INSTITUTO TECNOLÓGICO DE BUENOS AIRES - ITBA
ESCUELA DE INGENIERÍA Y GESTIÓN

PREDICCIÓN DEL START-UP TIME DE UN ENSAYO CLÍNICO

AUTOR/ES:
Colantonio, Santiago (Leg. Nº 58724)
Fichendler, Sebastián Alejandro (Leg. Nº 60112)

DOCENTE/S TITULAR/ES:
Rodriguez Varela, Juan Pablo
Gonzalez, Rubén Darío

PROYECTO FINAL PRESENTADO PARA LA OBTENCIÓN DEL TÍTULO DE LICENCIADO/A EN ANALITICA EMPRESARIAL Y SOCIAL

BUENOS AIRES
SEGUNDO CUATRIMESTRE 2022

Índice

Resumen Ejecutivo

El siguiente informe es un compendio de toda la labor hecha durante todo el segundo semestre del año 2022 en el marco del proyecto de la predicción del *start up time* de un ensayo clínico junto a la compañía farmacéutica Empresa ABC. Esto con el fin de obtener un modelo predictivo que ayude a la hora de estimar el tiempo que se demora en arrancar un ensayo clínico.

En primer lugar se brindó información sobre la farmacéutica en sí. Su presencia a nivel mundial, negocios e historia de la misma. Todo esto para poder encuadrar en contexto la figura de Empresa ABC. A su vez, se informó sobre lo que es un ensayo clínico, sus etapas y sus características.

Se procedió a explicar cuál era el objetivo y el enfoque que se le dio al proyecto, como así también las metodología utilizadas para encarar el mismo y con qué datos se contaban para la realización del mismo .

Luego, se hizo un análisis exploratorio de los datos que brindó Empresa ABC. En este se buscó mostrar con qué tipo de datos se contaba, que significaban y qué relación tenían con la variable objetivo del estudio. En base a este análisis es que también se hizo una limpieza de la base de datos como así también la agrupación de los valores que podían tomar las variables en nuevas variables. A continuación se procedió a codificar las variables para dejarlas listas para el desarrollo de los modelos.

El desarrollo de los modelos fue el proceso que demoró más tiempo en realizarse, debido a las iteraciones realizadas sobre el mismo y por la búsqueda de obtener el mejor modelo predictivo posible. Luego de probar con diferentes modelos de *machine learning* se obtuvo que el mejor modelo resultó ser un Árbol de Decisión de Regresión. Este entregó los mejores estadísticos de todos los modelos probados, sin embargo, se encuentra un poco lejos de los resultados óptimos que se podrían tener. Se cree que esto se debió a la poca cantidad de registros con los que se contaban y que a futuro podría mejorarse en gran medida el modelo.

Ya obtenido el modelo predilecto, se hizo un análisis de las variables que componen al mismo y cómo estas afectaban al *start up time.* A modo de recomendación se brindaron una serie de directrices que serían de utilidad para la farmacéutica a la hora de planear un nuevo ensayo clínico.

A modo resolutivo, se hizo una conclusión que resumió el desarrollo del trabajo, cuál fue el modelo elegido y como se podría continuar mejorándolo.

Contexto

Empresa ABC es una de las empresas farmacéuticas más importantes a nivel mundial. Cuenta con una de las mejores reputación de la industria en cuestión. Hoy en día, cuenta con alrededor de 81000 empleados a lo largo del mundo. Siendo Basilea en Suiza su sede central, Empresa ABC nace de la fusión de dos empresas,

Ciba-Geigy y Sandoz. Dicha fusión sucede en 1996 y en su momento representó la fusión de empresas más grande del mundo.

Luego de esta fusión, la empresa decide enfocarse en el negocio del cuidado de la salud. En 2005, los productos enfocados a la salud representan el 90% de los ingresos de la compañía.

Empresa ABC cuenta con una estructura de división de negocio muy marcada. Todas funcionan de manera independiente, pero esto no determina que no puedan complementarse entre ellas. Estas son:

- Fármacos: componen el 60% de las ventas de la compañía. Busca el desarrollo de nuevos medicamentos y productos innovadores. Es aquí donde se realizan la mayoría de los ensayos clínicos y hacia dónde irá dirigido el desarrollo del proyecto.
- Genéricos: representan el 14% de las ventas de la compañía.
- Vacunas & Diagnósticos: es el 4% de las ventas de la compañía. Se enfoca en la prevención e impacto de enfermedades de transmisión humanitaria de gran escala.
- Salud del Consumidor: son el 22% de las ventas de la compañía. Se enfoca en productos que generen un rápido retorno.

En el año 2022, Empresa ABC se posicionó como una de las principales compañías farmacéuticas en cuanto a ingresos totales. Claro indicio de la importancia de la misma en el cuadro macroeconómico de la salud a nivel mundial. En los últimos años, la empresa se ha encontrado siempre en los primeros puestos de los rankings de importancia mundial.

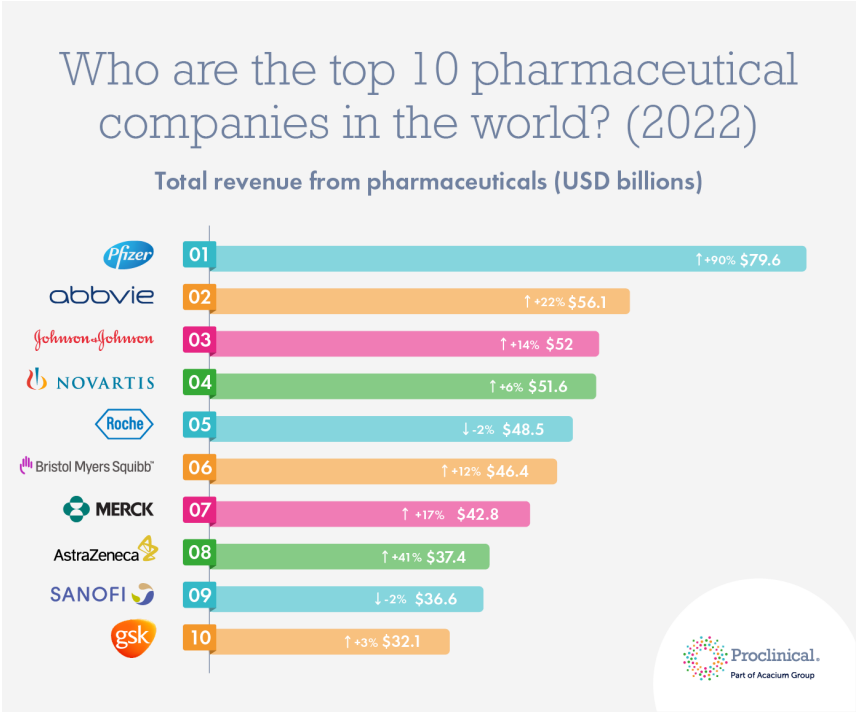


Figura 1. Ingresos de farmacéuticas en 2022.

Como puede verse tanto en la Figura 1 y la Figura 2, Empresa ABC estuvo presente en las empresas farmacéuticas de mayor magnitud, aprovechando momentos claves del panorama mundial como lo fue la pandemia del Covid-19, para seguir revalorizando su impronta. Los ingresos de la compañía se mantuvieron alrededor de los mismos valores en estos últimos años.

Sin embargo, también se observa que el gasto de la compañía en Investigación y Desarrollo no se acerca a los valores de sus competidores. No se encuentra entre los que menos gastan pero tampoco está entre los que más lo hacen. Esto es algo que las autoridades de Empresa ABC tienen muy en cuenta y lo controlan con mucha rigurosidad. Ya que este no es el foco principal de la operativa de Empresa ABC, pero cumple la función de soporte para otras aristas de la compañía.

Figura 2. Ranking de ventas y gastos de farmacéuticas en 2021.

¿Qué es un ensayo clínico y su *start-up time*?

Ninguna medicación ni ningún tratamiento es lanzado al mercado sin un previo control y estudio. Como la mayoría de las empresas farmacéuticas, Empresa ABC lleva a cabo distintos ensayos clínicos para las pruebas de sus medicamentos y tratamientos. Con esto en mente, surge la pregunta: ¿qué es un ensayo clínico?.

Un ensayo clínico es una evaluación experimental controlada de un fármaco o un tratamiento sobre una población de pacientes. Esto se realiza para conocer y comparar los efectos del fármaco o tratamiento, que busca ser introducido al mercado, con respecto a otros ya existentes. Se quiere observar qué diferencias sustanciales positivas presentan los nuevos productos en cuestión.

Previo al comienzo de un ensayo clínico se da una etapa conocida como etapa pre-clínica, que se corresponde con el descubrimiento de la droga a probar y

la correcta selección entre todas las moléculas de sus componentes. Una vez ya determinadas las moléculas y la droga en cuestión se da comienzo al ensayo clínico en sí. Es de estimación y aceptación general que un ensayo clínico tiene una duración de alrededor de 12 años. En total, se estima que el costo de desarrollo de una droga y sus ensayos clínicos es de 2.6 billones de dólares. Los ensayos clínicos cuentan con 3 fases. En donde en cada fase se va escalando gradualmente la cantidad de pacientes en los cuales se prueba la droga. Cada ensayo clínico se lleva a cabo en distintos sitios (hospitales o centros especializados para ensayos) en simultáneo para mayor validez estadísticas a la hora de extraer conclusiones.

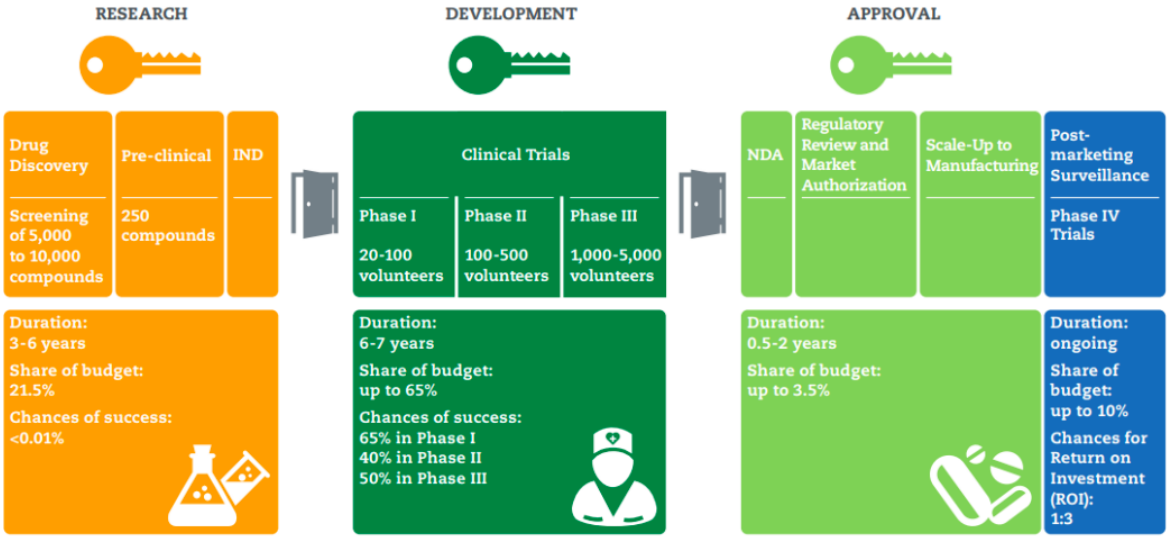


Figura 3. Proceso de un ensayo clínico.

Lo que se conoce como *start up time* es el tiempo que se tarda en comenzar un ensayo clínico, en cualquiera de sus fases. Una vez ya terminada la etapa pre-clinica Este comienza una vez que llega el protocolo final enviado por la farmacéutica y controlado por las instituciones sanitarias correspondientes. Y finaliza con la visita de inicio al centro de investigación. Este ciclo puede observarse en la Figura 4 y el detalle de cada uno de los pasos que lo componen se encuentran en la Tabla 1 del Anexo. En charlas con representantes de Empresa ABC, fue comentado que este tiempo es muy variable y se calcula en alrededor de unos 10 meses para lo que es la industria farmacéutica en general. Sin embargo, Empresa ABC se encuentra por debajo de este valor. Cuenta hoy en día, con un promedio de 8 meses de tiempo de *start up*. Vale destacar, que esto varía dependiendo del sitio en donde se desarrollará, qué persona estará a cargo del ensayo, cuántas personas participarán del estudio y varias otras variables que afectan a la determinación del tiempo de *start up*.

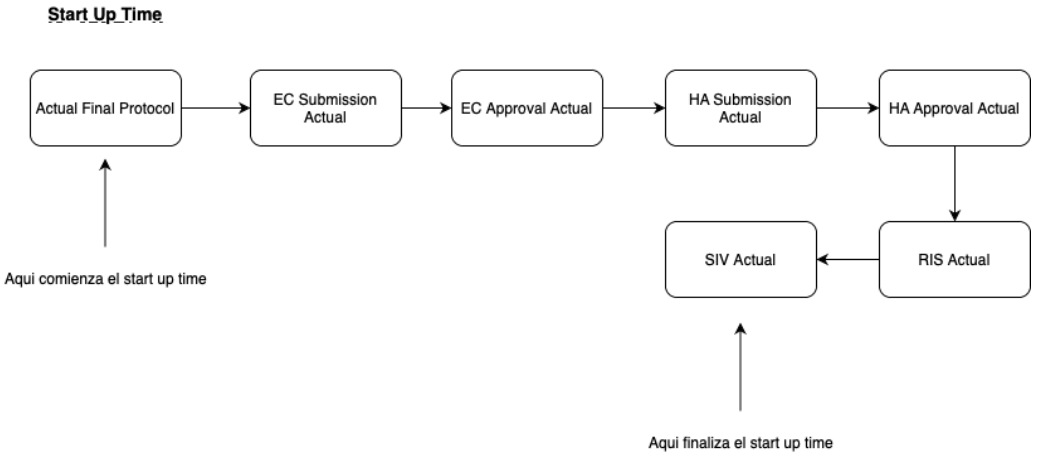


Figura 3. Flujo del *start up time.*

Objetivo del Proyecto

Con el desarrollo de este proyecto se buscó poder predecir el tiempo de *start up* de un ensayo clínico y entender que variables tenían más peso a la hora del cálculo del *start up*. Esto en base a las diversas características que presenta cada ensayo; ya sean desde la ubicación del mismo, la organización con la cual se desarrolla el ensayo, qué tipo de patología se trata, entre varias otras variables. A su vez, se buscó poder brindar ciertas recomendaciones para llevar a cabo elecciones que permitan resultar en tiempos de *start up* menores. Ya sea desde el mejor sitio para llevar a cabo un ensayo o qué institución se encargue de supervisar el mismo. Lo mencionado representa una inversión no solo monetaria sino de tiempo que es crucial para la compañía. Esto debido a los tiempos que se manejan en el sector farmacéutico, ya que el tiempo es oro para poder estar por delante de la competencia. Esto sirve no solo para el desarrollo más veloz de una droga, sino que también sirve como indicador de eficiencia a la hora de buscar inversiones para el desarrollo del tratamiento. El poder determinar y evaluar cómo poder reducir el *start up time* es un beneficio muy importante, ya que permite llegar al mercado de la venta de la droga antes que el resto de la competencia.

Éxito del Proyecto

Todo proyecto debe tener un punto de éxito al cual apuntar. Este sirve como un driver que orienta y da forma al desarrollo del mismo. En esta situación, el éxito del proyecto es que Empresa ABC gane tiempo operativo cuando se busque arrancar un ensayo clínico. Para esto se pusieron esfuerzos en lograr un modelo que pueda determinar correctamente el *start up time.* Ya con este modelo, se

pudieron realizar recomendaciones para poder lograr reducir este *start up time.* Esto no solo será una ayuda en materia de tiempo, sino que también el campo monetario va de la mano con el campo del tiempo. A mayor tiempo que se tarda en comenzar un ensayo, los costos van en aumento, ya que las ineficiencias tienen un costo alto. El éxito temporal tendrá impacto directo en el éxito monetario.

KPI a Impactar

El KPI por excelencia que se buscó impactar fue el del tiempo de *start up.* Principalmente se quiso investigar y analizar aquellas razones por las cuales se reduce el tiempo de arranque de un ensayo clínico. Observar qué medidas podrían tomarse y opciones a elegir que permitiese disminuir este tiempo.

A su vez, el impacto de este KPI tiene consecuencias directas sobre otras áreas de la organización. En cuanto a lo monetario, al mejorar este KPI implicaría un impacto monetario debido a la llegada antes que el resto de las farmacéuticas al mercado con un producto. También se mejoran ciertas cuestiones organizativas en cuanto a la preparación de los ensayos clínicos, ya que saber de antemano qué medidas tomar en función de lo que se quiera obtener agilizar proactivamente el flujo de trabajo.

¿Cómo se abordó el problema?

Para el abordaje del problema se buscó desarrollar y ajustar un modelo de *machine learning* o aprendizaje automático que permitiera predecir el tiempo de *start up* de los diferentes ensayos . En primer lugar, se separó en tres partes el conjunto de datos, sean los conjuntos de entrenamiento, validación y testeo. Estos conjuntos se conformaron por el 60%, 20% y 20% respectivamente del conjunto de datos para poder continuar con el análisis.

En segundo lugar, se desarrolló el modelo encargado de predecir el valor de *start up* de los estudios clínicos, que es la variable de interés para el negocio. Se decidió utilizar algoritmos de regresión, ya que estos son los que permiten predecir los valores requeridos en este caso. Algunos de los algoritmos que se probaron fueron regresiones lineales , árboles de decisión y un modelo de *support vector* enfocado en la regresión. Para la construcción del mismo se utilizó el conjunto de entrenamiento y se modificaron los diferentes hiperparámetros de cada tipo de algoritmo con el conjunto de validación. Una vez elegida la mejor combinación de hiperparámetros para cada tipo de algoritmo diferente, se compararon los mejores modelos, utilizando las métricas de performance elegidas. Y entonces se selecciona el modelo ganador entre todos los modelos de la experimentación utilizando el conjunto de validación.

Una vez elegido el modelo ganador se lo procedió a correr con el set de testeo para obtener los estadísticos del mismo y a su vez se le realizaron diversos análisis tanto al modelo en sí como al resto de las variables predictoras. Se buscará que valores de las variables predictoras son aquellas que permiten obtener tiempos de *start up* más bajos. Y así, poder recomendar ciertas pautas a la farmacéutica a la hora de la elección de ciertas variables.

Entregables y Outputs

Al finalizar el proyecto, se habrán obtenido diversos modelos predictivos en cuanto al cálculo del tiempo de *start up* de un ensayo clínico. Se entregará uno solo de estos modelos y será aquel cuyas métricas resulten mejores y convenientes con la situación en cuestión. Este modelo será brindado en formato de código python para que la farmacéutica pueda modificarlo a su parecer y ajustarlo a sus necesidades. A su vez, se entregará este mismo informe, que contiene los resultados obtenidos y las recomendaciones brindadas para lograr menores tiempos de *start up*.

Estado del Arte

Como parte de la investigación del estado del arte, se buscaron y analizaron diversas fuentes académicas que permitieron tener un punto de partida para el desarrollo y entendimiento de los modelos de predicción o *machine learning*.

Se destacó el proyecto titulado *Comparison of Machine Learning Algorithms for Software Project Time Prediction,* realizado por WanJiang Han , LiXin Jiang , TianBo Lu y XiaoYan Zhang para la School of Software Engineering, *Beijing University of Posts and Telecommunication*. En la investigación se compararon siete modelos de *machine learning* para ver como predecían el tiempo de desarrollo de un proyecto de software. Luego, midieron su desempeño con diferentes métricas. Una de ellas viene a ser la magnitud del error relativo o MRE por sus siglas en inglés. Como conclusión llegaron que el mejor modelo era el *Gaussian Process*. Pero para este proyecto se destaca también el desempeño de los otros dos mejores modelos, esto incluye una red neuronal de Multilayer Perceptron y el REPTree.

Por más de que no se hayan utilizado los modelos que el proyecto mencionado presentaba, estos sirvieron como guía para la elección de los que sí se utilizaron. Gracias a ellos, se pudo tener mayor conocimiento sobre cómo encarar la parte técnica del proyecto, qué métricas y qué metodologías usar.

Plan de Trabajo

Para la realización del proyecto se contó con un tiempo estimado de alrededor de 3 meses. Como puede observarse en la Figura 5, se subdividió todo el proceso en distintas etapas con sus respectivas fechas a cumplir. Se consideró un error de ± 3 días para cada etapa descrita. Ciertas etapas están superpuestas unas con otras para poder permitir una mirada más holística del proceso y que la toma de decisiones se volviese iterativa y sinérgica entre las distintas etapas.

Los encargados del proyecto trabajaron a la par en ciertas etapas mientras que en otras se dividieron tareas para un progreso más continuo del mismo. Esto se definió cuando se comenzaba una etapa.

Las comunicaciones con los representantes de Empresa ABC fueron constantes, ya que fue una buena forma de obtener información para el correcto desarrollo del proyecto y para poder llegar a nuevas conclusiones o *insights* sobre cómo seguir abordando el trabajo. Las comunicaciones mediante mensaje y correo electrónico ocurrían con una frecuencia casi semanal.

Primeramente, se hizo un reconocimiento del dataset y del uso que se le da al mismo en la organización. Observando los datos que presenta, la forma en que los presenta y todas las aristas de información que esconden estos datos.

Luego, fue hora de realizar el data wrangling y cleaning, donde se trataron los datos para poder utilizarlos de manera más optimizada. A su vez, fue necesario hacer una limpieza de datos para poder obtener los datos más pulcros posibles. Estos fueron pasados como inputs para el modelo predictivo.

Luego, se comenzó con la construcción de los modelos utilizando únicamente los conjuntos de entrenamiento y validación que permiten customizar los diferentes tipos de algoritmos de machine learning para que se ajusten de la mejor manera a los datos y al objetivo de negocio del proyecto. De esta etapa se eligió al modelo “ganador” y luego se midió su performance final con el conjunto de testeo.

Además, se realizó una revisión completa de todos los puntos hechos para verificar que se estén cumpliendo los objetivos planteados. Esto sirvió para observar cómo se encontraba el grupo con respecto a los objetivos propuestos y analizar todo el camino recorrido y cómo continuar con el mismo.

A continuación, se eligió el mejor modelo por su desempeño y su ajuste al objetivo de negocio.

Posteriormente, se realizó un informe exponiendo los hallazgos encontrados a lo largo del desarrollo del proyecto para dejar constancia de lo que se analizó, probó y que funcionó finalmente.

En último lugar, se llevará a cabo la presentación al board de negocio de Empresa ABC para que puedan comprender qué fue lo que se realizó y cómo pueden usar los hallazgos para la toma de decisión del negocio y llegar al objetivo que se planteó en un principio.

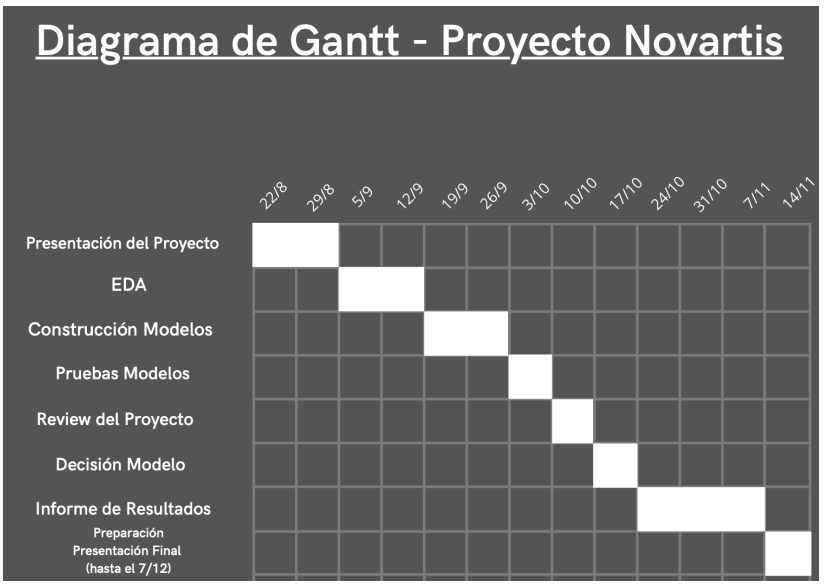


Figura 4. Diagrama de Gantt

Herramientas Utilizadas

En adición, para el desarrollo de los modelos de *machine learning* del proyecto se utilizó el lenguaje de programación *Python* soportado por el editor de texto *Jupyter Notebook*. El software ML Flow de control de modelos de *machine learning* fue usado y aportó un gran valor organizativo a la hora de analizar los modelos desarrollados. Todas las anotaciones que el grupo realizó se anotaron en un correspondiente documento de Google Docs para que todos los participantes puedan escribir sus comentarios, ideas y propuestas en cuanto al proyecto.

Fuentes y Tipos de Datos

Los datos sobre los ensayos clínicos, sus tiempos y características fueron obtenidos mediante un archivo *.xlsx* brindados por contactos dentro de Empresa ABC. Los estudios clínicos que allí se detallan fueron realizados en tres países diferentes: Chile, Argentina y Estados Unidos. A su vez de estos se contaba información de la ciudad y la provincia o estado en la cual fueron desarrollados o se están llevando a cabo. El detalle del archivo en cuestión posee su propia sección en este informe.

El correspondiente archivo contaba con datos estructurados de una variada cantidad de los mismos. Fechas, datos categóricos nominales y descripciones, tanto como valores numéricos representan en su totalidad los tipos de datos presentes en el dataset. Para ejemplificar cómo están presentes estos tipos de datos en el dataset en cuestión está presente la columna “*Indication*”, que explica cuál es la enfermedad sobre la cual se está tratando en ese ensayo clínico.

¿Con qué datos se contaba?

Como bien fue mencionado previamente, se contaba con un archivo *.xlsx* otorgado por los representantes de Empresa ABC. Este archivo poseía información sobre distintos estudios clínicos, ya sean activos o no. Los datos se presentaban de manera estructurada y bastantes claros en cuanto a su comprensión luego de un avistaje general al archivo. Se contaban con datos de ensayos clínicos desde el año 2008 hasta el 2022. Una gran variedad de columnas describían en detalle los tiempos, locaciones y encargados de los ensayos clínicos, entre varias otras columnas que a su vez hacían referencia a indicaciones médicas sobre el asunto. A pesar de que existían nomenclaturas propias del sector médico y de Empresa ABC en algunos de los nombres de las variables, los representantes de la farmacéutica comentaron en distintas reuniones varias de las definiciones de los mismos para que puedan comprenderse mejor esas variables.

La variable del objeto del proyecto es el *start up time,* la cual no aparecía explícitamente explicada dentro del dataset. Sino que hubo que llevar a cabo diferentes transformaciones de datos para poder obtenerla, que serán explicados en detalle posteriormente.

Preguntas que surgieron

Luego de haber realizado una mirada general y abarcativa del dataset en cuestión, surgieron ciertas preguntas que sirvieron de orientadoras en el desarrollo del análisis exploratorio y en la limpieza y selección de variables del dataset. Preguntas tales como: ¿cuáles de todas las variables son realmente importantes para el *scope* de este análisis? ¿Es necesario omitir variables? ¿Agregar nuevas? ¿Hay suficientes variables y son explicativas, como para lograr que cualquier persona ajena al equipo entienda lo que es iniciar un ensayo clínico? ¿El dataset presenta un entendimiento claro del problema en cuestión? ¿Cómo llevar a cabo el cálculo del *start up time?* ¿Se ven correlaciones entre variables consideradas relevantes para los investigadores y el start up time? ¿Estas correlaciones ameritan el análisis que se hará?

Todas estas preguntas fueron de gran ayuda a la hora de pensar cómo encarar y abordar el análisis del dataset.

Análisis Exploratorio de Datos

A continuación se detalla el análisis realizado para conocer más en profundidad la base de datos a la cual se tratará en busca de lograr el objetivo propuesto.

El Dataset

El dataset cuenta con un total de 12253 filas y 85 columnas, como puede verse en la Figura 5. Es decir, una cantidad bastante considerable de variables para analizar.

```
dfs_shape= df_s.shape
print("El dataset tiene {0} filas y {1} columnas.".format(dfs_shape[0],dfs_shape[1]))
✓ 0.8s
El dataset tiene 12253 filas y 29 columnas.
```

Figura 5. Estructura del dataset.

Luego de un análisis con foco en aquellas variables que son realmente relevantes para el scope del proyecto y del negocio, se decidieron eliminar varias variables que en sí no aportaban demasiada información. Ya que estas o presentaban valores poco relevantes o no tenían sentido dado el contexto en el cual se encuadra el proyecto. Muchas de ellas eran variables de fechas, que luego de charlarlo con los representantes de la farmacéutica, se notó que no tenían impacto o interés alguno para el objetivo del proyecto. Eran fechas por fuera del scope del *start up time*. Es así que el dataset pasó a contar con un total de 12253 filas y 29 columnas, como puede verse en la Figura 6.

```
dfs_shape= df_s.shape
print("El dataset tiene {0} filas y {1} columnas.".format(dfs_shape[0],dfs_shape[1]))
✓ 0.8s
El dataset tiene 12253 filas y 29 columnas.
```

Figura 6. Estructura del dataset luego de eliminar variables.

De esta forma se permitió un manejo más simple y dinámico de los datos. Pudiendo comprender con mayor amplitud cada campo y a que se refería cada uno. En cuanto a las variables en su momento presentes en el dataset, se tenían las siguientes:

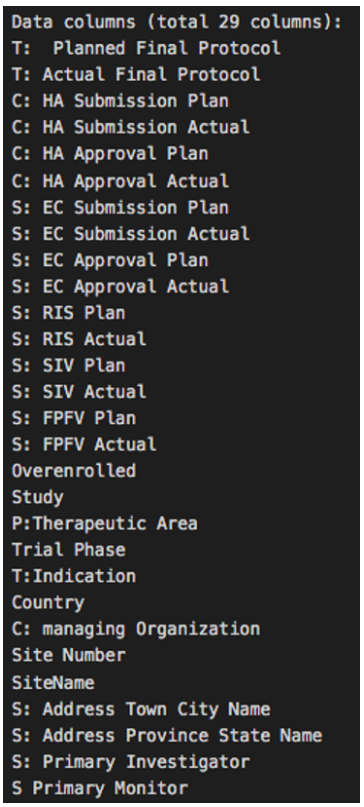


Figura 7. Columnas del dataset.

Dentro de estas columnas se pueden ver variables como "*Country*" que determina en qué país se está llevando a cabo el ensayo clínico. A su vez se establece la ciudad en la que se está llevando a cabo y a que provincia pertenece ("*Address Town City Name*", "*Address Province State Name*"). La variable "*Overenrolled*" indica si en el ensayo se tuvieron más pacientes de los que se había planeado en principio. "*P:Therapeutic Area*" corresponde a la variable que al área médica a la cual se le atribuye el ensayo clínico. "*Study*" es la variable que indica el ID del ensayo clínico en sí. La "*Trial Phase*" indica en qué fase del ensayo clínico se encuentra el estudio. La "*T:Indication*" es el tratamiento/enfermedad que se está estudiando. "*C:Managing Organization*" es la organización que se encuentra a cargo de la supervisión del ensayo clínico. "*Site Number*" es el número asociado al centro de investigación, y "*SiteName*" es el nombre del mismo. "*S: Primary Investigator*" es el encargado principal de llevar a cabo el ensayo clínico. "*S: Primary Monitor*" es aquel individuo que se encarga de la supervisión del ensayo clínico. Todas las variables que son fechas se encuentran detalladas en la Tabla I del Anexo. Por cuestiones de comodidad, las variables fueron renombradas como se muestra a continuación.

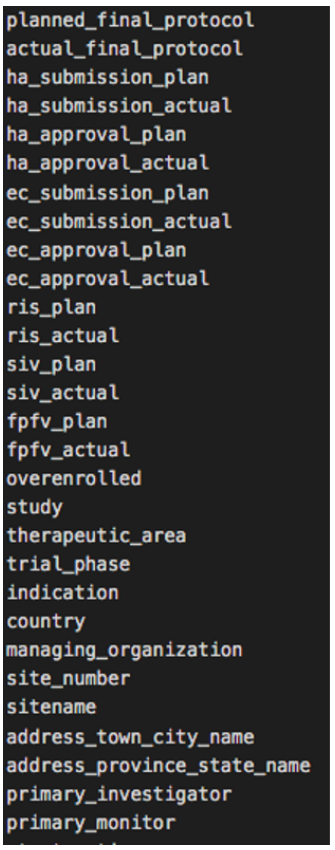


Figura 8. Variables Renombradas.

En las variables en formato fecha, se tienen dos variantes de cada variable. Por un lado están las que son las fechas “planeadas”, estimadas según la farmacéutica que hace referencia a las fecha en que calculan que se lleve a cabo el proceso que indica la variable. Estas variables “planeadas” se arman mediante datos históricos de cuánto tiempo conlleva en promedio según ciertas características cada proceso del ensayo en cuestión. Mientras que las que son “actual”, son las fechas reales en que ocurrió el proceso correspondiente. Luego de analizarlas, se decidió omitir las variantes “plan” de las variables “*ha_submission*”, “*ha_approval*”, “*ec_submission*”, “*ec_approval*”, “*ris*”, “*siv*” y “fpfv”. Lo que sí se realizó con estas variables fue armar una nueva variable llamada “*startup_time_plan*” que servirá a futuro cuando se quiera comparar esta variable con el *start up time* real de cada uno de los ensayos. Cabe destacar que la variable de *start up time* no venía calculada dentro del dataset. Esta surge de hacer la diferencia entre las variables “*actual_final_protocol*” con “*siv_actual*”. El mismo cálculo se realizó para calcular “*startup_time_plan*”. A su vez, también se eliminaron las variables de fechas del dataset ya que estas son parte de la variable objetivo y en el caso de utilizarlas como variables predictoras, sesgarian los resultados y desarrollo del proyecto. Por lo que las variables “*ha_submission*”, “*ha_approval*”, “*ec_submission*”, “*ec_approval*”, “*ris*”, “*siv*” y “fpfv”, tanto de “actual” como de “planned” , bien comentado previamente, fueron eliminadas del dataset.

Por lo tanto las variables que conformaban el dataset para esta altura eran las que pueden verse en la Figura 9.

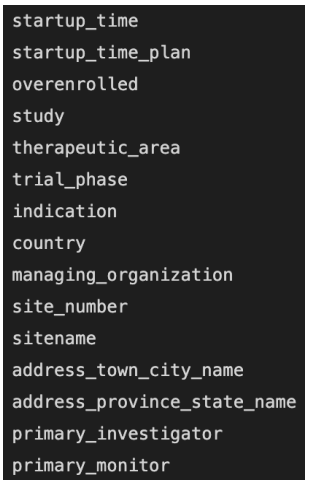


Figura 9. Variables del dataset.

De la misma manera, se observa que existen una gran cantidad de registros que cuentan con valores de *start up time* negativos o iguales a 0. Esto puede ocurrir debido a un error en los ingresos manuales de los datos o a que el ensayo se encuentra transitando su tiempo de *start up* en la fase en la que se encuentren. Estos valores no eran de utilidad a la hora de realizar los modelos de *machine learning* a posteriori. Por lo que estratégicamente se decidió continuar con el análisis exploratorio con aquellos registros que sí serían de utilidad. De la misma manera, luego de charlas con Empresa ABC, se nos fue comentado y luego cerciorado en el dataset, de que alrededor del 90% de los ensayos presentes ocurrían en Estados Unidos, ya que de Chile y Argentina hace muy poco tiempo que se había comenzado a recabar información. Por lo que se decidió y acordó hacer el análisis para solo dentro de Estados Unidos. De esta manera, luego de filtrar la base de datos por aquellos registros en donde la variable objetivo resultó ser mayor a 0 y se llevasen a cabo en Estados Unidos, se obtuvo un dataset con 4507 filas.

Con respecto a los missings, no se encontraban en el dataset valores del tipo *null* per se, excepto en la variable "*trial_phase*", que si tenía valores nulos pero en una cantidad bastante pequeñas de registros. Solo 36 registros no presentaban valores, como puede observarse en la Figura 10 a continuación.

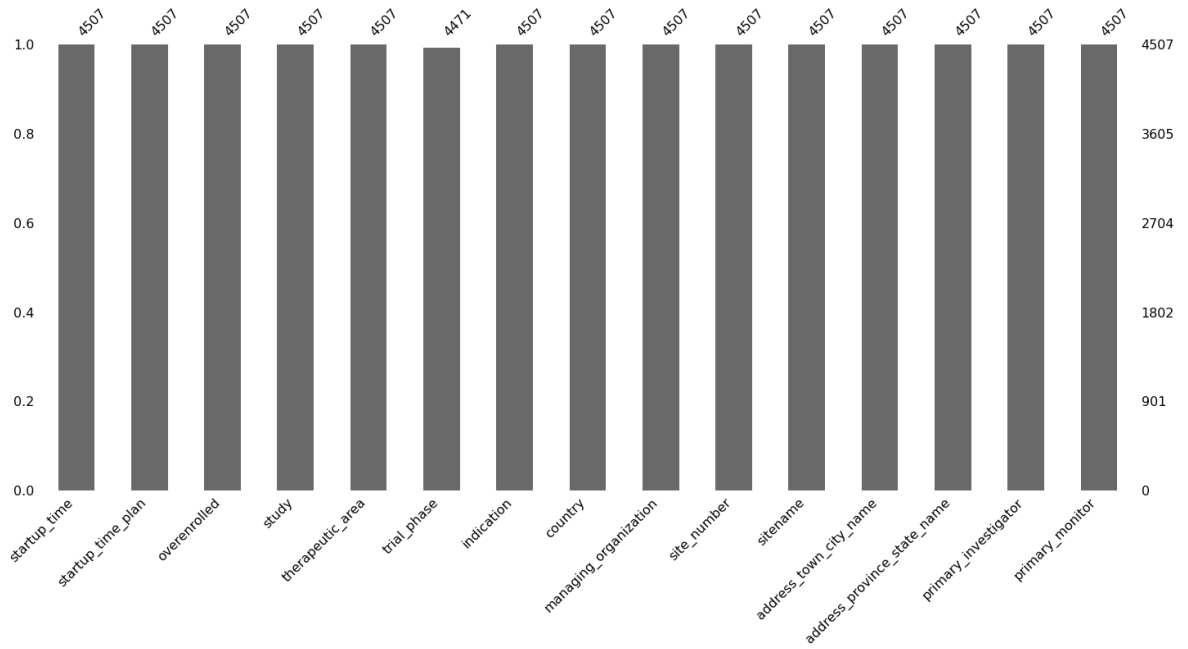


Figura 10. Missings del dataset.

Una vez realizado esto se comenzó a investigar más en detalle qué valores presentaban y de qué manera se mostraban los valores de los registros de cada una de las variables.

Análisis Univariado

Comenzando con la variable "*indication*", que indicaba el término médico para el cual se estaba realizando un ensayo clínico, se vio que había muchos registros donde había valores separados por un ";". Esto ocurría ya que el primer valor que aparecía correspondía a la indicación principal del estudio, y la segunda, al secundario. Se decidió solo mantener la indicación primaria y no la secundaria, debido a que esta era la indicación más general y la cual describe con mayor envoltura el foco que tomaba el ensayo clínico. La indicación secundaria era más específica y no tan relevante para el desarrollo del proyecto. En la Figura 11 pueden verse algunos de los valores que presentaba esta variable primaria de "*indication".* Son en total 132 opciones de indicaciones que aparecen en su totalidad .

Multiple sclerosis	0.066563
Breast cancer	0.053916
Breast cancer metastatic	0.052807
Cardiovascular event prophylaxis	0.049257
Non-small cell lung cancer	0.044154

Figura 11. Porcentajes de frecuencia de una determinada cantidad de valores de la variable "*indication*".

En la Figura 12 se muestran algunas de las tantas indicaciones que aparecían dentro del *dataset*. Se notó una clara presencia de indicaciones relacionadas al cáncer dentro de los puestos superiores.

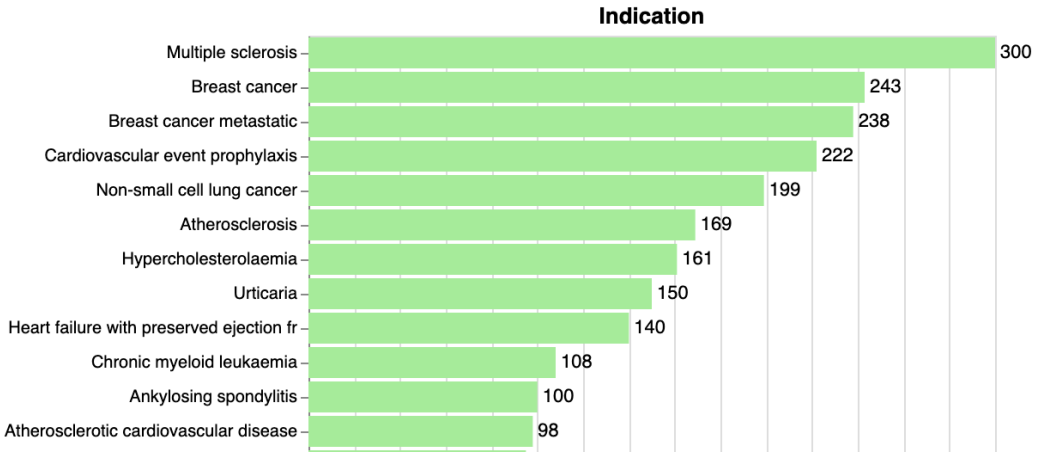


Figura 12. Indicaciones de los ensayos clínicos.

Dentro de la variable "*primary_monitor",* que denotaba a la persona responsable de monitorear un ensayo clínico, se encontraron una gran cantidad de valores en forma de "-". Un 26% de los valores presentaban este símbolo, lo que se consideró como que esos valores eran faltantes. Esto puede verse en la Figura 13. A su vez, el siguiente valor que se presenta con mayor frecuencia solo lo hace en un 0.6% de los registros. Lo que indica que había una gran variedad de posibles valores que podía tomar esta variable. Esto hubiera complicado mucho todo lo que tenía que ver con la codificación de esta variable para poder utilizarla en los modelos de *machine learning*. Fue justamente esta variabilidad de posibilidades lo que decidió el prescindir de esta variable para el resto del análisis, ya que no sería de gran utilidad a futuro.

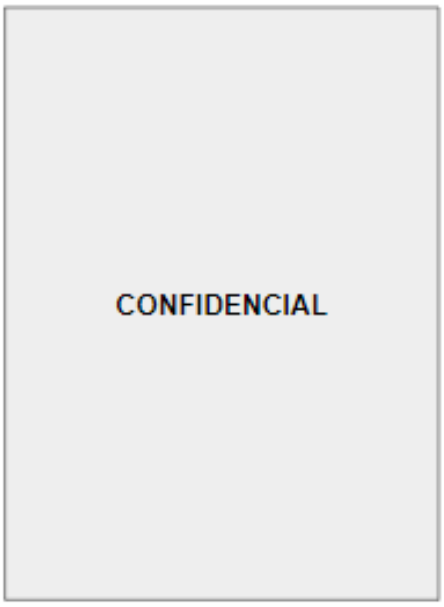


Figura 13. Porcentajes de frecuencia de una determinada cantidad de valores de la variable “primary_monitor”.

En la misma senda, en la variable “*primary_investigator*”, siendo esta la persona encargada de la investigación del ensayo clínico, se tenían 3220 opciones de valores. Esto demostraba que esta variable tenía un valor de uso insignificante, debido a la enorme cardinalidad que presentaba. En la Figura 14 puede observarse que el investigador que aparece más reiteradamente solo lo hace en un 2% de los registros. Es por esto que también se decidió prescindir de esta variable.

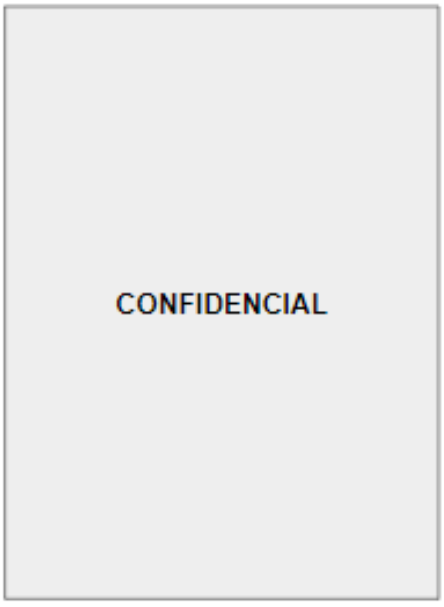


Figura 14. Porcentajes de frecuencia de una determinada cantidad de valores de la variable “primary_investigator.”

La variable “*address_province_state_name*” presentaba 109 valores posibles que podía tomar. Esta indicaba el estado en donde se llevó a cabo el ensayo clínico. Un porcentaje muy pequeño de los valores tomaban la opción de ‘-’, que fueron considerados como valores faltantes. Estos valores faltantes fueron reemplazados por “*Other*”. A pesar de su gran cardinalidad, esta variable se mantuvo porque posteriormente fue encodificada en distintos grupos para disminuir esta cardinalidad. En la Figura 15 pueden observarse los principales valores que toma la variable con sus porcentajes.

```
CA    0.118260
FL    0.105170
TX    0.104948
NY    0.049700
OH    0.036388
MA    0.033725
IL    0.031285
NC    0.030175
MD    0.028178
MI    0.027291
```

Figura 15. Porcentajes de frecuencia de una determinada cantidad de valores de la variable “*address_province_state_name*.”

La Figura 16 se encarga de mostrar algunos de los tantos estados de EEUU en donde se llevaron a cabo ensayos clínicos. Las ciudades más importantes del país hacen su lugar entre los primeros puestos. California, Florida, Texas y Nueva York fueron los estados en donde se desarrollaron la mayoría de los ensayos. Hubo una clara concentración de los ensayos en las grandes urbes de los Estados Unidos.

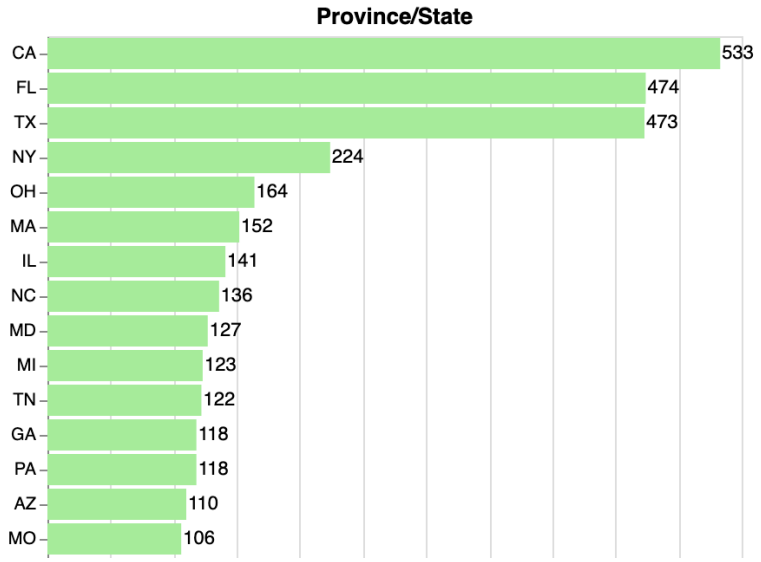


Figura 16. Estados donde se desarrollaron ensayos clínicos.

La variable “*address_town_city_name*” presenta una cantidad de diversidad de registros similar a “*address_province_state_name”*. Lo que tiene sentido, ya que ambas variables están muy relacionadas. Esta variable indicaba la ciudad en la que se llevó a cabo el ensayo clínico. Esta relación casi simbiótica se confirma en la Figura 17. Estos valores faltantes son reemplazados por “*Other*”.

CA	0.118260
FL	0.106501
TX	0.105392
NY	0.049922
OH	0.036388
MA	0.033725
IL	0.031285
NC	0.030175
MD	0.028178
MI	0.027291

Figura 17. Porcentajes de frecuencia de una determinada cantidad de valores de la variable “*address_town_city_name*.”

La Figura 18 no hace otra cosa que confirmar lo comentado en el párrafo previo. Se repiten los mismos valores y las cantidades eran parecidas a los de la variable *“address_province_state_name”.*

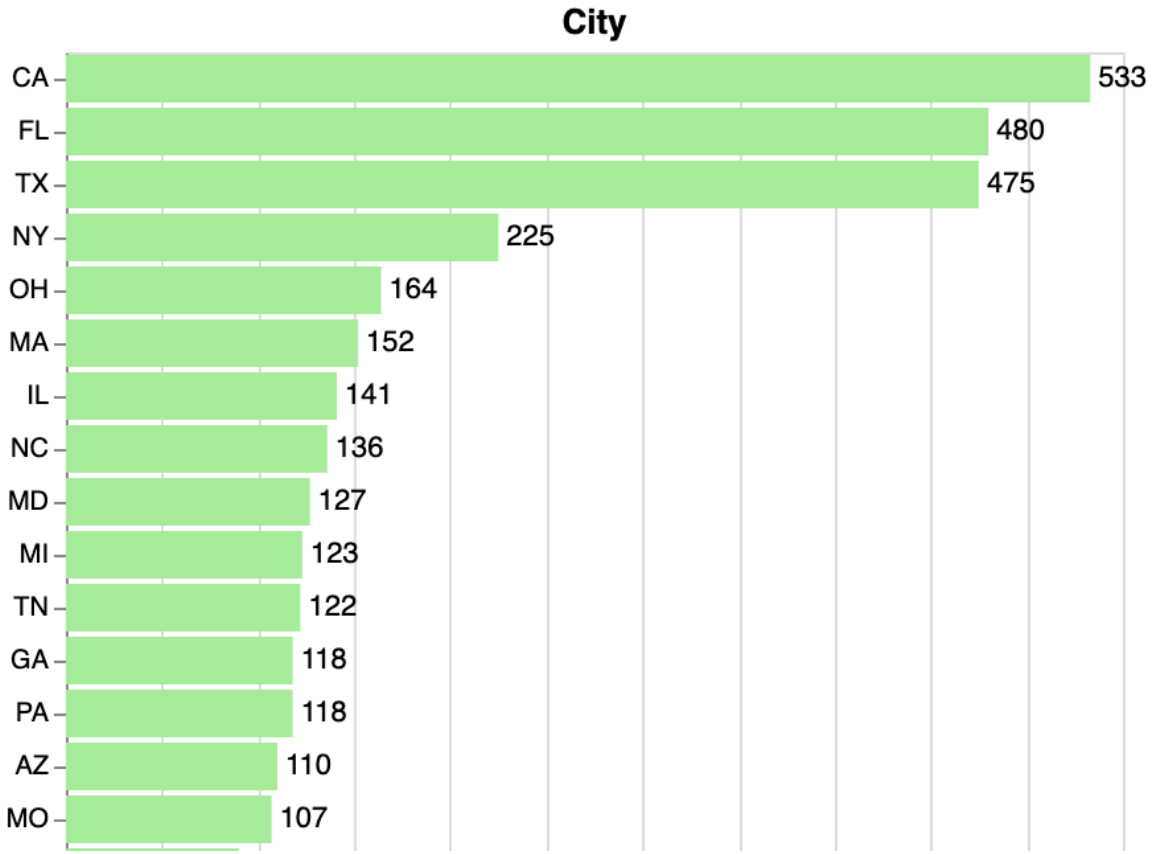


Figura 16. Ciudades donde se desarrollaron ensayos clínicos.

La variable "*study"* presentaba 259 valores que podía tomar. Sin embargo, luego de analizarlo se decidió prescindir de esta variable ya que en sí, no aportaba nada de valor al desarrollo del proyecto. Solo eran las nomenclaturas con las que Empresa ABC llamaba a sus estudios. En la Figura 17 puede observarse la diversidad de posibilidades con sus porcentajes de aparición.

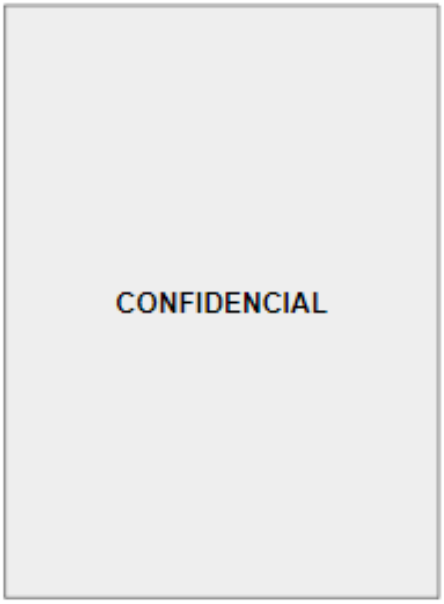


Figura 17. Porcentajes de frecuencia de una determinada cantidad de valores de la variable "*study*."

La variable "*overenrolled"* indicaba si el ensayo en cuestión había superado el cupo estimado de pacientes con los cuales contar. Las opciones que se presentaban en el dataset eran "*yes*" o "*no*". La Figura 18 muestra que 78% de los ensayos no fueron "*overenrolleados*" y un 22% sí lo fueron.

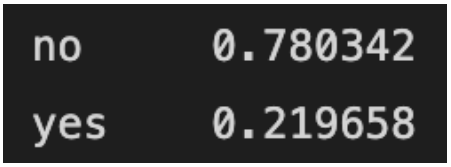


Figura 18. Porcentajes de frecuencia de una determinada cantidad de valores de la variable "*overenrolled*."

La Figura 19 muestra los porcentajes previamente mencionados en sus valores absolutos. La mayoría de los ensayos, 3517 para ser exactos, no contaron con sobre enrolamiento de pacientes.

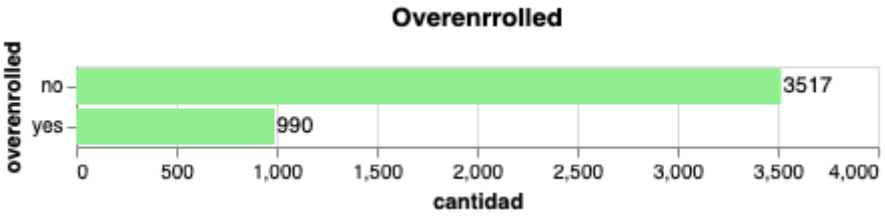


Figura 19. Cantidad de ensayos con y sin "*overenrrolled*".

Tanto la variable "*sitename*" y "*sitenumber*" no entregaban demasiada información para el proyecto. La variable "*sitenumber*" indicaba el nombre del centro donde se llevó a cabo el ensayo clínico, y "*sitename*" era un número asociado al mismo. Debido a la enormidad de posibilidades que "sitename" tomaba, más de 3000 posibilidades, y a que adicionalmente, no hubiera permitido direccionar el proyecto hacia el rumbo que el trabajo de equipo estaba buscando, se decide eliminar ambas variables.

La variable "trial_phase", como fue comentado previamente, fue la única variable que presentaba *missings.* Solamente eran 36 registros, por lo que se procedió a imputar con el valor "other". En la Figura 20 puede observarse que alrededor del 70% pertenecían a la Fase III, secundado por la Fase II. Luego, los valores porcentuales del resto de las fases eran mucho más bajos.

```
III      0.684491
II       0.216552
I        0.044819
IV       0.043044
other    0.007988
I/II     0.003106
```

Figura 20. Porcentajes de frecuencia de una determinada cantidad de valores de la variable "*trial_phase*".

La Figura 21 muestra claramente como la cantidad de registros de ensayos en fase III supera ampliamente al resto de las fases. Esto pudo haberse debido a que la fase III es la que mayor tiempo de duración posee, ya que es la que mayor cantidad de pacientes conlleva. Esto implicaría mayores cuestiones organizativas y de ahí su demora.

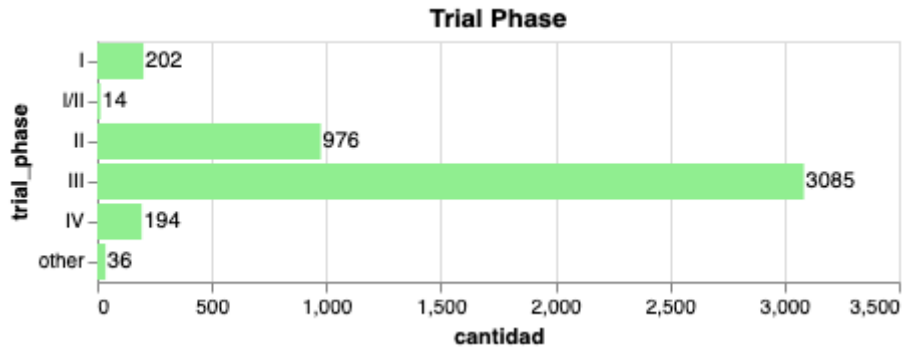


Figura 21. Cantidad de fases de los ensayos clínicos.

Con respecto a la variable "*therapeutic_area*", encargada de establecer y encuadrar a qué área terapéutica correspondía el ensayo, presentaba valores porcentuales de frecuencia de aparición bastante distribuidos entre sí. Se notó, sin embargo, que diversas áreas relacionadas con la oncología eran las que más aparecían entre los primeros puestos. Esto puede observarse en la Figura 22.

P0 Oncology (Solid Tumors, all MA Onc projects)	0.261815
P9 Cardiovascular, Renal and Metabolism (CRM)	0.221211
P7 Immunology	0.175061
P4 Neuroscience	0.097404
PA Oncology (Hematology not for MA projects)	0.081873
P5 Ophthalmology	0.054138
N7 Oncology	0.035057
P8 Respiratory and Allergy	0.014644
N5 DAx	0.014422
P2 SZ Biopharma	0.010206
P6 Global Health	0.009984
N8 Ophthalmology	0.007100

Figura 22. Porcentajes de frecuencia de una determinada cantidad de valores de la variable "*therapeutic_area*".

La Figura 23 muestra gráficamente lo mencionado previamente sobre las áreas terapéuticas. Se pudo observar que en los primeros 7 puestos del ranking es donde se encuentran la mayoría de los registros de la base de datos. Con una fuerte presencia de lo oncológico.

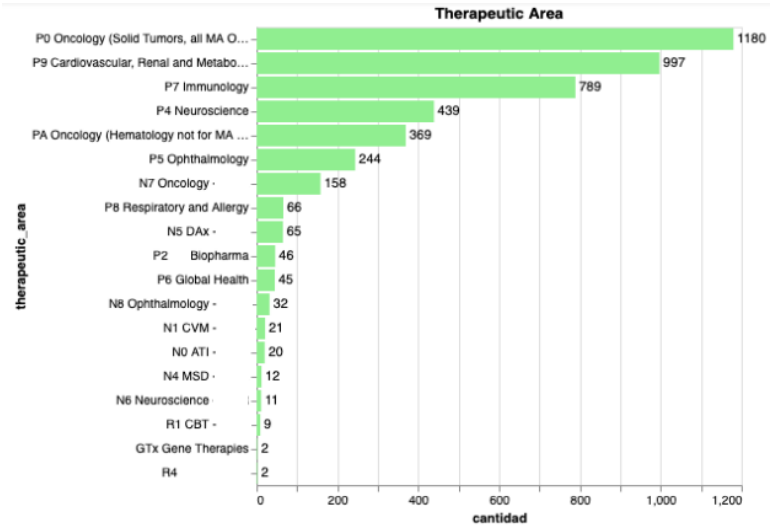


Figura 23. Áreas terapéuticas de los ensayos clínicos.

La variable "*manganing_organization*" mostró una clara presencia de la managing_organization_1, con un 78% de las variables con aquella opción. Esta variable indica a la organización encargada de asistir, ayudar y monitorear los ensayos clínicos junto a Empresa ABC. El resto de las opciones tuvieron porcentajes de presencia no tan dispares entre sí, como puede observarse en la Figura 24.

Managing Organization 1	0.785223
Managing Organization 2	0.072998
Managing Organization 3	0.072110
Managing Organization 4	0.059463
Managing Organization 5	0.010206

Figura 24. Porcentajes de frecuencia de una determinada cantidad de valores de la variable "*managing_organizations*".

La Figura 25 muestra la importancia de la organización Managing Organization 1 por sobre el resto de las organizaciones. Superaba ampliamente al resto de las organizaciones, con un total de 3539 registros sobre los 4507 que presentaba la base de datos.

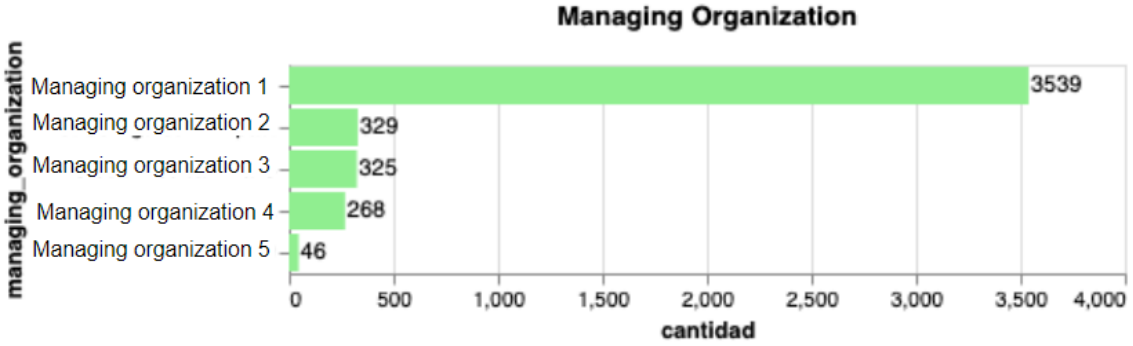


Figura 25. Organizaciones de los ensayos clínicos.

La variable objetivo "*startup_time*", como muestra la Figura 26, presentaba una acumulación de registros alrededor del valor de los 10 meses de duración. Sin embargo, había una considerable cantidad de registros que contaban con duraciones mayor a los 20 meses, tiempos demasiados altos de *start up time*.

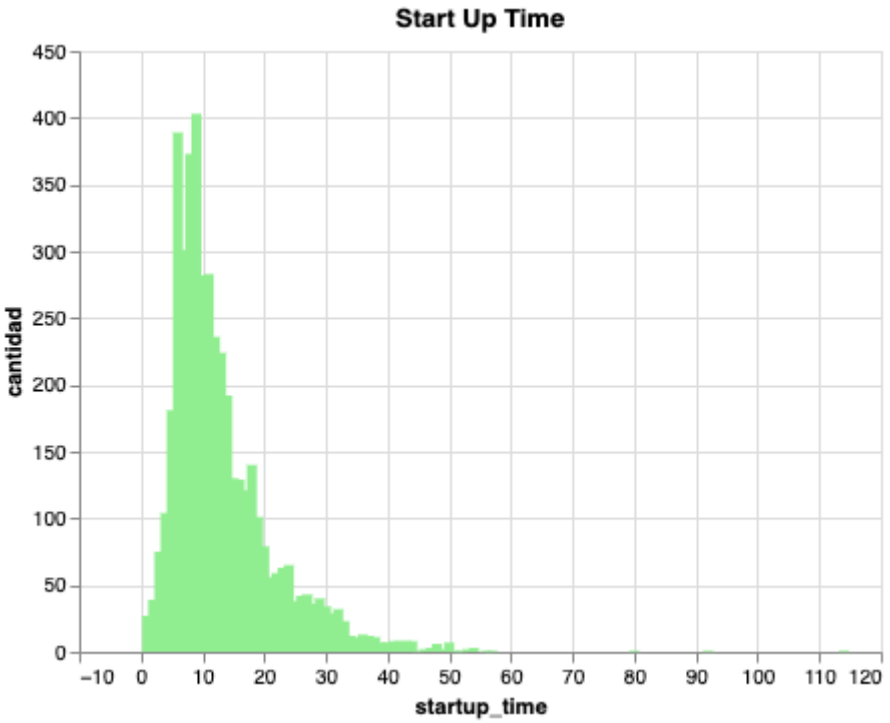


Figura 26. Histograma de la variable "*startup_time"*.

La variable "*startuptime_plan*" tuvo la particularidad de contar con estimaciones de las duraciones con valores negativos. Esto puede haber ocurrido debido a un error en la carga de datos manuales de las fechas utilizadas para las estimaciones o errores de otra índole. Son varias las estimaciones que presentan una duración negativa de 20 meses. Fuera de esto, se observó que en el sector de valores positivos, las estimaciones hechas por la farmacéutica seguían una distribución similar a la del tiempo de *start up* real. Con una clara acumulación alrededor de los 10 meses. Todo esto puede observarse en la Figura 27.

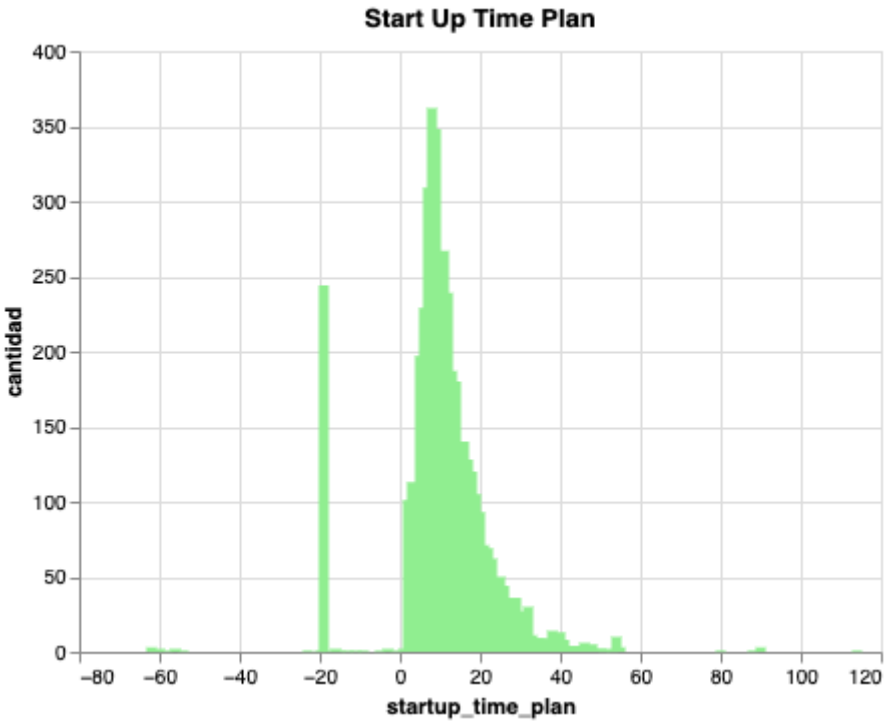


Figura 27. Histograma de la variable "*startup_time_plan"*.

Finalmente, luego de toda la limpieza de variables consideradas no pertinentes, el dataset queda conformado por las variables que pueden verse en la Figura 28 junto a la cantidad de registros del *dataset*.

startup_time	4507
startup_time_plan	4507
overenrolled	4507
therapeutic_area	4507
trial_phase	4507
country	4507
managing_organization	4507
address_province_state_name	4507
indication	4507

Figura 28. Variables del *dataset*.

Análisis Bivariado

A posteriori, se realizó un análisis de las variables con respecto a la variable objetivo. Esto para observar el comportamiento de las mismas y poder atisbar una idea de su relación con el *start up time.*

Como no se contaban con variables numéricas, se realizó un análisis de las variables categóricas con respecto a la variable objetivo. Para esto, se estandarizó

la variable de *start up time* para poder llevar a cabo un análisis equitativo con las distintas variables.

En cuanto a la variable "*managing_organization"* se pudo observar, en la Figura 29, que la managing_organization_1 presentaba mayores outliers en cuanto a la duración del *start up time,* teniendo una mayor amplitud de valores que podía tomar la variable objetivo si esta organización se encargaba de ayudar en la administración de los ensayos clínicos. La managing_organization_3 seguía la misma línea que managing_organization_1, pero poseía un cuerpo central un poco mayor con respecto a esta última. El resto de las organizaciones se veían con cuerpos centrales muy parecidos entre sí, a diferencia de managing_organization_5, cuya amplitud era mucho menor. Es decir, los *start up times* tenían una duración menor.

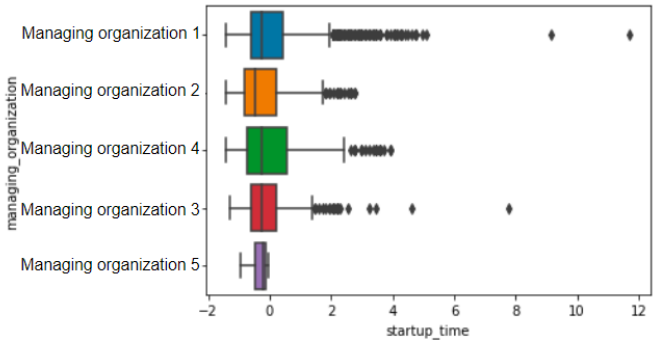


Figura 29. Boxplot de organizaciones.

La Figura 30 muestra el comportamiento de las distintas fases de un ensayo clínico frente al tiempo de *start up*. Se pudo ver que la Fase IV es la que conllevaba, en principio, una mayor duración de arranque, con un cuerpo central mucho más amplio que el resto de las fases. La Fase III es la que poseía un cuerpo más acotado, y sin embargo, es la Fase que contenía la mayor cantidad de registros. Esto sirvió para observar, en principio, que a pesar de tener casi el 70% del total de los registros, su tiempo de *start up* parecía ser el menor comparado al resto de las variables. La Fase I y II eran muy parecidas en estructura, solamente que la Fase I contenía outliers más elevados que el resto de las fases.

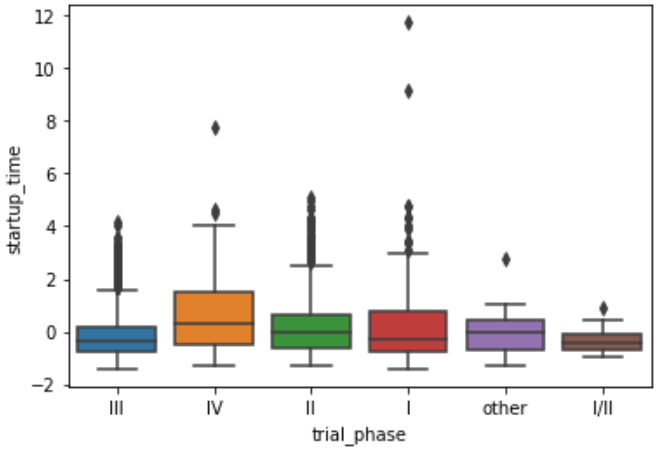


Figura 30. Boxplot de fases de ensayos clínicos.

La variable "*overenrolled",* como muestra la Figura 31, trajo un *insight* interesante. Parecía que el hecho de tener sobre-enrolado un ensayo clínico conllevaba menor tiempo de *start up time.* Esta opción contenía un cuerpo central más pequeño que si no se sobre-enrolaba un ensayo. A la vez que poseía una menor cantidad de outliers.

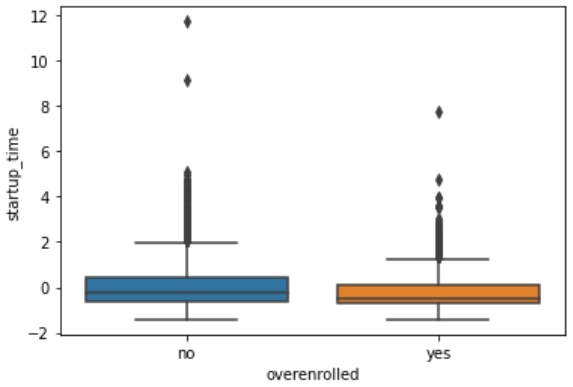


Figura 31. Boxplot de sobre-enrolamiento de un ensayo clínico.

Debido a la gran variedad de opciones que podía tomar la variable "*address_province_state_name*", se decidió agrupar en regiones los distintos estados de EEUU con el fin de contar con datos que permitieran tener un mejor manejo y una visualización de los mismos. Es así como se crea la variable "*region"* con las agrupaciones previamente mencionadas. De todas formas se mantuvo la variable "*address_province_state_name*", nada más que fue renombrada como "*state*". En la Figura 32, se muestran las regiones por las que se agruparon los distintos estados. A pesar de que los cuerpos centrales del boxplot no mostraban diferencias sustanciales entre sí, se pudo observar que la región "*South*" tenía duraciones más cortas del *start up time,* pero también contaba con una gran cantidad de outliers. En antítesis, la región "*North East"* parecía contar con tiempos de *start up* más altos. Aunque tampoco en gran medida.

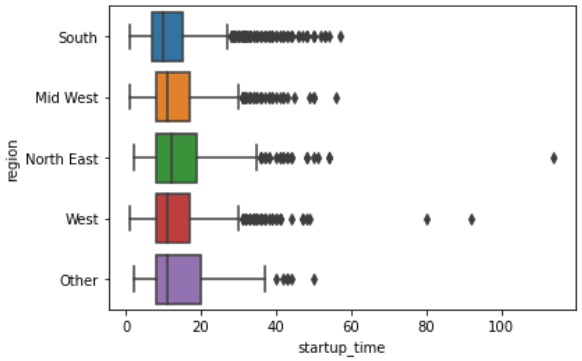


Figura 32. Boxplot de regiones de EEUU.

En la misma senda, se agruparon tanto la variables "*indication*" como "*therapeutic_area*", debido a la cantidad bastante alta de valores que ambas podían tomar. Por lo que se decidió englobar las opciones de cada una en denominaciones generales.

En la Figura 33 se puede observar lo mencionado previamente con la variable "*indication".* Se pudo ver que si un ensayo tiene la indicación de leucemia o esclerosis, la duración del *start up time* suele ser mayor que en el resto de las indicaciones. Esto puede deberse a que tal vez los procesos de arranque con estas indicaciones suelen demorarse más. Por el contrario, cuando se trata de cáncer, la duración daba la impresión de que el tiempo de arranque era menor. Pero contaba con una mayor cantidad de outliers. Lo mismo pudo observarse con la indicación "cardiovascular".

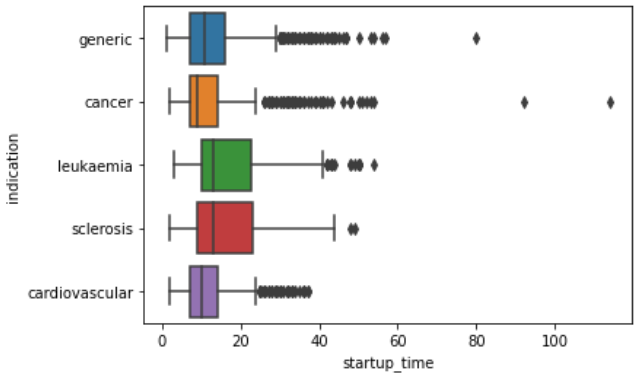


Figura 33. Boxplot de indicaciones de un ensayo clínico.

La Figura 34 muestra los grupos hechos con la variable "*therapeutic_area*" con respecto al *start up time*. El área de oncología se destacó por ser uno de los que menor tiempo de duración de *start up* tenía en relación a la cantidad de registros que este presentaba en el dataset. Sin embargo, esta contaba con muchos outliers. Algo similar se notó con el área de la neurociencia y lo relacionado a lo cardiovascular. Poseían menores tiempos de *start up* pero con una importante cantidad de outliers. El área de oftalmología presenta características similares y sus outliers se reducen. Podría deberse a que organizativamente esta área sea más

expeditiva y/o las habilitaciones necesarias para el arranque de los ensayos clínicos conlleva menos tiempo.

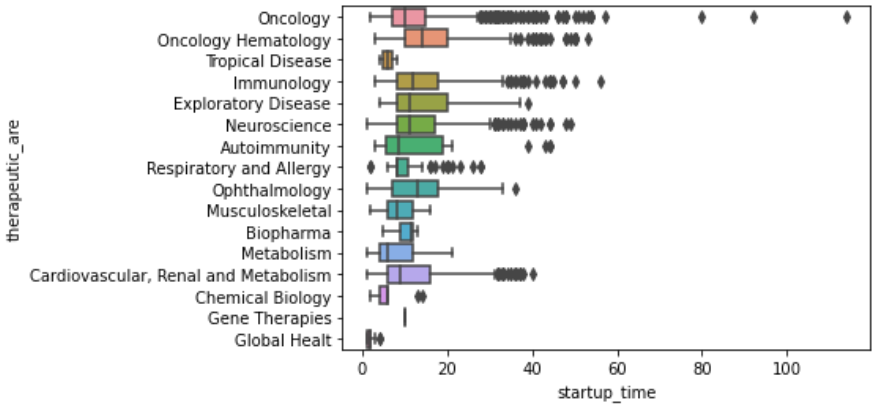


Figura 34. Boxplot de áreas terapéuticas de un ensayo clínico.

Transformaciones

Una vez realizado los análisis univariados y bivariados correspondientes fue hora de preparar las variables y los datos considerados necesarios para la realización de los modelos de *machine learning*.

Como fue comentado en el apartado anterior, a las variables "*indication*", "*therapeutica_area*" y "*address_province_state_name*" se las agruparon en nuevos grupos debido a la alta cardinalidad de valores que presentaba cada una de las variables. En la variable "*indication*" se decidió agrupar a los valores en grupos tales como "*cancer*", "*leukaemia*", "*cardiovascular*", "*sclerosis*" y "*generic*". En la variable "*address_province_state_name"* se encasillaron los estados en las regiones de EEUU correspondiente, "*South*" "*Mid West*", "*North East*", "*West*" y "*Other*". Con esta separación se procedió a crear la variable "*region"* a su vez manteniendo la "*address_province_state_name"* pero renombrando como *"state"*. A la variable "*therapeutic_area"* se la desmembró entre los grupos que pueden verse en la Figura 34 del apartado previo.

Ya realizado lo previo, fue hora de realizarle *encodings* a las variables para estructurarlas y establecerlas como inputs que los modelos de *machine learning* pudieran comprender.

A las variables "*therapeutic_area", "state",* "*trial_phase*" se les decidió realizarles una codificación conocida como *frequency encoding* debido a su alta cardinalidad. Este encoding reemplaza el valor que toma la variables por un número asociado a la frecuencia de aparición de ese valor en el total de los registros. Los nuevos valores asociados pueden verse en las Tablas 2, 3 y 4 del Anexo.

A las variables "*indication_group", "region", "managing_organizations"* se les decidió aplicar una codificación *one hot encoding* debido a la baja cardinalidad que

presentaban estas variables. Esta codificación le asigna el valor ‘1’ a una nueva columna derivada de la variable original, donde el valor del registro corresponde al valor que especifica el nombre de la nueva columna. Sino se le asigna un ‘0’.

Luego de realizar estas codificaciones se contaba con un dataset con las variables mostradas en la Figura 35.

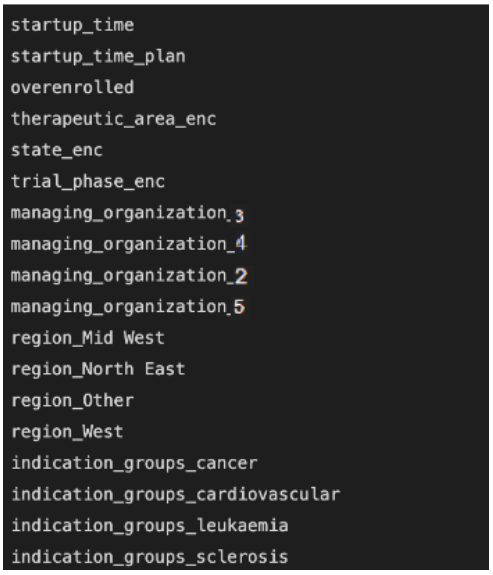


Figura 35. Variables finales del dataset.

Modelos de Machine Learning

Luego de todo el análisis, preparación, limpieza y codificación de las variables del dataset llegó la hora de las pruebas con los diferentes modelos de *machine learning* elegidos. Se optó por utilizar y probar los siguientes modelos orientados a la regresión:

- Árbol de Decisión de Regresión
- Support Vector Regression (SVR)
- Regresión Lineal
- Ridge Regression

Se probaron distintas versiones de los modelos en base a las variaciones de los hiperparámetros que cada modelo tenía para modificar. Se subdividió al dataset en sets de entrenamiento, validación y testeo. Los porcentajes utilizados para esta división fueron 60%, 20% y 20% en concordancia con los sets previamente mencionados. Primeramente se entrenaron los modelos y con el set de validación se decidió que configuración de hiperparametros de cada modelo era la mejor según el criterio de los estadísticos elegidos para la comparación. Y se eligió al mejor modelo entre los 4 tipos de algoritmos. Una vez elegido el modelo “ganador” se lo corrió con el conjunto de testeo para ver su performance final. Los estadísticos

elegidos para observar el comportamiento del modelo y para la comparación con el resto de los modelos fueron el R^2 y el WAPE (Weighted Absolute Percentage Error). El R^2 mide cómo se ajusta un modelo de regresión a los datos reales. El WAPE mide la desviación global de los valores predecidos con respecto a los valores reales. Cuanto más alto el R^2 mejor ajusta el modelo y cuanto más bajo el WAPE mejor se comporta.

Dentro del dataset se contaba con la variable "*startup_time_plan",* que informaba la estimación realizada por la farmaceútica sobre cuánto iba a demorar el tiempo de *start up* para un ensayo clínico. En base a esto es que se calcularon los estadísticos previamente mencionados sobre esta variable. Esto para poder tener valores con los cuales comparar con lo que se obtuvo de los modelos de *machine learning.* Se obtuvo que para la predicción en meses cuentan con un R^2 de -0.7 y un WAPE de 0.22 . Las estadísticas previas se obtuvieron teniendo en cuenta el set de testeo de la variable, esto para poder comparar con los modelos de *machine learning*. En el set de entrenamiento se obtuvo un R^2 de -0,48 y 0.22 . Como puede observarse, el planeamiento de Empresa ABC no ajustaba muy bien con los valores reales de los tiempos de *start up.* Su R^2 es un indicio de lo previamente mencionado.

A continuación, se analizaron los resultados obtenidos a partir de la experimentación para obtener la mejor configuración de hiperparametros para cada tipo de modelo.

En primer lugar, la Ridge Regression quedó parametrizado con un *alpha* de 0, *iteration* en 9000 y una *precision* en 4.1e-05. Su R^2 fue de 0.121 y un WAPE de 0.429 cuando se lo corrió con el set de validación, estos valores se ven reflejados en la Tabla 7 del Anexo. Ambos siendo valores poco representativos a la hora de considerar el modelo como oportuno.

En segundo lugar, la Regresión Lineal quedó determinada que entregaba sus mejores estadísticos a la primera potencia; y estos resultaron en un R^2 de 0.121 y un WAPE de 0.429. Estos valores fueron expuestos en la Tabla 6 del Anexo, obteniendo la misma conclusión que para la Ridge Regression.

El modelo de SVR entregó sus mejores resultados del set de validación cuando su kernel es "*rbf*" y su hiperparámetro "*C*" es 116 . Los estadísticos obtenidos fueron un R^2 0.093 y un WAPE de 0.398, estos valores están presentados en la Tabla 8 del Anexo. Para este modelo se iteró en la experimentación entre los dos kernels, "*rbf*" y "*poly*", y un valor "*C*" de 0 a 200, ya que se vio que el "C" por defecto estaba limitando la performance.

Por último, el Árbol de Decisión entregó sus mejores resultados del set de validación con un R^2 0.181 y un WAPE de 0.391, teniendo los hiperparametros *depth* igual a 10 , *min sample leaf* 16 y *min sample split* 38, estos resultados están

presentes en la Tabla 5 del Anexo. Este modelo es el único que poseía valores cercanos a ambas métricas para evaluación y entrenamiento.

Para la experimentación primero se optó por iterar los parámetro de 1 a 10 para los tres, pero se notó una limitación para algunos de los parámetros y se agrandaron los intervalos hasta 15 para los *depths*, 25 para *min sample leaf* y 55 para *min sample split*.

Luego, debido a lo mencionado y analizado previamente, se decidió elegir al Árbol de Decisión (10-16-38), ya que presentó los mejores estadísticos. Entonces fue este el que se probó con el set de testeo para medir su performance final. Este modelo resultó obtener un R^2 de 0.21 y un WAPE de 0.35 con el conjunto de testeo. De todas formas, todavía tiene un largo camino por recorrer para ser considerado un buen modelo.

Como comentario adicional, se trató de entrenar los modelo utilizando la metodología de *Cross Validation,* pero no se llegaron a resultados diferentes significativamente con respecto a lo mencionado previamente, entonces se decidió no implementarlo.

Feature Importance - Árboles de Decisión

Una vez escogido el Árbol de Decisión como modelo “ganador” se buscó reflejar qué variables tomaban mayor importancia dentro del modelo. Esto para poder saber cuales son las variables a tener en cuenta al momento de decidir la configuración de un ensayo clínico y su repercusión en el *start up time.*

Se puede observar en la Figura 37 a continuación, la importancia de cada variable predictora sobre la variable a predecir, *start up time*. La variable con más impacto es *therapeutic_area_enc* con un 31,65% de importancia, seguida por la *trial_phase_enc,* que describe en qué estado de fase clínica se encuentra el ensayo, con un 15,9%. Este modelo completo del Árbol de Decisión presentaba un R^2 de 0,21 y un WAPE de 0,35. Se decidió eliminar del modelo aquellas variables que presentaban menos del 1% de importancia. Esto para observar si los estadísticos tenían o no una mejoría en la reducción de variables y además para eliminar aquellas variables que no aportan a la predicción.

	features	importance
1	therapeutic_area_enc	0.316542
3	trial_phase_enc	0.159437
15	indication_groups_sclerosis	0.108673
0	overenrolled	0.088365
2	state_enc	0.085969
14	indication_groups_leukaemia	0.063590
12	indication_groups_cancer	0.046771
6	managing_organization_2	0.029114
4	managing_organization_3	0.026012
5	managing_organization_4	0.025067
13	indication_groups_cardiovascular	0.021912
9	region_North East	0.016610
11	region_West	0.011937
7	managing_organization_5	0.000000
8	region_Mid West	0.000000
10	region_Other	0.000000

Figura 37. Importancia de las variables de Árbol completo.

Por lo tanto se decidieron eliminar del modelo las variables *managing_organization_5, region_Mid West, region_Other.* Una vez eliminadas estas variables, se procedió a correr el modelo con el set de testeo y se obtuvieron los mismos valores para las métricas, 0,21 y 0,35 para el R^2 y el WAPE respectivamente. A su vez, se procedió a realizar un *feature importance* de este modelo. Se obtuvo un comportamiento similar a la importancia anterior, la importancia liderada por *therapeutic_area_enc* y seguida por *state_enc*, como se puede observar en la Figura 38.

	features	importance
1	therapeutic_area_enc	0.316542
3	trial_phase_enc	0.159437
12	indication_groups_sclerosis	0.108673
0	overenrolled	0.088365
2	state_enc	0.085969
11	indication_groups_leukaemia	0.063590
9	indication_groups_cancer	0.046771
6	managing_organization_2	0.029114
4	managing_organization_3	0.026012
5	managing_organization_4	0.025067
10	indication_groups_cardiovascular	0.021912
7	region_North East	0.016610
8	region_West	0.011937

Figura 38. Importancia de las variables del Árbol.

Se decidió probar nuevamente reducir las variables del árbol para observar si ocurría una mejoría en los estadísticos. De esta manera, se decidió quitar del modelo aquellas variables que posean una importancia menor al 5%. Las variables *indications_groups_cancer, managing_organization_3, managing_organization_4, indications_groups_cardiovascular, region_North East* y *region_West*. Este nuevo modelo resultó con un R^2 de 0,17 y un WAPE de 0,36. Estadísticos con peores valores que los modelos anteriores. De tal manera que se decidió finalizar aquí el recorte de variables por el hecho de que cada vez permanecen menos variables para eliminar y que los estadísticos resultaron en valores más bajos que en el resto de los modelos.

Modelo AdD	Test R2	Train R2	Test WAPE	Train WAPE
Completo	0.21	0.30	0.35	0.34
Sin 1%	0.21	0.30	0.35	0.34
Sin 5%	0.17	0.28	0.36	0.35

Tabla 1. Estadísticas de los Árboles de Decisión.

Como puede observarse en la Tabla 1, el árbol completo y aquel al que se le fueron sacadas las variables que no cumplían con un 1% de importancia poseen estadísticos muy similares. Se decidió elegir como modelo “ganador” al árbol al cual se le quitaron las variables menores al 1% de importancia. Esto se debe a que en el árbol completo se tienen variables que no son para nada importantes en el modelo pero sin embargo se encuentran dentro del mismo. Esto puede generar ciertas complicaciones a futuro al momento de interpretar los resultados. Por lo que el árbol recortado, al no contar con estas variables, termina siendo la mejor opción.

Comportamiento de Variables

Luego de realizar el *feature importance* del árbol de decisión, se decidió también realizar un análisis de los *shap values* de las variables del modelo. Esto se debe a que estos *values,* que se aplican sobre las variables predictoras, otorgan una interpretación más clara del impacto de las variables sobre la variable objetivo, en este caso, el *start up time.* Estos valores muestran la contribución de cada uno de las variables sobre aquella variable a predecir. Es decir que tanto “peso” tienen a la hora del cálculo del *start up time.* Este análisis fue realizado sobre el árbol considerado “ganador”.

En la Figura 39 se muestran las variables que se mantuvieron luego de hacer la reducción de variables post análisis de *feature importance.* Aquellas variables que finalicen con "_enc" son variables que poseen un encoding no dicotómico. Las que no presentan esa terminación, son dicotómicas. Estos *encodings* pueden observarse en el Anexo en sus tablas correspondientes.

Como puede observarse a mayor valor de *trial_phase_enc* menor es el impacto que se tiene en el tiempo de *start up.* En oposición, cuanto menor es el valor de la variable mayor impacto es en el cálculo de la variable predicha. Esto se condice con un comentario que realizó Empresa ABC, que es que cuando se avanza en las fases en un ensayo clínico, los tiempos de *start up time* se reducen debido a que ciertos procesos se aceleran.

Luego para la variable dicotómica *indication_groups_cancer*, se observa que al ser un ensayo de indicación "*cancer*" disminuye el valor predicho.

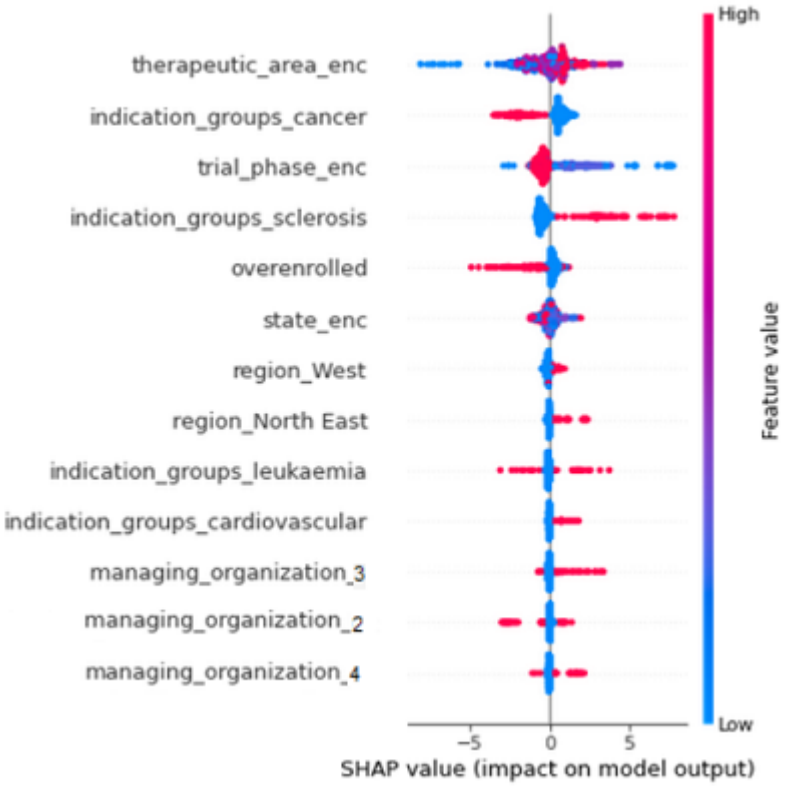


Figura 39. Beeswarm de los *shap values* de las variables.

Luego en la Figura 40, se puede observar el valor de la variable *therapeutic_area* de manera codificada representa Oncology y Cardiovascular que son las áreas más frecuentes , como se puede observar en la Tabla 3 del Anexo. disminuye el valor del *start up time* predicho. Además, para el estado donde se realiza el ensayo, se puede ver que al aumentar el valor *state_enc* el tiempo predicho aumenta, esto hace referencia a los estados más frecuentes sean California, Florida y Texas, obtenido de la Tabla 4 del Anexo . En adición, la managing_organization_3 y la indicación "*leukaemia*" tienen un comportamiento

similar de disminuir el valor predicho, en magnitudes similares. Luego cuando la variable overenrolled toma el valor 1, sea que hay *overenrollement*, aumenta el valor predicho.

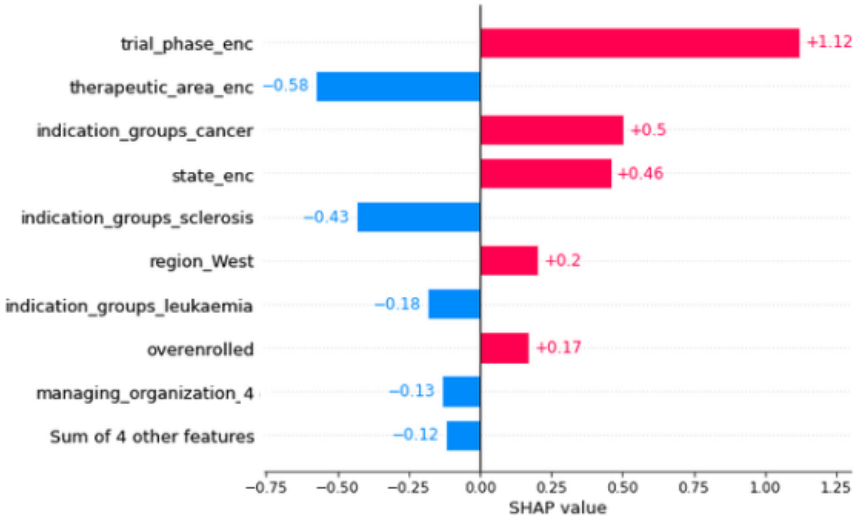


Figura 40. Gráfico de barras de los *shap values* de las variables.

Recomendaciones

A modo de recomendación, para poder contrarrestar un efecto no deseado, viniendo a ser este el aumento el tiempo de *startup* de los diferentes estados más frecuentes, sean California, Florida y Texas, se pueden realizar ensayos de las áreas terapéuticas "*Oncology*" y "*Cardiovascular*", que disminuyen el tiempo de *startup.* Luego para la indicación "*cancer*" se recomienda intentar realizar los ensayos clínicos junto a la managing_organization_3 ya que estas dos tienen efectos opuestos sobre la variable a predecir. Además, para toda la región Oeste (West) buscar realizar ensayos de indicación de "*leukaemia*", ya que este tiene el efecto de disminuir el tiempo de *startup*.

Y por otro lado, se recomienda que se haga especial atención a la variable *overenrolled*, ya que esta aumenta el tiempo de *startup.* A su vez, esta variable está directamente impactada por las decisiones de los encargados del ensayo clínico, ya que se podrían planear y organizar los ensayos correctamente para evitar que se necesite contar con más pacientes que los que fueron tenidos en cuenta en un principio para formar parte del ensayo. Sumado a que sobre enrolear un ensayo clínico conlleva más gastos para la farmacéutica.

A modo de resumen, se detallan las recomendaciones pensadas por el equipo en la siguiente lista:

- Realizar ensayos sobre Oncology y Cardiovascular en California, Florida y Texas.
- Con la managing_organization_3 realizar ensayos de indicación "cancer".

- En los estados de la región Oeste focalizarse en los ensayos de indicación “*leukaemia*”.

Conclusión

En conclusión, el Árbol de Decisión con una configuración de hiperparametros de *depth* igual a 10 , *min sample leaf* 16 y *min sample split* 38 fue el modelo que entregó los mejores resultados en materia de estadísticos. Este fue optimizado y mejorado levemente luego de eliminar ciertas variables que no presentaban tanta relevancia en el modelo.

De todas formas, los estadísticos del modelo no son lo suficientemente fuertes y confiables como para poder determinar que este es un modelo útil a la hora de determinar el *start up time* de un ensayo clínico. Para ser considerado útil se debería contar con un R^2 más alto y un WAPE más bajo. No obstante, se tiene la confianza de que si se contasen con una cantidad de registros mayor y más variables sin valores faltantes, el modelo sí se convertiría en un predictor confiable y robusto del *start up time*.

A pesar de todo esto, se llevó a cabo un análisis tanto descriptivo previo al desarrollo de los modelos de *machine learning* como un análisis post modelos de *machine learning* que permitieron conocer más en profundidad la base de datos y qué variables son las que impactan en mayor y menor medida al tiempo de *start up*.

Las recomendaciones brindadas en la última sección del informe buscan generar una mejoría en lo que respecta a las decisiones que se toman y afectan al tiempo de *start up.* Se espera y fuertemente recomienda que la farmaceútica las tenga en cuenta y pueda aplicarlas dentro de su organización.

En fin, a pesar de que no se logró llegar a un modelo “ideal” para el cálculo del *start up time*, sí se arribó a uno que tuviese mejores estadísticos que la variables *planned* de la farmacéutica. Por lo que, de manera inicial y a forma de prototipo, podría implementarse este modelo junto a otras prácticas, las cuales ya posee la empresa, para el cálculo del *start up time*.

Bibliografía

- *Ensayos Clínicos*. (2018, julio). [Informe]. https://seom.org/seomcms/images/stories/recursos/Ensayos_Clinicos_JUL18.pdf
- *Canva*. (s. f.). Canva. https://www.canva.com/
- School of Software Engineering, Beijing University of Posts and Telecommunication,. (s. f.). *Comparison of Machine Learning Algorithms for Software Project Time Prediction*. https://gvpress.com/journals/IJMUE/vol10_no9/1.pdf
- Python. (s. f.). [Software].
- Excel. (s. f.). [Software].
- Stack Overflow - Where Developers Learn, Share & Build Careers. (s. f.). Stack Overflow, https://stackoverflow.com
- Towards Data Science. (s. f.). Towards Data Science. https://towardsdatascience.com/
- MLflow - A platform for the machine learning lifecycle. (s. f.). MLflow.

Anexo

Cuando llega el protocolo Final, se toma como inicio.	T: Actual Final Protocol	
Cuando se presenta al comité de Ética del Centro de Investigación	S: EC Submission Actual	S: EC Approval Actual

Después de ser aprobado el comite de Etica pasa a HA (ANMAT) – Autoridad Sanitaria	C: HA Submission Actual	C: HA Approval Actual
Firma de Documento para poder iniciar el centro	S: RIS Actual	
Visita de Inicio - el comienzo del Centro para poder empezar a reclutar pacientes	S: SIV Actual	
Primera Paciente con la primera visita al tratamiento	S: FPFV Actual	

Tabla 1. Detalle de Fechas del dataset.

TRIAL PHASE	TRIAL_PHASE_ENC
III	0,68449
II	0,21655
I	0,04481
IV	0,04304
other	0,00798
I/II	0,00310

Tabla 2. Encoding de la variable “Trial Phase”

THERAPEUTIC AREA	THERAPEUTIC_AREA_ENC
Oncology	0,297
Cardiovascular, Renal and Metabolism	0,221
Immunology	0,175
Neuroscience	0,099
Oncology Hematology	0,082
Ophthalmology	0,061
Respiratory and Allergy	0,015
Exploratory Disease	0,014
Biopharma	0,010
Global Health	0,010
Metabolism	0,005

Autoimmunity	0,004
Musculoskeletal	0,003
Chemical Biology	0,002
Gene Therapies	0,000444
Tropical Disease	0,000444

Tabla 3. Encoding de la variable “Therapeutic Area”.

STATE NAME	STATE_ENC	STATE NAME	STATE_ENC
California	0,1187	Connecticut	0,0111
Florida	0,1072	Kansas	0,0107
Texas	0,1054	Nebraska	0,0102
New York	0,0499	Kentucky	0,0095
Ohio	0,0364	Arkansas	0,0091
Massachusetts	0,0337	Oklahoma	0,0075
Illinois	0,0313	Nevada	0,0064
North Carolina	0,0302	Iowa	0,0053
Maryland	0,0282	Mississippi	0,0049
Michigan	0,0273	Puerto Rico	0,0049
Tennessee	0,0271	District of Columbia	0,0042
Georgia	0,0262	Montana	0,0042
Pennsylvania	0,0262	New Mexico	0,0036
Arizona	0,0244	Idaho	0,0033
Missouri	0,0237	Hawaii	0,0031
Washington	0,0200	South Dakota	0,0031
Colorado	0,0195	Rhode Island	0,0022
New Jersey	0,0182	Maine	0,0018
Virginia	0,0175	New Hampshire	0,0018
Louisiana	0,0173	West Virginia	0,0018
South Carolina	0,0160	Delaware	0,0016
Alabama	0,0153	North Dakota	0,0016
Indiana	0,0153	Vermont	0,0011
Oregon	0,0138	Alaska	0,0009

Wisconsin	0,0126	Other	0,0007
Minnesota	0,0122	Wyoming	0,0002
Utah	0,0118		

Tabla 4. Encoding de la variable “State”.

Depth	Min Sample Leaf	Min Sample Split	Validation R2	Train R2	Validation WAPE	Train WAPE
10	16	38	0,181	0,265	0,391	0,387
10	16	36	0,179	0,267	0,392	0,386
10	16	42	0,177	0,261	0,392	0,388
10	16	40	0,177	0,262	0,392	0,388
10	14	40	0,176	0,270	0,392	0,386
10	16	50	0,176	0,259	0,392	0,389
10	16	44	0,176	0,260	0,393	0,389
10	16	46	0,176	0,260	0,393	0,389
10	16	48	0,175	0,259	0,393	0,389
10	18	36	0,175	0,261	0,397	0,391

Tabla 5. Valores del Árbol de Decisión.

Power	Validation R2	Train R2	Validation WAPE	Train WAPE
9	0,111	0,096	0,431	0,454
8	0,111	0,096	0,431	0,454
7	0,111	0,096	0,431	0,454
6	0,111	0,096	0,431	0,454
5	0,111	0,096	0,431	0,454
4	0,111	0,096	0,431	0,454
3	0,112	0,096	0,431	0,454
2	0,115	0,096	0,43	0,454
1	0,121	0,095	0,429	0,455

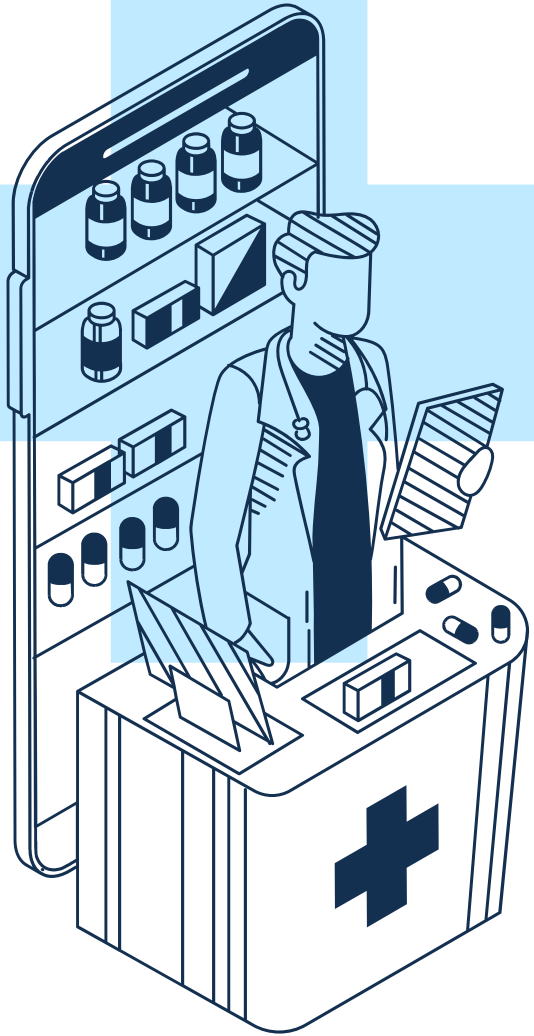
Tabla 6. Valores de la Regresión Lineal.

Iteration	Alpha	Precision	Validation R2	Train R2	Validation WAPE	Train WAPE
90000	0.0	4.1e-05	0,121	0,095	0,429	0,455
90000	0.0	3.1e-05	0,121	0,095	0,429	0,455
90000	0.0	2.1e-05	0,121	0,095	0,429	0,455
90000	0.0	1.1e-05	0,121	0,095	0,429	0,455
90000	0.0	1e-06	0,121	0,095	0,429	0,455
80000	0.0	4.1e-05	0,121	0,095	0,429	0,455
80000	0.0	3.1e-05	0,121	0,095	0,429	0,455
80000	0.0	2.1e-05	0,121	0,095	0,429	0,455
80000	0.0	1.1e-05	0,121	0,095	0,429	0,455
80000	0.0	1e-06	0,121	0,095	0,429	0,455

Tabla 7. Valores de la Ridge Regression.

C Value	Kernel	Validation R2	Train R2	Validation WAPE	Train WAPE
116	rbf	0,093	0,201	0,398	0,363
114	rbf	0,093	0,201	0,398	0,363
112	rbf	0,093	0,201	0,398	0,363
118	rbf	0,093	0,201	0,398	0,363
110	rbf	0,093	0,2	0,398	0,364
120	rbf	0,093	0,202	0,398	0,363
108	rbf	0,093	0,2	0,398	0,364
122	rbf	0,093	0,202	0,398	0,363
106	rbf	0,093	0,2	0,398	0,364
124	rbf	0,093	0,202	0,398	0,363

Tabla 8. Valores del SVR



Predicción del start-up time de un ensayo clínico

Santiago Colantonio - Sebastián Fichendler

Contenido

01

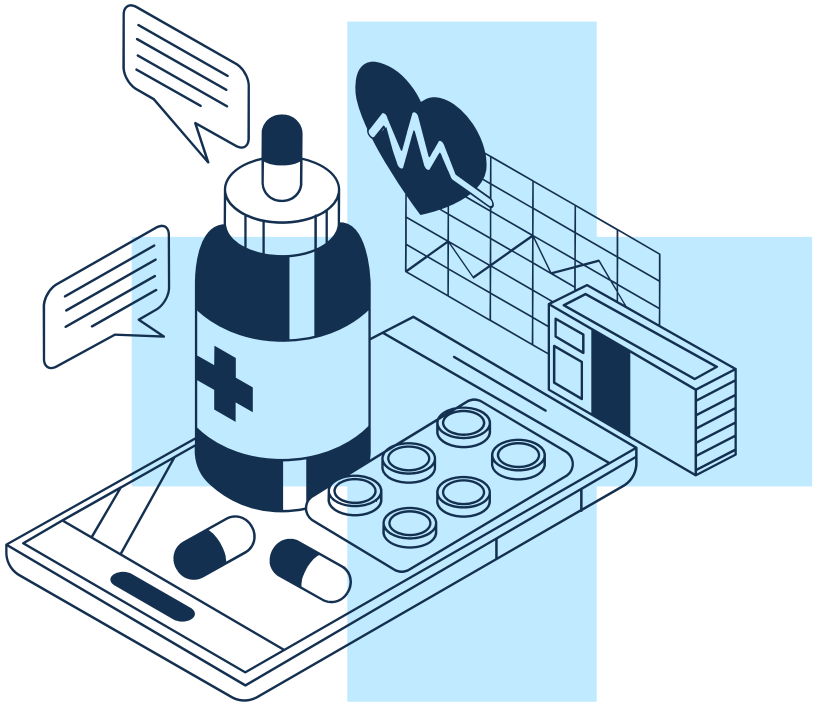
Problemática

El Start Up Time

¿Que es el *start up time*?

Tiempo en que se demora en arrancar una fase de un ensayo clínico

- Cálculo mediante **estimaciones históricas**
- Falta de **entendimiento concreto** de las causas



10 meses

Estándar de la Industria

Estándar de Novartis

8 meses

11,9 meses

Promedio Datos

¿De qué sirve saber de antemano este tiempo?

Mercado

Llegada al mercado antes que el resto de los competidores

Foco

Se puede poner especial atención a aquellas variables que tengan mayor peso

¿Cómo se puede predecir este tiempo?

¡Aprovechando los datos!

02

Solución

Proyecto de *Machine Learning*

¿Con qué datos se contaban?

- Archivo *.xlsx*
- Ensayos clínicos
- Gran cantidad de variables

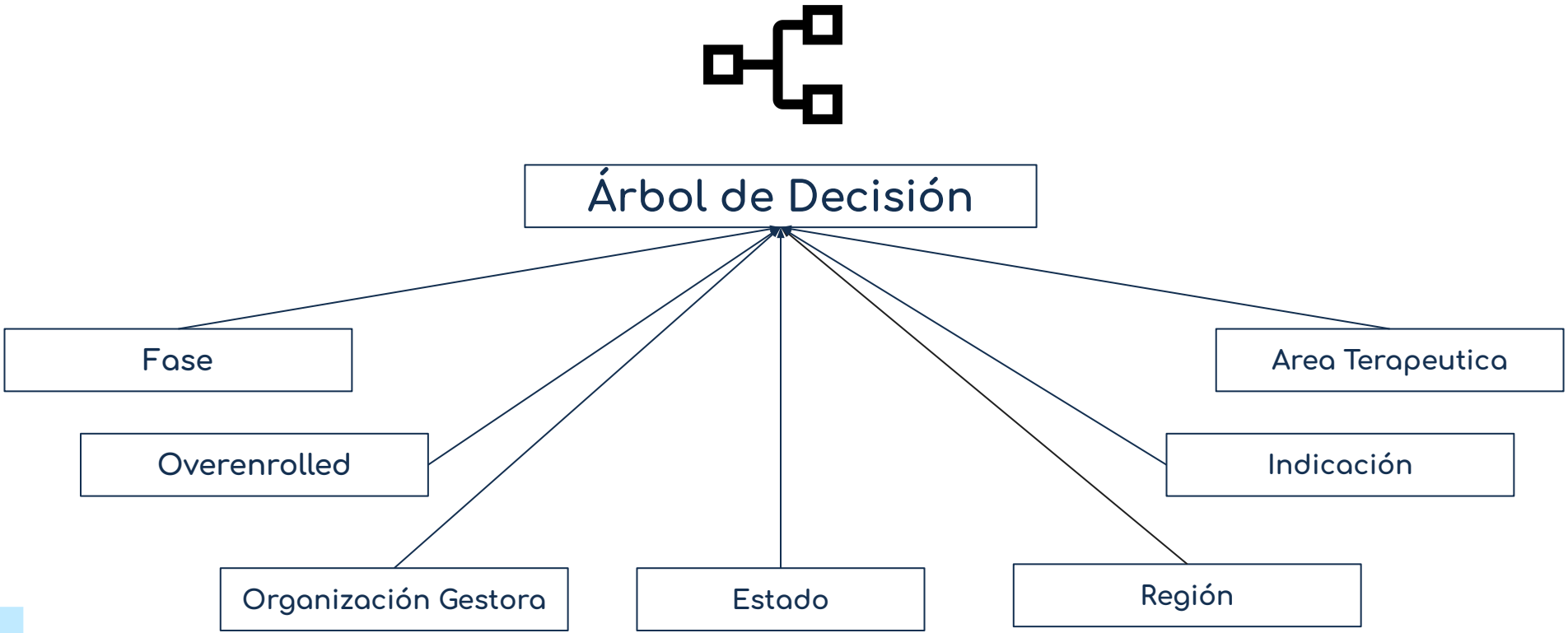
Datos Crudos

Exploración

Transformaciones

¡Listo!

Árbol de Decisión de Regresión



¿Cómo se midió la performance?

R^2

Explicabilidad del modelo

21%

WAPE

Exactitud de la predicción

35%

Impacto

> 18 meses

8 M dólares

En promedio

Descubrimientos y Recomendaciones

Organización 4

Indicación "cáncer"

-10 %

Florida

Área Terapéutica

Neurociencia

o

Cardiovascular

-3 %

03

Pasos a Futuro

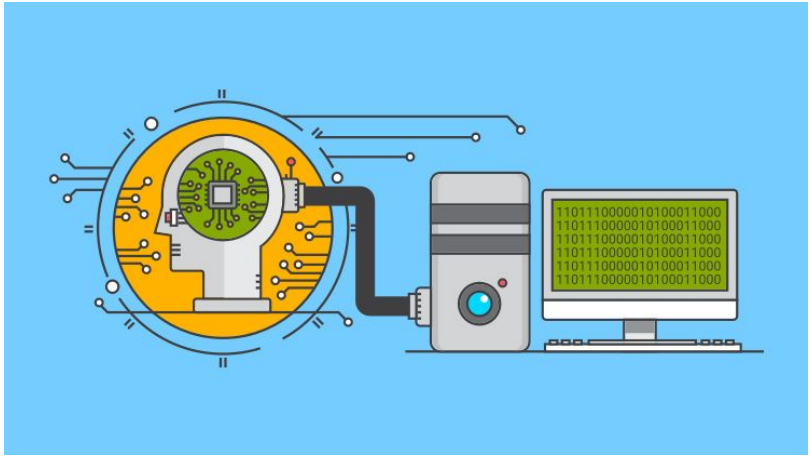
Propuestas a Futuro

Variables
Más información

Cluster
Agrupar por las características de los ensayos

PCA
Reducción de dimensionalidad

Modelo



Profundizar el modelo de *machine learning*

¿Preguntas?

Anexo

Florida
+4%

Organizacion 4
+3%

Cáncer
-14%

Neuroscience
-6%

Cardiovascular
-6%

ANEXO

Autorización para la publicación en el Repositorio Institucional

El/los autor/es firmante/s autoriza/n al Instituto Tecnológico de Buenos Aires (ITBA) a poner a disposición del público la obra detallada en el presente documento, a solo fin de divulgación de la producción científico-académica de la Universidad. El trabajo será de consulta libre y gratuita en el Repositorio Institucional, a través de Internet, y en todos aquellos repositorios digitales en los que participe la Universidad.
Esta autorización representa la cesión al ITBA de forma gratuita y no exclusiva, por el máximo plazo legal y con ámbito universal, de los derechos de reproducción, distribución y comunicación pública por cualquier medio o soporte de la obra.
Asimismo, se autoriza la transformación de la obra, sin producir cambios en el contenido, siempre que sea necesaria para permitir su preservación y uso en formato electrónico, incluyendo la realización de copias digitales y migraciones de formato necesarios para la seguridad, resguardo y preservación a largo plazo de la misma.

1. Datos del/os Autor/es[2]

Apellido y Nombre: Colantonio, Santiago
DNI: 41702626
Legajo: 58724
E-mail: scolantonio@itba.edu.ar

Apellido y Nombre: Fichendler, Sebastián Alejandro
DNI: 42300623
Legajo: 60112
E-mail: sfichendler@itba.edu.ar

2. Datos de la obra

Título completo del trabajo: Predicción del start-up time de un ensayo clínico
Palabras clave: Árbol de decisión, Salud, Recomendación, Tiempo
Carrera: Licenciatura en Analitica Empresarial y Social

3. Tipo de obra:

- Tesis doctorado [ ]
- Tesis maestría [ ]
- Trabajo final de Especialización [ ]
- Proyecto Final de Grado [X]
- Artículo de publicación periódica [ ]
- Libro [ ]
- Parte de libro [ ]

[2] Deberán firmar todos los autores de la obra.

4. Autorizo la publicación del:

Texto completo [X]

A partir de su aprobación/presentación [X]

Dentro de los 6 meses posteriores a su aprobación/presentación [ ]

Otro plazo mayor (detallar/justificar): .....

5. NO autorizo [ ]

Si Ud. Se encuentra comprendido en el caso de que su producción esté protegida por derechos de Propiedad Industrial y/o acuerdos previos con terceros que implique la confidencialidad[3] de los mismos, indique por favor el motivo: .....

.....

El período de confidencialidad o el secreto del trámite finaliza el: .....

El/los autor/es declara/n que la autorización realizada no infringe derechos de terceros y libera/n al ITBA de todo tipo de responsabilidad (sea civil, administrativa o penal) que pudiera surgir frente a cualquier reclamo o demanda referida a la obra por parte de terceros, asumiendo dicha responsabilidad de forma exclusiva.

Acepta/n y toma/n conocimiento de que en caso que la obra sea inédita perderá la condición de tal con su publicación en la web.

Lugar y fecha: .....Buenos Aires, Argentina, 24 de febrero 2023.....

Firma y aclaración del/os autor/es .....

.....

Sebastián Alejandro Fichendler

Santiago Colantonio

A ser completado por el Departamento de Grado /Posgrado/Doctorado: Nro. de Acta ..... Calificación: .....7..... Jurado (Apellido, Nombre): Rodríguez Varela, Juan Pablo Rodríguez González, Rubén ..... Fecha de defensa/aprobación: 07/12/2022	..... Firma y sello (Director de Departamento)

Juan Pablo Rodriguez Varela

[3] NOTA: Se incluyen acuerdos con terceros, políticas institucionales, leyes, reglamentaciones, decisiones unilaterales, etc. En cualquiera de estos casos, se deberá acompañar copia del acuerdo de confidencialidad, del acuerdo que contiene cláusulas de confidencialidad o de la solicitud de derecho de propiedad industrial, o documentación correspondiente.