Script

Introduction (Galo)

Our topic is: Models for Management of Congenital Heart Disease in Mexico (ABC Hospital).
Our team members: Daniela Soto, Alberto Valdez and Galo Cazares

Congenital Heart Diseases
Congenital heart diseases (CHD) are a relatively common problem with an incidence of 6/1000 newborns. It is a critical fact that 30% of these CHD are never diagnosed.

In Mexico, approximately 18,000 to 21,000 children are born each year with a congenital heart disease.. Only 5% to 8% of the population have access to private insurance and 40% will have access to government-run hospitals.

State of Care for Congenital Heart Diseases
According to the Society of Thoracic Surgeons (STS) Congenital Heart Surgery Database, in the United States with almost 40,000 operations per year. The percentage of mortality after surgery is 8.8% in neonates.

In Mexico it is estimated that 90% of CHD patients receive suboptimal treatment due to financial issues. The mortality rate in Mexico is 15%.

Kardias Private–Public (Hybrid) Program
We choose to work with Kardias Foundation from Mexico which has a specialized program in two hospitals: the National Institute of Pediatrics (INP) and the ABC-Kardias Pediatric Heart Center (in partnership with the ABC Medical Center).

U.S. Food & Drug Administration's Artificial Intelligence Challenges
From the U.S. Food & Drug Administration's Artificial Intelligence and Machine Learning Challenges we chose to aim our project to help with more accurate diagnosis, prognosis, and risk assessment.

Congenital Heart Disease in Mexico (Kardias Patients)
The top places in Mexico with CHD are: State of Mexico and Mexico City, we can find 306 cases for the former and 342 for the latter.

Zooming into the area we can observe that the counties with more CHD incidence are Iztapalapa with 52 patients and GAM with 44 respectively.

World Database for Congenital Heart Surgery 2011-2012
It is important to mention that in the year of 2011 there was an initiative to create a World Database of Congenital Heart Surgery. The study was done only for 1 year, we used some data from this study to compare our project results.

Cleaning ETL (Dany)

The data given by Kardias required multiple changes, it can be seen that all the records were added by different medical staff and it included a high number of capturing mistakes.
The exploratory data analysis started in Excel where we modified some small orthographic errors, changed names and verified the differences between rows.
-
In python, some unnecessary columns were deleted and all columns were renamed. Then, we dropped null values and changed the data types.
Also, a new Python library was created for the project with Kardias. The objective of this package is to normalize records (like medical notes) that include Natural Language, by using approximate string matching and various ETL methods like no-stopwords and basic RegEx. It can be used for data cleaning, which in this case resulted in a 20% variance reduction.
After all the cleaning processes, the data still required some changes.
So we created a dictionary of main diagnosis categories with their RegEx pattern; it was used to verify if a string contains specific patterns to avoid duplicated categories.
Data were divided in two groups, high and low frequency, having all the high frequency results as unique categories and creating a new category named Others for the low frequency data.
After all the cleaning phases, we finally have a useful database for visualizations and machine learning.
-

Visualizations (Galo)

surgical procedures

The next graph shows some important main surgeries and the number of patients that had each surgery during 2011 to 2012 (extracted from a Clinical Investigation paper.) and during 2012 to 2022. We can see here that the Closure of Arterious Conduct and the number of Closure of Inferior Cava Vein increased in 2022, the Repair of the Tetralogy of Fallot was reduced and some procedures had similar numbers of patients in both ranges of years like the AV Canal Repair.
-
rachs vs mortality

The next bar chart shows RACHS vs Mortality. RACHS means Risk Adjusted Classification for Congenital Heart Surgery, it is a classification applied to surgical procedures and it's grouped in 6 numerical categories, being 6 the most risky.
The data shows that the hospital had a small number of patients that required a risky procedure, and the results from RACHS 6 shows that from 7 risky procedures, the mortality number was 4, corroborating that RACHS 6 has the highest mortality number.

ML / Deployment (Beto)

ML

So how can we use Machine Learning to improve clinical decision-making?
We chose to focus on predicting two variables:
Mortality Rate and Number of days that the patient will stay in the Intensive Care Unit.
In the case of a Neural Network, we get high accuracy but low sensitivity. Turns out that having a low representation of positives in our population doesn't help on getting a good sensitivity score. Here is the prediction matrix after a few iterations of trying to improve the model performance.
We used K-Means clustering to help us understand the relationship between all the variables and reduce the dimensions of our data, so whenever we get a new patient, we can input their data and get a corresponding cluster with an approximate number of stay days.
Finally, as a way to assist the interpretation of the clusters, we used Linear Regression on the Rachs score and the Stay Days for each cluster, which showed a fairly good relationship between the chosen variables.
We can always improve the outliers before and after the clustering as well as focusing on more accurate categorization of the Diagnosis and Procedure data by consulting medical experts.

Deployment

Thinking on this iterative collaboration, we want to have a way for the end-user to try the model and even input Patient data in the future. So we deployed our model and REST API to Amazon Web Services with a temporary front-end on GitHub pages. This is key if we want to have a full-circle data pipeline where we can improve data quality, feature engineering, and model performance. **DEMO**.