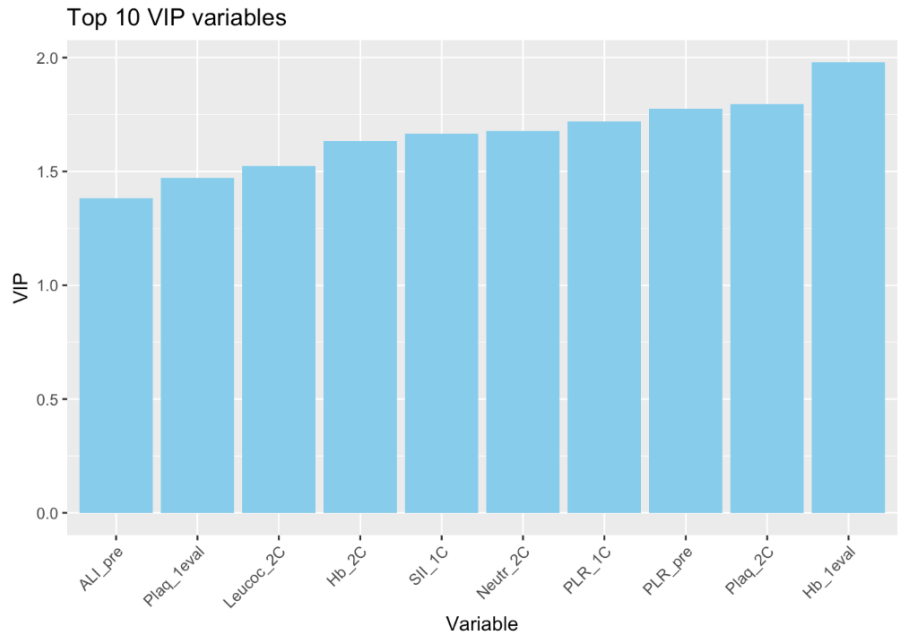

Figure XIII. Best Response PLS

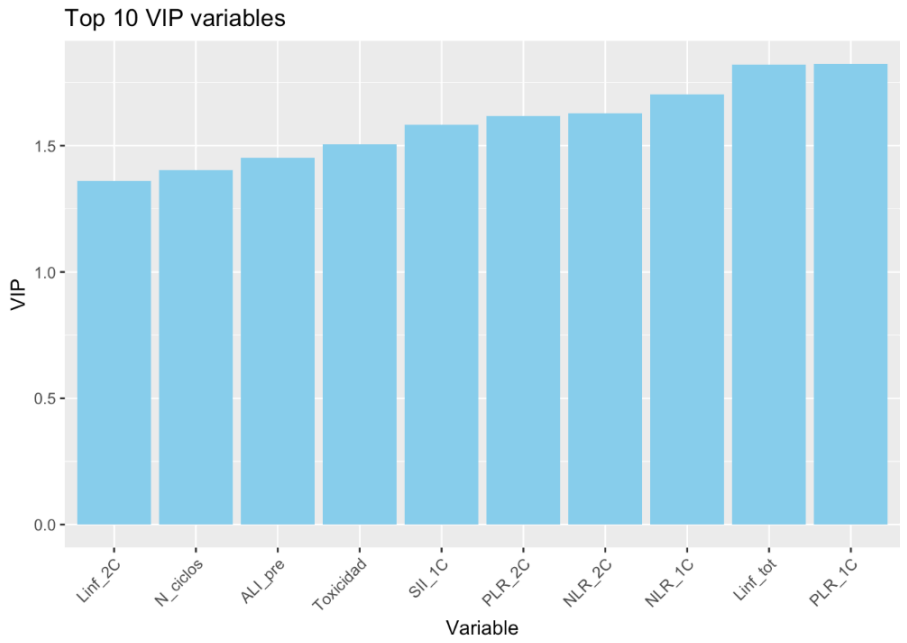

Figure XIV. Primera Eval PLS-DA

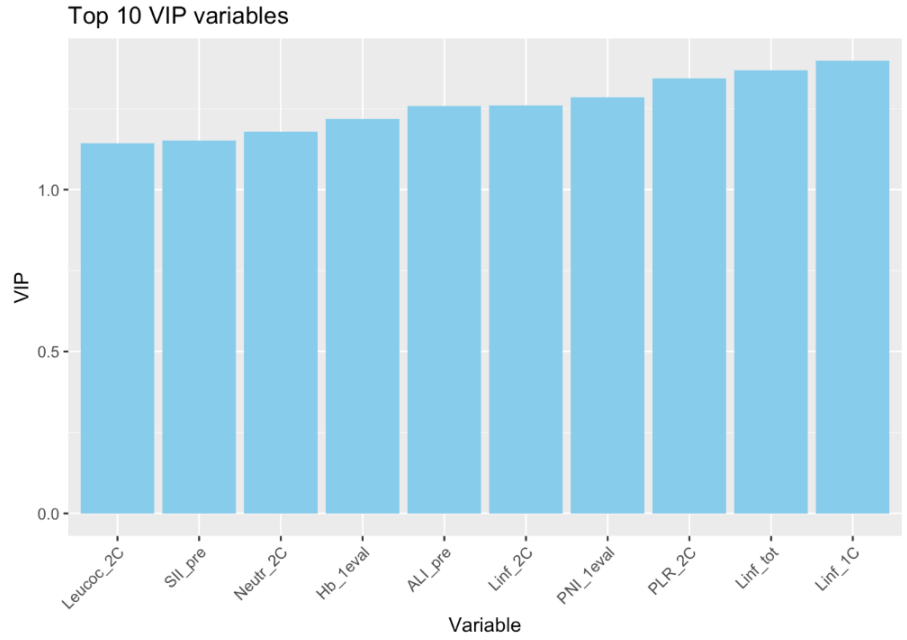

Figure XV. Best Response PLS-DA

PLS-DA: Metrics For imbalaced data Pri Eval

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  1  2  3
##          1 11  4  0
##          2  2 11  0
##          3  0  1  5
##
## Overall Statistics
##
##                Accuracy : 0.7941
##                  95% CI : (0.621, 0.913)
##     No Information Rate : 0.4706
##     P-Value [Acc > NIR] : 0.0001154
##
##                   Kappa : 0.6708
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3
## Sensitivity            0.8462   0.6875   1.0000
## Specificity            0.8095   0.8889   0.9655
## Pos Pred Value         0.7333   0.8462   0.8333
## Neg Pred Value         0.8947   0.7619   1.0000
## Prevalence             0.3824   0.4706   0.1471
## Detection Rate         0.3235   0.3235   0.1471
## Detection Prevalence   0.4412   0.3824   0.1765
## Balanced Accuracy      0.8278   0.7882   0.9828
```

Figure XVII. Metrics For imbalaced data

RANDOM FOREST: VARIABLE IMPORTANCE

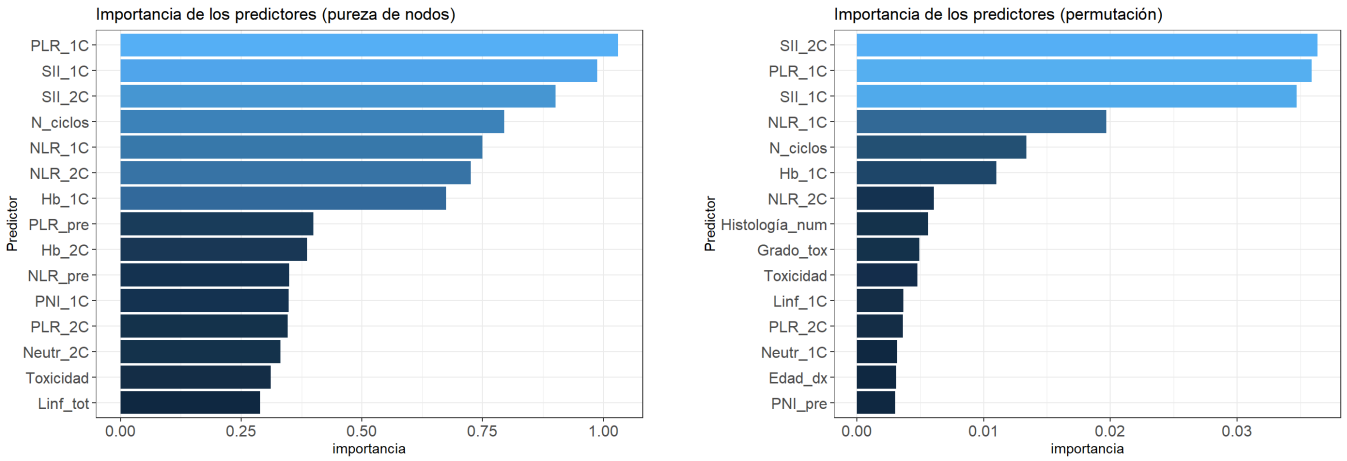

Figure XVIII. RF regression Pri Eval | Best response

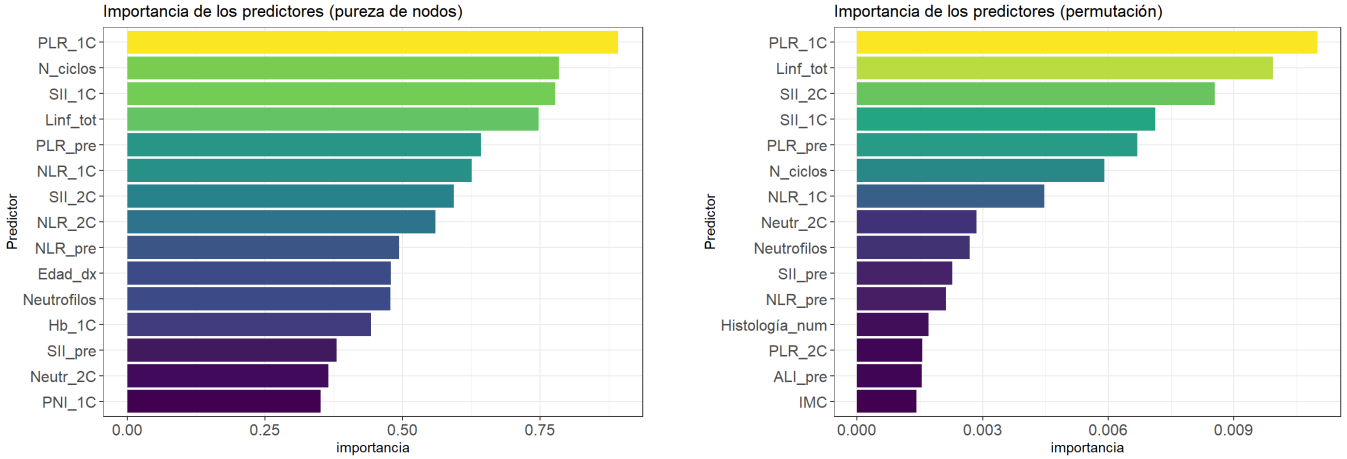

Figure XIX. RF Classification Pri Eval | Best response

PLS-DA: Metrics For imbalaced data Best Response

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1  2  3
##          0  0  0  0  0
##          1  4 17  3  0
##          2  0  0  3  0
##          3  1  0  1  5
##
## Overall Statistics
##
##                Accuracy : 0.7353
##                  95% CI : (0.5564, 0.8712)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : 0.004521
##
##                   Kappa : 0.5578
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 0 Class: 1 Class: 2 Class: 3
## Sensitivity            0.0000   1.0000  0.42857   1.0000
## Specificity            1.0000   0.5882  1.00000   0.9310
## Pos Pred Value            NaN   0.7083  1.00000   0.7143
## Neg Pred Value         0.8529   1.0000  0.87097   1.0000
## Prevalence             0.1471   0.5000  0.20588   0.1471
## Detection Rate         0.0000   0.5000  0.08824   0.1471
## Detection Prevalence   0.0000   0.7059  0.08824   0.2059
```

Figure XXVII. Metrics For imbalaced data Best Response

NEW DATA FRAME FOR SURVIVAL ANALYSIS:

Edad_dx (Age at Diagnosis): This variable represents the age of the patients at the time of lung cancer diagnosis. Age is a critical factor in cancer prognosis and can influence treatment outcomes.

IMC (Body Mass Index): BMI is included to evaluate the nutritional and health status of patients, which can affect their overall survival and response to cancer therapy.

Porcentaje_perdpeso (Percentage of Weight Loss): This indicates the percentage of weight loss experienced by the patient before the treatment initiation, serving as an indicator of cancer cachexia and overall health deterioration.

Exp_tab (Smoking Exposure in Pack-Years): Given the strong association between lung cancer and tobacco use, this variable quantifies the extent of smoking in pack-years, providing insights into the risk profile of each patient.

PD_L1 (Expression Level): PD-L1 expression levels are measured to determine the likelihood of response to anti-PD-L1 therapies like Prelozumb. Higher levels are often associated with a better response to immunotherapy.

LDH (Lactate Dehydrogenase): An enzyme that, when elevated, can indicate tissue damage and has prognostic significance in cancer.

Prot_tot (Total Protein) and Albumina (Albumin): These markers of nutritional status and liver function are essential for assessing the patient's general health and potential complications.

Fecha_nac (Date of Birth) and Fecha_dx (Date of Diagnosis): These dates are fundamental for calculating the age at diagnosis and understanding the timeline of each patient's disease progression.

Fecha_inicio_pem (Date of Treatment Initiation): The start date of Pembrolizumab treatment, critical for survival analysis.

Fecha_SLP (Date of SLP Event) and Fecha_exitus (Date of Death or Last Follow-up): These endpoints are used to calculate survival times, crucial for the analysis of treatment effectiveness.

Censoring Indicators (SLP_cens, SG_cens): These binary variables indicate whether the survival data for SLP (and overall survival, SG) are censored, meaning the event of interest (death or progression) has not been observed during the study period.

REFERENCES

[1] “Cancer”. World Health Organization (WHO). 2019 July. Accessed May 21, 2024. [Online]. Available: https://www.who.int/health-topics/cancer#tab=tab_1

[2] K. C. Thandra. “Epidemiology of lung cancer”. PubMed. 2023 February. Accessed May 21, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33911981

[3] B. Myre. “What’s the difference between small cell and non-small cell lung cancer?” Edward-Elmhurst Health. 2020 February. Accessed May 21, 2024. [Online]. Available: https://www.eehealth.org/blog/2020/02/lung-cancer-types/

[4] N. Franceschini, A. Frick and J. B. Kopp. “Genetic Testing in Clinical Settings”. *Amer. J. Kidney Diseases*, vol. 72, n.º 4, pp. 569–581. 2018 October. Accessed May 21, 2024. [Online]. Available: https://doi.org/10.1053/j.ajkd.2018.02.351

[5] R. Vaseq, A. Sharma, Y. Li and I. G. H. Schmidt-Wolf. “Revising the Landscape of Cytokine-Induced Killer Cell Therapy in Lung Cancer: Focus on Immune Checkpoint Inhibitors”. *Int. J. Mol. Sci.*, vol. 24, n.º 6, p. 5626. 2023 March. Accessed May 21, 2024. [Online]. Available: https://doi.org/10.3390/ijms24065626

[6] N. Hotz .“What is CRISP DM?”. datascience-pm. 2024 April. Accessed May 21, 2024. [Online]. Available: https://www.datascience-pm.com/crisp-dm-2/

[7] “THE 17 GOALS | Sustainable Development”. Sustainable Development. (s.f.). Accessed May 21, 2024. [Online]. Available:: https://sdgs.un.org/goals

[8] M. Zuil. “70.000 un trasplante de pulmón, 4.000 la cesárea… ¿Cuánto nos cuesta la sanidad?”. *El confidencial*. 2018 January. Accessed May 21, 2024. [Online]. Available: https://www.elconfidencial.com/espana/2018-01-29/coste-operaciones-medicas-hospital_1512125/

[9] “Analítica General en Valencia”. tuMédico. (s.f.). Accessed May 21, 2024. [Online]. Available: https://www.tumedico.es/analisis-clinicos/analitica-general/valencia#:~:text=39€%20Precio%20desde:,prevenir%20y%20cuidar%20tu%20salud

[10] J. Abal Arca, M. Á. Blanco Ramos, R. G. de la Infanta, C. Pérez López, L. González Pérez and J. Lamela López. “Coste hospitalario del diagnóstico del cáncer de pulmón”. *Arch. Bronconeumol.*, vol. 42, n.º 11, pp. 569–574. 2006 November.Accessed May 21, 2024. [Online]. Available: https://doi.org/10.1016/s0300-2896(06)70710-4

[11] S. v. Buuren and K. Groothuis-Oudshoorn. “mice: Multivariate Imputation by Chained Equations inR”. *J. Statistical Softw.*, vol. 45, n.º 3. 2011. Accessed May 21, 2024. [Online]. Available: https://doi.org/10.18637/jss.v045.i03

[12] J. A. Rodrigo. “Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE”. Cienciadedatos.net. 2017 June. Accessed May 21, 2024. [Online]. Available: https://cienciadedatos.net/documentos/35_principal_component_analysis

[13] K. H. Liland, B. Mevik, R. Wehrens and P. Hiemstra. “Partial Least Squares and Principal Component Regression”. 2017 November. Accessed May 21, 2024. [Online]. Available: https://cran.r-project.org/web/packages/pls/pls.pdf

[14] S. Kucheryavskiy. “Partial Least Squares Discriminant Analysis”. (s.f.). Accessed May 21, 2024. [Online]. Available: https://search.r-project.org/CRAN/refmans/mdatools/html/plsda.html

[15] A. Vehtari, A. Gelman and J. Gabry. “Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC”. *Statist. Comput.*, vol. 27, n.º 5, pp. 1413–1432. 2016 August. Accessed May 21, 2024. [Online]. Available: https://doi.org/10.1007/s11222-016-9696-4

[16] B. Greenwell and B. Boehmke. “Variable Importance Plots—An Introduction to the vip Package”, *R J.*, vol. 12, n.º 1, p. 343. 2020. Accessed May 21, 2024. [Online]. Available: https://doi.org/10.32614/rj-2020-013

[17] B. Hamner, M. Frasco and E. LeDell. “Evaluation Metrics for Machine Learning”. 2022 October. Accessed May 21, 2024. [Online]. Available: https://cran.r-project.org/web/packages/Metrics/Metrics.pdf

[18] L. Breiman. “Random Forest”. *Mach. Learn.*, vol. 45, n.º 1, pp. 5–32. 2001. Accessed May 21, 2024. [Online]. Available: https://doi.org/10.1023/a:1010933404324

[19] M. N. Wright and A. Ziegler. “ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R”. J. Stat. Soft., vol. 77, no. 1, pp. 1–17. 2017 March. Accessed May 21, 2024. [Online]. Available: https://doi.org/10.18637/jss.v077.i01

[20] T. Chen et al. “Extreme Gradient Boosting”. 2024 January. Accessed May 21, 2024. [Online]. Available: https://cran.r-project.org/web/packages/xgboost/xgboost.pdf

[21] T. G. Clark, M. J. Bradburn, S. B. Love y D. G. Altman, “Survival Analysis Part I: Basic concepts and first analyses”. *Brit. J. Cancer*, vol. 89, n.º 2, pp. 232–238. 2003 July. Accessed May 21, 2024. [Online]. Available: https://doi.org/10.1038/sj.bjc.6601118

[22] A. Allignol and A.Latouche. “CRAN Task View: Survival Analysis”. 2023 September. Accessed May 21, 2024. [Online]. Available: https://CRAN.R-project.org/view=Survival.