Kidney Transplant Prediction Model: A risk calculator web app.

Problem, Opportunity, and Impact

Overview of kidney transplants in Colombia

More than 7 Million people worldwide have chronic kidney disease [5] and it is related to multiple diseases, one of the most common is diabetes [3]. Between 2017 and 2018 there were 1'479,733 people in Colombia who had chronic kidney disease [4], which in its worst stage (when is lost about 85% percent of the kidney function [2]) led to kidney failure. The majority of patients with end-stage renal disease must receive a kidney transplant [1]. According to kidney transplant research carried out in Colombia studying 10-year experience over 1002 patients, 145 of them (14,4%) lost the transplanted kidney [3]. Also, another study on 78 patients of the Colombian Caribbean region between 2013 and 2018, showed that 46.2% of them experienced transplant rejection [4].

COLOMBIANA DE TRASPLANTES SAS is a company founded in 2002 that provides public health benefits through the provision of health services in transplants, improving the standards of life of patients who suffer from acute or chronic diseases and are susceptible to these treatments [6].

Currently, COLOMBIANA DE TRASPLANTES SAS has a database of more than 60,000 medical records, with information about patients such as age, gender, geographic location, blood group and laboratory tests. However, this data has been not used to predict the correct functioning of a transplant in patients who require it.

Define the specific problem that should be solved

The project will be focused on studying the information of approximately 1,500 patients in order to identify, how the collected variables affect positively or negatively the success rate of a kidney graft transplant surgery and to find the probability that patients have of receiving the desired diagnosis after treatment, because historically it has been observed that in some

patients the initial conditions of health prior to transplantation remain. Additionally, studies in kidney transplants using artificial intelligence approaches [5] have shown promising results.
The main purpose is to build a prediction model based on the clinical background included in the datasets provided, using all the variables available that show significant predictive value.
Finally, our motivation as a team is to contribute to improving the tools that doctors and patients have available to make the decisions, in order to improve the health of all those who require a transplant.

Also, given the Colombian Caribbean region study [4] from Colombiana de Trasplantes, we would like to include in our tool a map that allows this company to have a global view of their patients and determine the regions with the highest rate of rejection and success, after receiving the kidney transplant.

Why does this problem matter?

Patients who had kidney transplant surgery usually suffer from several chronic diseases such as diabetes, arterial hypertension (HTN), kidney chronic disease (CKD), etc., which generates excessive costs to Health Systems. Colombia has a similar situation as many countries, where there is a low availability of organs to donate for CKD patients [5].
Even after a positive immunological test between the patient and the donor, some experience a transplant rejection. Then, finding an innovative manner through analyzing the existing data that allow us to predict the rejection rate of kidney transplant patients will be of great use by Colombiana de Trasplantes.

Datasets + Data Wrangling & Cleaning

First Dataset

COLOMBIANA DE TRASPLANTES SAS provides us with the dataset 'historias_clinicas_anonimizado.csv', which is anonymized and has an unique identifier for each patient. In this dataset, we found 1,947 patients who received transplants, 75 patients who died and 75 patients who lost their transplanted kidneys.

The structure of the dataset is shown below:

Dataset columns	Type	Description
ENCODED_ID	int	Unique Identifier for each patient.
FECHA_TX	Datetime-Object	Date on which the patient is treated.
FECHA_CONSULTA	Datetime-Object	Date on which the patient attends.
EDAD_TX	int	Patient age at the date of treatment.
MUERTE	Bool	The patient has died until the consultation date.
FECHA_MUERTE	Datetime-Object	Date on which the patient has died.
RIÑON_FUNCIONAL	Bool	The kidney continues to work until the date of the medical consultation
FECHA_PERDIO_INJERTO	Datetime-Object	Date on which the patient lost the kidney
ENFERMEDAD_ACTUAL	String-Object	
OBSERVACIONES	String-Object	Text reported by the doctor about the general state of health of the patient
PLAN	String-Object	It corresponds to the data reported by the doctor about the treatment plan of each patient

Second Dataset

The second dataset received by Colombiana de Trasplantes has:

- **1616 records**
- **35 variables**

This dataset contains crucial information such as IMC, KDPI, diseases of patients and donors, and much more clinical data.

Dataset columns	Type	Description
ENCODED_ID	int	Unique Identifier for each patient
Edad (Años)	int	Patient Age at the moment of treatment
Sexo	object	Binary value, male or female
Fecha Tx	object	Date on the patient was treated
Etiología IRC	object	Causes related to the Chronic kidney disease
Meses de diálisis	float	Number of months of dialysis treatment before the treatment
Diabetes mellitus preTx (si/no)	object	Binary value, existence or not of diabetes
IMC (kg/m2)	float	Body mass index of the patient
Número de tx	int	Number of treatments the patient has had related to the disease before the surgery.
Tipo de donante	object	Category of the donor
Edad del donante (años)	float	Age of the donor at the moment of the treatment
Género del donante	object	Gender of the donor as binary value.
Peso Donante (kg)	object	Donor weight
Talla donante (m)	object	Donor height
IMC donante (kg/m2)	object	Body mass index of the donor
Isquemia fría (horas)	float	Time of cooling of the graft after the blood supply has been

		reduced or cut off.
Criterios expandidos	object	The presence of expanded criteria was defined as those donors aged 60 years or older or older than 50 years with at least two of the following conditions: history of arterial hypertension, serum creatinine > 1.5 mg/dl or cause of cardiovascular death. All other donors who did not meet these criteria were classified as the standard criteria group.
KDPI	object	Value of the donor KDPI: Kidney Donor Profile Index
Tiempo qx (horas)	object	Surgery duration
Lateralidad del injerto	object	Transplantation kidney laterality
Isquemia caliente	float	The time that passes from the interruption of the circulation of the donated organ until the moment in which it is perfused with the hypothermic preservation solution.
Complicaciones urológicas intraoperatorias	object	Binary value of complications that occurred during surgery (yes or no)
Complicación urológica (SI/NO)	object	Binary value of whether a urological complication occurred (yes or no)
Temporalidad complicaciones urológicas	object	Whether the urologic complications occurred soon or late.
Pérdida del injerto	object	Binary value of whether or not the graft was lost
Fecha de pérdida	float	Date when de graft was lost
Causa de pérdida	object	Description of the kidney loss causes. (13 types of causes)
Causa de pérdida.1	int	Code of the loss cause, from 0 to 13.
Nefrectomía del injerto	int	Binary value whether the graft was removed or not.
Fecha nefrectomía	float	Date when the graft was removed.

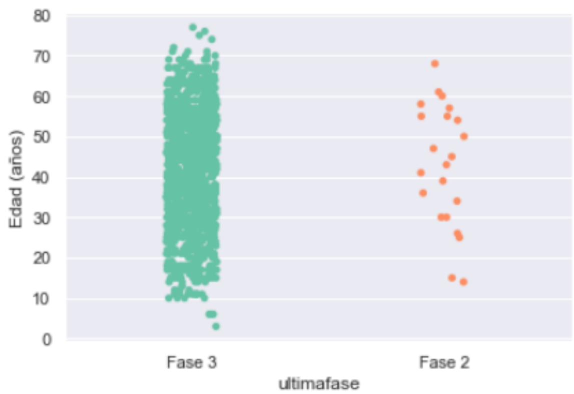
Mortalidad	object	Whether the patient died or not.
Fecha mortalidad	float	Date of the patient death
Rehospitalización	object	Whether the patient was admitted to the hospital again after the surgery or not.
Fecha rehospitalización	object	Date on the readmission of the patient.

Third Datasets

The third dataset received from Colombiana de Trasplantes contains:

- **1614 records**
- **3 variables**

This dataset allows us to know about the department where the patients lived at the moment of the treatment. It also gives information about the phase in which each patient was at the moment of the treatment.



Dataset columns	Type	Description
ENCODED_ID	int	Unique Identifier for each patient
departamento	object	Department where the patient resides
Última Fase	object	Phase in which the patient was at the moment of the treatment.

Exploratory Data Analysis (EDA)

With the three datasets, we proceed to join all of the files by the Encoded_ID, removing duplicate values and using all the possible variables in one unified dataset.

According to the exploratory data analysis, we can summarize the following basic information of the patients:

- 1617 Patients.
- 275 Patients lost the graft after the procedure.
- 260 Patients died.

Below is the description of the statistical data for the numerical variables of the base:

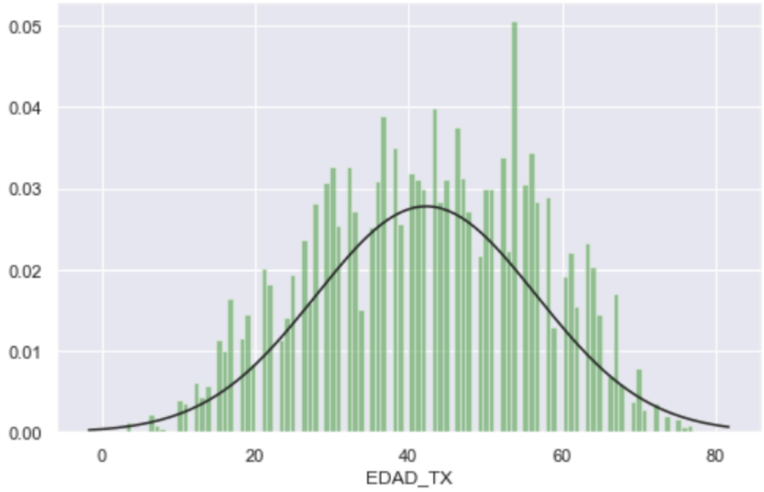
- Age: The age ranges from 3 to 77 years and the average age of patients is approximately 41 years.
- Dialysis Time: Dialysis time ranges from 0 to 276 months and the average duration of dialysis for patients is 33 months.
- IMC: IMC ranges from 12.9 (Kg/m2) to 39.6 (Kg/m2) and the average IMC of patients is 23.7 (Kg/m2).
- Donor Age: The age ranges from 4 to 73 years and the average age of donors is approximately 40 años.

	ENCODED_ID	Edad (años)	Meses de diálisis	IMC (kg/m2)	Número de tx	Edad del donante (años)	Isquemia fria (horas)	Isquemia caliente	Fecha de pérdida	Causa de pérdida.1	Nefrectomía del injerto
count	1.617000e+03	1617.000000	1568.000000	1591.000000	1617.000000	1614.00000	1617.000000	1617.000000	277.000000	1617.000000	1617.000000
mean	9.733382e+07	41.520099	33.293367	23.710277	1.040816	40.22119	13.708602	36.773655	41907.884477	0.691404	0.048856
std	2.077956e+07	14.827298	39.262661	4.415069	0.218719	13.87742	16.029890	9.325923	1180.921773	1.936014	0.215633
min	5.051201e+07	3.000000	0.000000	12.979989	1.000000	4.00000	0.000000	0.000000	39675.000000	0.000000	0.000000
25%	8.104325e+07	30.000000	6.000000	20.553804	1.000000	28.00000	1.500000	30.000000	40777.000000	0.000000	0.000000
50%	9.707142e+07	42.000000	18.000000	23.335466	1.000000	41.00000	14.000000	35.000000	42040.000000	0.000000	0.000000
75%	1.128637e+08	53.000000	48.000000	26.552855	1.000000	52.00000	19.500000	40.000000	42893.000000	0.000000	0.000000
max	1.542550e+08	77.000000	276.000000	39.622468	3.000000	73.00000	275.000000	125.000000	44089.000000	13.000000	1.000000

Exploring the variable 'AGE' / 'EDAD'

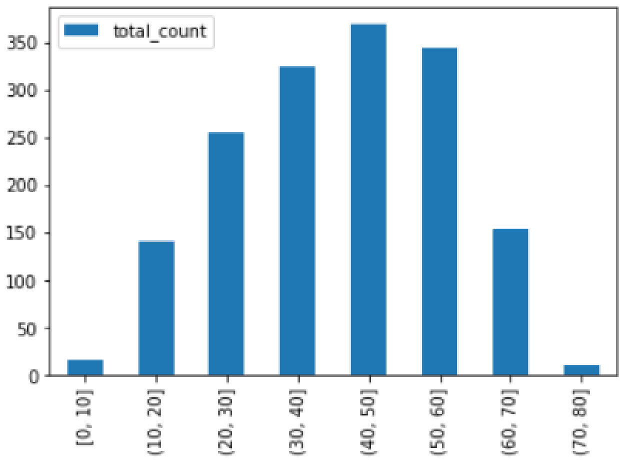
First it is important to show some basic exploratory information about the variable:

1) Distribution by age ranges:



2) A grouping is made according to age ranges as follows:

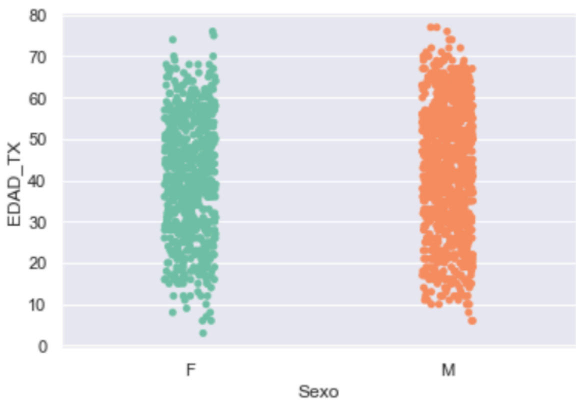
	intervals	total_count
0	(0, 10]	16
1	(10, 20]	141
2	(20, 30]	256
3	(30, 40]	324
4	(40, 50]	369
5	(50, 60]	345
6	(60, 70]	154
7	(70, 80]	12



Exploring the variable ‘Genre by age’

59% of patients are male and 41% female, which could lead us to think that males have more diseases that lead to kidney failure.

Genre by age




```
M    959
F    657
Name: Sexo, dtype: int64
```

Exploring the variable 'Donor type by age'

There are two types of donors: death / cadaveric or alive. Also alive donors can be divided into 4 categories:

Alive-related (Vivo relacionado): when it is a family member of the patients who donate the kidney.

Alive- not related (Vivo no relacionado): when it is not a family member of the patients who donate the kidney.

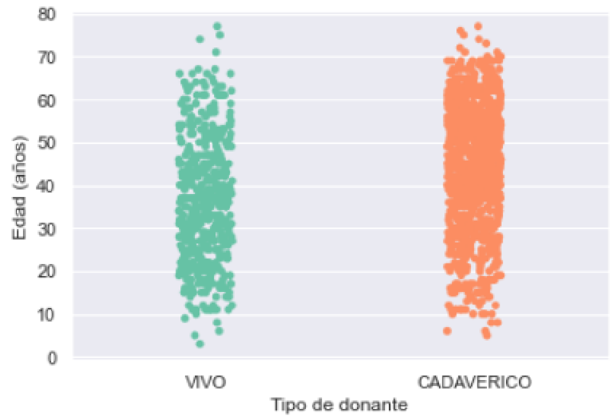
Alive: when people donated their kidney but it is not possible to know if it is related or not to the patient.

Autotransplant: these cases are not really a kidney transplant, these are cases where a patient had a malfunction in their kidney and they had to perform a surgery. The autotransplant case should be removed from the dataset.

```
CADAVERICO               1096
VIVO RELACIONADO          421
VIVO NO RELACIONADO        54
VIVO                       43
AUTOTRASPLANTE              2
Name: Tipo de donante, dtype: int64
```

To continue with the treatment of the data, after removing auto transplants, a grouping is carried out in such a way that there are two LIVE and CADAVERIC segments, in order to seek to minimize noise to make predictions and thus be able to test prediction models for donors, (ALIVE and CADAVERIC):

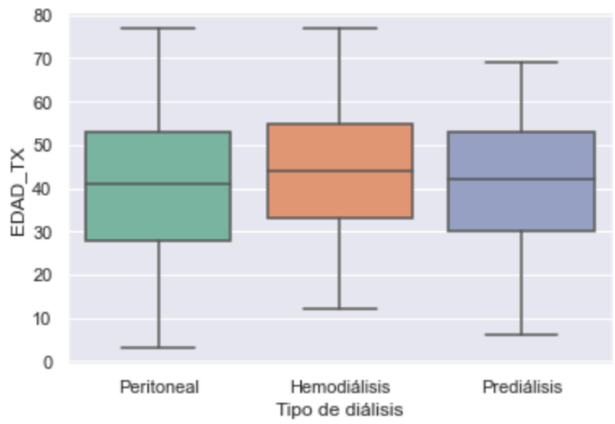
ALIVE = ALIVE-RELATED, ALIVE -NOT RELATED AND ALIVE
CADAVERIC = CADAVERIC



Exploring the variable 'type of dialysis by age'

There are 3 types of dialysis:

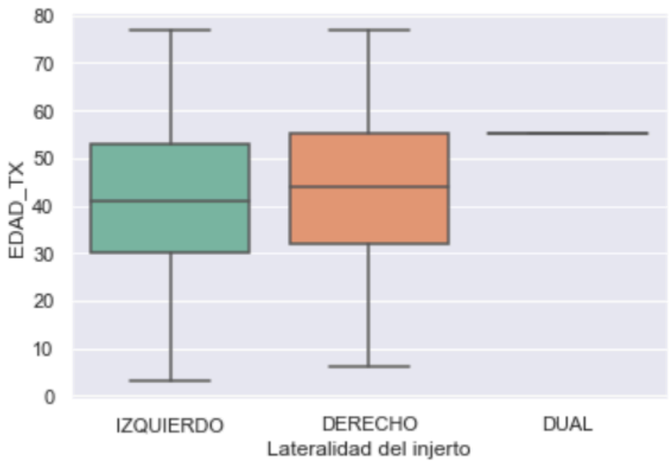
```
Hemodiálisis    808
Peritoneal      641
Prediálisis     166
Name: Tipo de diálisis, dtype: int64
```



Exploring the variable ‘Graft Laterality of the kidney's donor by age’

The original graft location of the kidney's donor; there are only two possible options: left or right:

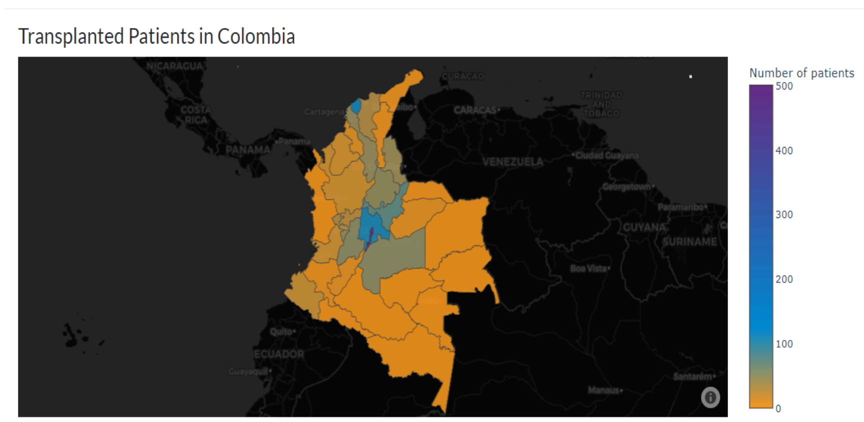
```
IZQUIERDO    930
DERECHO      685
DUAL           1
Name: Lateralidad del injerto, dtype: int64
```



57% of patients that received a kidney graft were on the left side of the donor, while 43% of patients received a kidney graft on the right side. The dual are special cases that need further analysis.

Colombia's Map of patient location

Given the patient's demographics data, their location is described by state, where the states with the most quantity of patients are Bogotá D.C and Cundinamarca. This concentration is shown in the map. In the center region, there are Caldas, Tolima, Quindío and Risaralda. Furthermore, the points highlighted are ones with the most quantity of patients are Atlántico, Boyacá, Meta, Santander and Bolívar.




```
BOGOTA D.C            501
ATLANTICO             112
CUNDINAMARCA          107
BOYACA                 74
TOLIMA                 53
META                   53
QUINDIO                49
SANTANDER              48
BOLIVAR                47
NORTE DE SANTANDER     43
RISARALDA              34
MAGDALENA              31
CALDAS                 26
VALLE DEL CAUCA        18
ANTIOQUIA              17
CORDOBA                15
SUCRE                   9
```

Modeling Phase

The classification models that we are going to explore in this project to determine the association between the **clinical variables** and the **graft loss of kidney** are:

- **Logistic Regression**: we expect a non-linear model explaining the relationship between the graft loss of kidney and the most important clinical variables.
- **Decision Trees**: we intend to use several different techniques because they are easy to understand, perform well and can handle numerical and categorical variables, which is what we have in the datasets provided.
  - **DecisionTreeClassifier:** this is an implementation from scikit-learn 1.1.1 of the CART algorithm.
  - **Random Forest:** is a meta estimator that fits a number of decision tree classifiers. As this technique looks for decreasing the variance of the forest estimator and we need a very precise estimation of the graft loss this is one of the classifiers we are exploring.
  - **Gradient Tree Boosting**: is an accurate and effective off-the-shelf procedure that we are going to try. There are different ways to use it as far as we know: Standalone Random Forest With XGBoost API and Standalone Random Forest With Scikit-Learn-Like API. We expect to use it next week.
- **Neural Networks:** As a more robust model and one that could be able to extract little variations in the data we will try a neural network to classify the graft loss of the kidney.

To start the data treatment in a prediction model, a segmentation by groups must be carried out in order to identify if it is possible to classify by age ranges and thus carry out the segmentation according to the groups obtained, but given the results obtained from the clustering , it is concluded that using these models as segmentation for these data is not the best decision, for this reason it is decided to discard the option of using clustering models and start the prediction with the general base and later divide it by type of donor classified as LIVING and CADAVERIC.

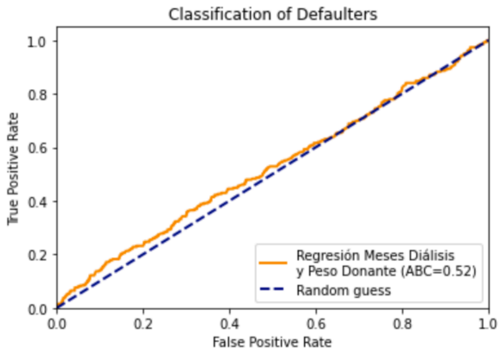
This modeling phase starts by first observing the distribution that the variables follow. Second cleaning the dataset where missing data and outliers were eliminated; we identified senseless data according to the description of their variables, for example, negative weights.

We are going to start by testing a **Logistic Regression Model** in order to determine how accurately it is predicting whether the kidney is functional or not.

Logit Regression Results

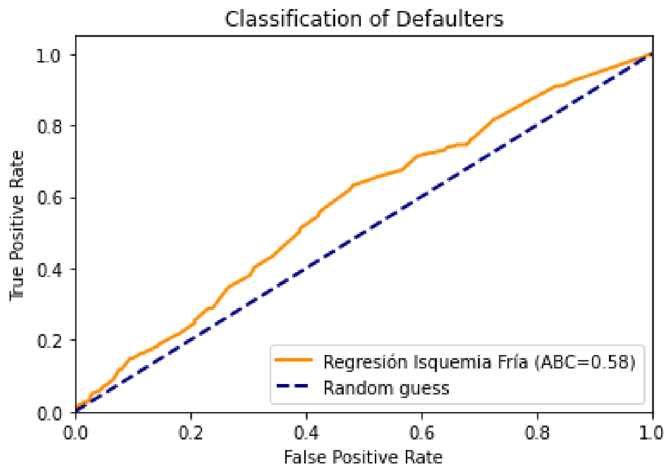
Dep. Variable:	Pérdida del injerto	**No. Observations:**	1547
Model:	Logit	**Df Residuals:**	1544
Method:	MLE	**Df Model:**	2
Date:	Mon, 06 Jun 2022	**Pseudo R-squ.:**	0.003046
Time:	21:18:51	**Log-Likelihood:**	-703.12
converged:	True	**LL-Null:**	-705.27
Covariance Type:	nonrobust	**LLR p-value:**	0.1167

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-1.8026	0.359	-5.026	0.000	-2.506	-1.100
Peso Donante (kg)	0.0014	0.005	0.286	0.775	-0.008	0.011
Meses de diálisis	0.0033	0.002	2.094	0.036	0.000	0.006



The results of training indicate that the model is not fitting well, that is, the interpretability of the model is not recommended for the prediction of graft loss, because the residual of the R squared is very low, and the ROC curve is 52%. This indicates that to improve the accuracy of the model, more variables must be included or a data transformation is needed.

Graft Loss ~ Cold Ischemia (hours)

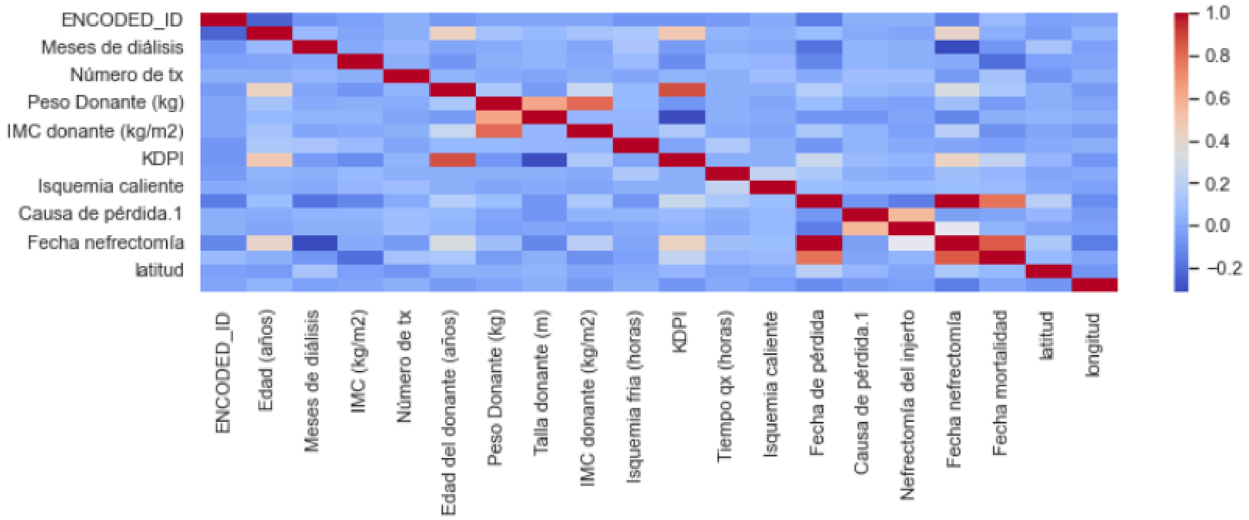


Given that the ROC curve is 58%, training graft loss with a single variable shows that the model does not predict well. Because of the mentioned above, it is suggested to train including all the variables or those with the highest correlation. However, depending on the results obtained if they are good enough we will move on to modeling phase two.

Feature Selection

A correlation between the numerical variables was observed where `Fecha nefrectomía`, `Fecha mortalidad`, `Nefrectomía del injerto` and `Edad del donante`, have a major correlation.

For the selection of variables, it must be taken into account that the donor's weight and donor's height are correlated with the BMI (KG/M2), since in order to calculate the BMI it is necessary to measure these two variables. Connections between KDPI and Months of dialysis are observed.



Since the correlation map for this case does not allow for the best selection of variables, since low correlations are observed, we look for methods that allow us to identify which variables are the most appropriate in order to have the best prediction. Among these methods we found the following:

1) **K-Best** (Supervised Method): This method shows the best k variables to fit the model.
2) **RFE** (Intrinsec Method): This method shows with respect to the objective variable, which independent variables will adjust with the best precision in order to get the best fit.

About these methods, we found that

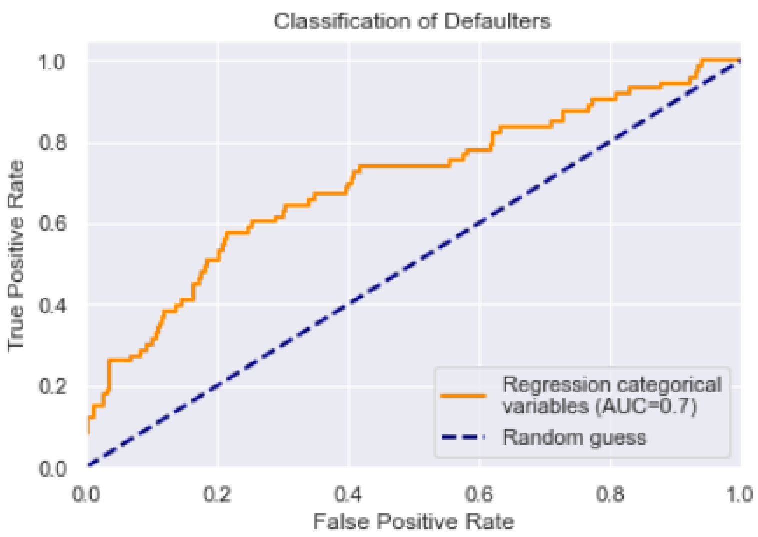
- With the K-Best method, a test was made with **5 numerical variables** with the objective of identifying the results's improvement, evaluating it on a logistic regression.

The variables that are categorical were passed to dummies to be able to adjust a logistic regression, and thus be able to evaluate the choice of variables made by the different methods proposed. Also we realized a change to the name of the variables, so we can have a more organized dataset.

Doing the feature selection with Random Forest Classifier from sklearn.ensemble and a logistic regression, having as a parameter that the selection of variables was repeated in more than 70% of iteration times, the following was obtained:

```
['Edad (años)',
 'Meses de diálisis',
 'IMC (kg/m2)',
 'Edad del donante (añc
 'Peso Donante (kg)',
 'Talla donante (m)',
 'IMC donante (kg/m2)',
 'Isquemia fria (horas)
 'KDPI',
 'Tiempo qx (horas)',
 'Isquemia caliente',
 'ultimafase_Fase 3']
```

After the feature selection we did tests with the selected variables with logistic regression models, and we observe that:



A division of the data was made **10%** for **testing; 90%** for **training,** and 3 models were made (Logistic Regression, Decision Tree and Random Forest).

When the three models are trained with the variables selected by Future Selection Random Forest, the three models show overfitting with the test data and when cross validation is applied, the models do not present a good classification.

Given this situation, it is decided to return to the initial variable selection step, which is based on the correlation of the data, and select the variables again with this criterion, for which the following variables were selected:

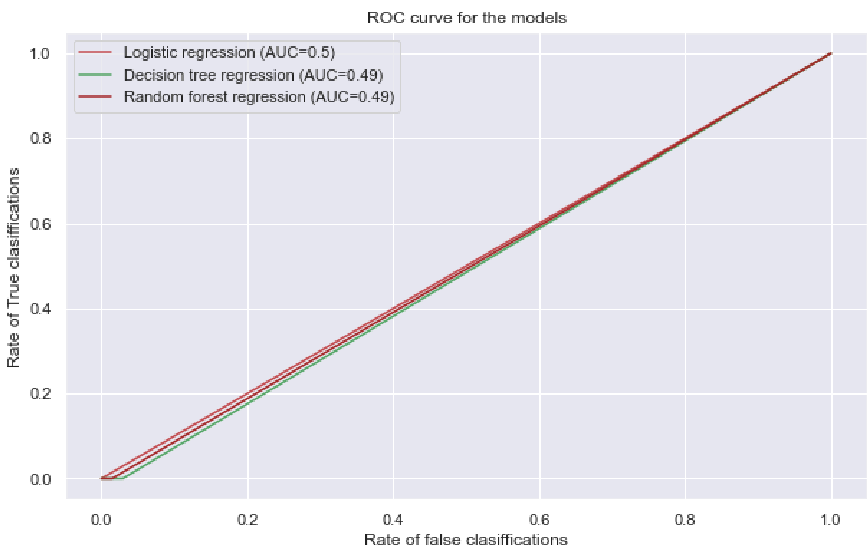
['Edad (años)', 'IMC (kg/m2)',

'Edad del donante (años)','Sexo','Género del donante','Etiologia IRC',
'IMC donante (kg/m2)', 'Isquemia fria (horas)','ultimafase',
'Tiempo qx (horas)', 'Isquemia caliente','KDPI','Lateralidad del injerto']

We obtained the following results:

```
Puntaje de precisión modelo logístico:  0.907
Puntaje de precisión modelo de arbol de decisiones:  0.88
Puntaje de precisión modelo de bosque aleatorio:  0.893


Matriz de confusión modelo logístico:
 [[68  7]
 [ 0  0]]
Matriz de confusión modelo de arbol de decisiones:
 [[66  7]
 [ 2  0]]
Matriz de confusión modelo de bosque aleatoria:
 [[67  7]
 [ 1  0]]
```

Carrying out the evaluation of the models using the AUC and ROC methods, we obtain the following graph:



Both the ROC curve, the AUC and the precision scores suggest **overfitting**, since the precision is high while the AUC is low. The main cause of the overfitting may be due to the lack of data, however, we will apply cross validation in order to seek to improve the accuracy of the test data.

Overfitting:

The overfitting that is occurring in the ROC curve when evaluating the data may be due to various reasons, including the selection of variables. Before modifying the selected variables, it is recommended to follow these steps in order to improve the overfitting that is occurring:

- Use cross validation.
- Collect more data.
- Introduce a complexity penalty with some regularization technique.
- Optimize model parameters with grid search.
- Reduce the dimension of the data.
- Apply attribute selection techniques.
- Use assembled models.

Given the above, we are going to use cross validation, that is applied to the general base in the following models and we got:

Random Forest:

For the random forest model it is decided to complement the random forest process with Hyperparameter tuning, also called hyperparameter optimization, is the process of finding the configuration of hyperparameters that results in the best performance. The process is typically computationally expensive and manual.

```
Random Forest Model Accuracy Score:  0.907
=== Confusion Matrix ===
[[68  0]
 [ 7  0]]


=== Classification Report ===
              precision    recall  f1-score   support

           0       0.91      1.00      0.95        68
           1       0.00      0.00      0.00         7

    accuracy                           0.91        75
   macro avg       0.45      0.50      0.48        75
weighted avg       0.82      0.91      0.86        75



=== All AUC Scores ===
[0.71268657 0.72574627 0.47947761 0.64179104 0.89552239 0.64179104
 0.78464819 0.43283582 0.66950959 0.68443497]


=== Mean AUC Score ===
Mean AUC Score - Random Forest:  0.6668443496801706
```

Logistic Regression:

```
Puntaje de precisión modelo de Logistic Regression:  0.947
=== Confusion Matrix ===
[[68  0]
 [ 7  0]]


=== Classification Report ===
              precision    recall  f1-score   support

           0       0.91      1.00      0.95        68
           1       0.00      0.00      0.00         7

    accuracy                           0.91        75
   macro avg       0.45      0.50      0.48        75
weighted avg       0.82      0.91      0.86        75



=== All AUC Scores ===
[0.49253731 0.57276119 0.50559701 0.55863539 0.68017058 0.7206823
 0.7228145  0.62260128 0.5010661  0.52238806]


=== Mean AUC Score ===
Mean AUC Score - Logistc Regression:  0.5899253731343282
```

Linear Discriminant Analysis:

```
Puntaje de precisión modelo de linear discriminant Analysis:  0.947
=== Confusion Matrix ===
[[67  1]
 [ 7  0]]


=== Classification Report ===
              precision    recall  f1-score   support

           0       0.91      0.99      0.94        68
           1       0.00      0.00      0.00         7

    accuracy                           0.89        75
   macro avg       0.45      0.49      0.47        75
weighted avg       0.82      0.89      0.86        75



=== All AUC Scores ===
[0.48507463 0.60074627 0.52238806 0.60341151 0.6673774  0.63113006
 0.75479744 0.6119403  0.49253731 0.554371  ]


=== Mean AUC Score ===
Mean AUC Score - Linear Discriminant Analysis:  0.5923773987206823
```

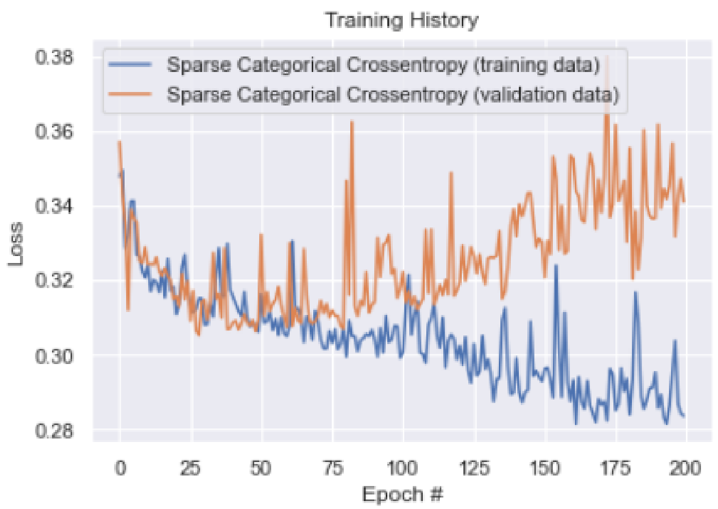
Gradient Boosting:

```
Puntaje de precisión modelo de Gradient Boosting Classifier:  1.0
=== Confusion Matrix ===
[[66  2]
 [ 7  0]]


=== Classification Report ===
              precision    recall  f1-score   support

           0       0.90      0.97      0.94        68
           1       0.00      0.00      0.00         7

    accuracy                           0.88        75
   macro avg       0.45      0.49      0.47        75
weighted avg       0.82      0.88      0.85        75



=== All AUC Scores ===
[0.61940299 0.66604478 0.55970149 0.67803838 0.68230277 0.70362473
 0.68443497 0.59061834 0.62686567 0.57569296]


=== Mean AUC Score ===
Mean AUC Score - Gradient Boosting Classifier:  0.6386727078891259
```

Based on the results obtained after applying cross Validation with hyperparameters tuning, it is observed that the model with the best learning after training is the Random Forest, with an accuracy of 90% and an AUC of 66%. It is always desirable to have a higher AUC, but since the dataset is not robust, the AUC does not have enough data to learn, for this reason it is recommended to strengthen the data in order to have better results.

As it was initially commented, the idea to train the models was to divide the types of donors into different models for CADAVERIC and for ALIVE. However, after training the model, it was observed that the results were similar, for this reason, a single model is selected for the general base.

After testing the models mentioned above, it is decided to train a neural network and the following results were obtained:



The neural network fits around 90%. Therefore, it is suggested that in order to have a neural network with greater precision, more robust data is necessary and thus the model can learn from the initial layer and as its adjustment progresses it improves, otherwise of the observed with these data, where the model starts at 90% and is maintained with the same precision and with a loss of 30%. For future projects related to the prediction of functional kidney transplants, it is recommended to have as many patients as possible, in order to have more learning data.

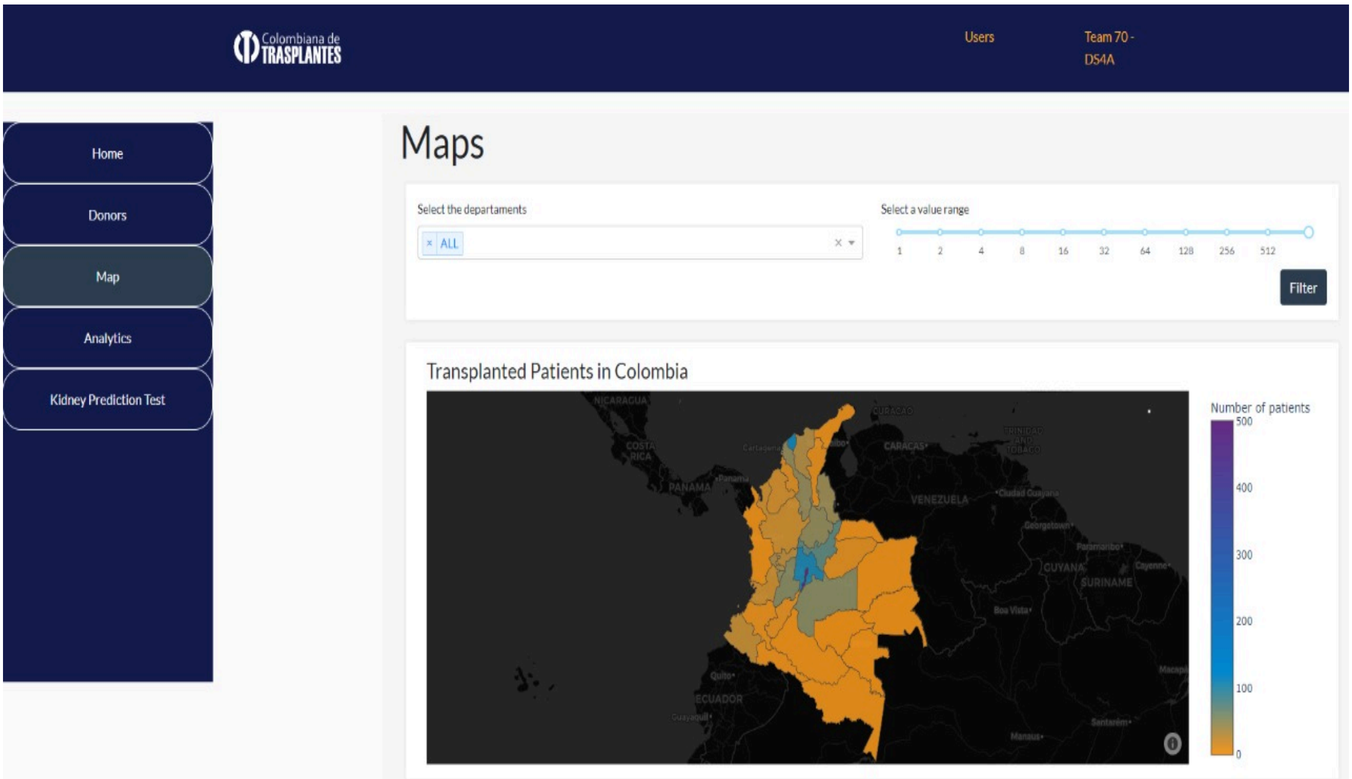
WEB APP

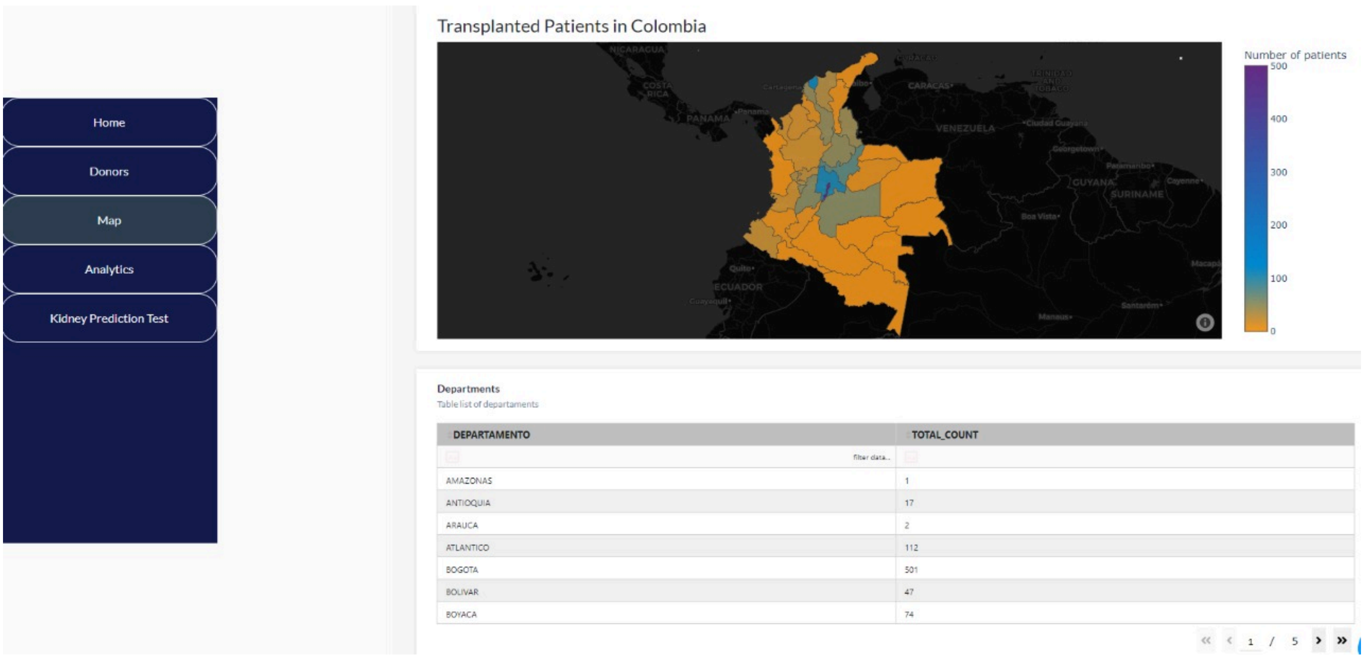
For the WEB APP was implemented a dashboard that allows the navigation through 5 sections:

- ***Home***: Here is presented some basic demographics:

- ***Donors***: In this section is shown some other demographic info about the Donor population:

- ***Map***: Here we show the heat map of the location of the patients by department. It has a filter that allows the user a more specific visualization.





- ***Analytics:***

  The analysis of the COLOMBIANA DE TRASPLANTES SAS database, was made over 1617 patients, where it get found categorical variables as, sex, type of dialysis, etiology, type of donor, laterality of the graft, diabetes and numerical variables, such as, BMI, age, months of dialysis, cold ischemia and warm ischemia. The main idea was to partition the database by type of donor: Cadaverico and Vivo, where the category Vivo includes the categories alive-related, alive-not related and alive. On the other hand, we searched by clustering and k-means to group by age range without major results, furthermore the model did not have greater assertiveness for this reason it was decided to do the segmentation with this database and the general database where the results were similar, the AUC it was 66%.

  Carrying out a more detailed analysis of the variables available here, a similarity was observed in the laterality of the graft, since 58% of the patients had
  left graft laterality, while 42% of patients had right graft laterality, there is a record that corresponds to dual laterality, but is not taken into account, referring to the Gender variable, it is evident that 59% of the data correspond to male patients, while 41% correspond to female patients, regarding the types of dialysis performed, it is highlighted that 50% of the types of dialysis are Hemodialysis, 40% Peritoneal and the remaining 10% predialysis.

Concerns

Here we summarize the main concerns of the team during all the project:

- The way this tool is going to be used by the doctors of COLOMBIANA DE TRASPLANTES, taking into account that due to the short time of the project we can not guarantee the best prediction model.
- We felt that the time given to build a complete web application tool was quite short, since even though most of the team members have programming experience, none of us have ever implemented a web application in a real environment with a backend in python and with all the processes that are implicit in a prediction model.

References

[1] "Dialysis," *nhs.uk*, Oct. 19, 2017. https://www.nhs.uk/conditions/dialysis/ (accessed May 26, 2022).

[2] "What is Dialysis?," *National Kidney Foundation*, Dec. 24, 2015. https://www.kidney.org/atoz/content/dialysisinfo (accessed May 26, 2022).

[3] D. Espitia, A. García-López, N. Patino-Jaramillo, and F. Girón-Luque, "Desenlaces a largo plazo en pacientes trasplantados renales con donantes de criterios expandidos: experiencia de 10 años," *Rev Colomb Cir*, Feb. 2022, doi: 10.30944/20117582.1052.

[4] A. Estupiñán-Bohórquez, J. Acosta-Reyes, D. Viasus-Pérez, A. García-López, N. Patino-Jaramillo, and F. Girón-Luque, "Trasplante renal de donantes con criterios expandidos en la región Caribe colombiana," *NEFRO*, vol. 18, no. 2, p. 7533, Dec. 2021, doi: 10.24875/NEFRO.21000028.

[5] M. Raynaud *et al.*, "Dynamic prediction of renal survival among deeply phenotyped kidney transplant recipients using artificial intelligence: an observational, international, multicohort study," *The Lancet Digital Health*, vol. 3, no. 12, pp. e795–e805, Dec. 2021, doi: 10.1016/S2589-7500(21)00209-0.

[6] "Centro de Trasplantes de riñón e hígado | Colombiana de Trasplantes," *Colombiana de Transplantes*. https://colombianadetrasplantes.com/web/ (accessed May 26, 2022).