|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Decorative | | | | |
|  |  |  | |  |
| MODELO DE identificación de licitaciones públicas | | |
| Benjamin kutz | | **Stefani Pereira** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Decorative | | | | |
|  |  |  | |  | |
|  | presentación  El modelo de *machine learning* desarrollado que se detalla a continuación, tuvo como objetivo principal automatizar y facilitar la identificación de licitaciones públicas de Chile, en las cuales, empresas o personas naturales, puedan ofrecer sus productos o servicios. Esto mediante la clasificación de caso entregando una probabilidad donde alguno de sus productos o servicios coincida con la descripción del requerimiento licitado. | |  | |

Contenido

[Beneficiarios del Proyecto 4](#_Toc110080021)

[Caracterización del Objeto de Investigación 4](#_Toc110080022)

[Justificación 5](#_Toc110080023)

[Objetivo General 6](#_Toc110080024)

[Objetivos Específicos 6](#_Toc110080025)

[Ejecución del Proyecto 7](#_Toc110080026)

[1. Entendimiento del Negocio y Acercamiento Analítico 8](#_Toc110080027)

[1.1. Entendimiento del Negocio 8](#_Toc110080028)

[1.2. Acercamiento Analítico 9](#_Toc110080029)

[1.3. Herramientas Utilizadas 9](#_Toc110080030)

[2. Obtención de los Datos 10](#_Toc110080031)

[2.1. Datos Mercado Público 10](#_Toc110080032)

[2.2. Datos ClientMetrica (cliente) 11](#_Toc110080033)

[3. Análisis de los Datos 12](#_Toc110080034)

[4. Preprocesamiento de Datos 13](#_Toc110080035)

[4.1. Limpieza de Datos 13](#_Toc110080036)

[4.2. Creación de Variables 13](#_Toc110080037)

[Impacto y Resultado 15](#_Toc110080038)

[5. Modelamiento 15](#_Toc110080039)

[6. Análisis de Resultados 15](#_Toc110080040)

[Análisis de errores 17](#_Toc110080041)

[Conclusiones 18](#_Toc110080042)

[Líneas de Continuidad 19](#_Toc110080043)

[Anexos 20](#_Toc110080044)

|  |  |  |  |
| --- | --- | --- | --- |
|  | |  | |
| Beneficiarios del ProyectoCaracterización del Objeto de Investigación Las licitaciones públicas son el método utilizado por la administración pública para la compra de productos y obtención de servicios. Es decir, cuando algún órgano del estado tiene una necesidad de comprar o contratar algún tipo de servicio, hace un llamado público en la cual detalla su necesidad entregando características específicas para que proveedores puedan aplicar y ofrecer sus productos y servicios. Así se espera lograr transparentar a la población sobre el gasto público, facilita la conexión entre las compras del estado con los proveedores, y dar un buen uso de recursos del estado a través de la selección del postor que ofrezca el mejor precio de un producto y/o servicio. En el caso particular del Estado de Chile, el Gobierno ha habilitado una plataforma de licitaciones públicas llamada Mercado Público, la cual se puede acceder mediante el navegador escribiendo la dirección [www.mercadopublico.cl](http://www.mercadopublico.cl).  Mercado Público habilita en su plataforma filtros de búsqueda que permita facilitar empresas o personas buscar licitaciones en las cuales pueda ofrecer sus productos o servicios. Tales filtros incluyen el título o ID de la licitación, nombre del organismo público, ubicación geográfica, fecha cierre, y su estado. El resultado de la búsqueda entrega una grilla con campos que detallan los filtros ingresados y adicionalmente un resumen que describe información acerca de la licitación. Siendo este último el parámetro más importante ya que posee en detalle los productos solicitados (o servicios) y sus cantidades respectivas. (Ver figura1)  Lamentablemente tales filtros no son suficientes para acotar la búsqueda de oportunidades, lo que termina dificultando la identificación de productos y resultando en un proceso de búsqueda muy costoso. El buscador de la plataforma termina siendo limitado en su facilidad de búsqueda al no entregar como criterio el producto o servicio que se licita. Lo cual convierte el trabajo de una persona en una actividad titánica al tener que leer diariamente la descripción de licitaciones nuevas tratando de identificar si alguno de sus productos o servicios es requerido por el Estado.  Y si fuera poco, a este problema de búsqueda se debe sumar el alto volumen de licitaciones que se publican diariamente, inconsistencia que existen en la data detallada en la licitación (diferencias entre el título y la descripción), el grado de conocimiento sobre el catálogo de productos y servicios que debe tener la persona, concentración y comprensión lectora del individuo que busca. Justificación Este problema es fácilmente abordable con una máquina/robot, a la cual caracterizaremos nombrándolo: **Juanito** a modo de representar a esa persona que realiza el trabajo de búsqueda de oportunidades en licitaciones públicas. Este robot debe ser capaz de resolver tales dificultades a un **costo** **menor en esfuerzo y tiempo** en comparación al que tenga para un ser humano.  La empresa ClientMetrica es quien nos presenta tal desafío dado que detectó este problema como un patrón que se repetía en la mayoría de los procesos de prospección de sus clientes. De esta forma, el modelo resultante servirá para ofrecer un servicio a las empresas el cual les permita reducir sus tiempos de búsqueda, mejore su tasa de éxito de encontrar oportunidades y no perder oportunidades, así como redirigir esfuerzos del equipo comercial a preparar la oferta para ganarse la licitación.  Por lo tanto, nuestro trabajo se enfocará en construir una herramienta de inteligencia artificial la cual **automatice el proceso de búsqueda** de productos y servicios **clasificando** aquellas **licitaciones** en las cuales una empresa o persona **pueda participar**.  Para las personas naturales o jurídicas que desean participar en las licitaciones públicas, este modelo les permitirá obtener información sobre las licitaciones en las que puedan optar teniendo probabilidades de éxito, lo que ayuda a reducir el trabajo manual que lleva al **desgaste físico** y **emocional** de la persona quien realiza la búsqueda, **disminuir** **el tiempo de exploración y errores** que se puedan cometer en el proceso. Objetivo General Crear un modelo de **machine learning** que identifique, dentro de un listado de licitaciones públicas (de Chile), aquellas donde una empresa o persona pueda participar con su catálogo de productos y/o servicios. Objetivos Específicos  * Identificar los procedimientos actuales de participación en las licitaciones públicas de Chile. * Establecer variables claves que faciliten el proceso de búsqueda de licitaciones. * Construir un algoritmo que permita clasificar si una licitación aplica o no para participar con los productos o servicios del cliente. * Lograr que el modelo tenga un grado de confianza tal que nos permita llevarlo a producción. | | | |
| Lupa con relleno sólido | Ojo con relleno sólido | | Libros con relleno sólido |
| Filtrado  No existe en las opciones de búsqueda el criterio de filtrar por producto para las licitaciones de Mercado Público. | Reconocimiento  Es difícil identificar licitaciones en las que un proveedor pueda postular dentro del resultado de búsqueda. | | Alto volumen  Diariamente se ingresa un alto volumen de licitaciones en la plataforma. |

|  |  |
| --- | --- |
|  |  |

# Ejecución del Proyecto

Para construir a Juanito nos basamos en la metodología **CRISP-DM** para guiar, estructurar, y organizar nuestro trabajo. Bajo este marco es que definimos las distintas acciones que nos permitiera lograr los objetivos planteados. La metodología nos permite **iterar distintas soluciones e ir buscando mejoras** sobre los distintos resultados obtenidos. Así corregir de forma ágil y disminuir en cada iteración errores que se presenten sin tener que llegar al final y tener que comenzar todo de nuevo. Siempre teniendo presente que la solución a la cual se llegue genere resultados verídicos y deseables (respaldado en métricas y grados de confianza aceptables).

Ciclo Metodologíco de Data Science

Diagram

Description automatically generated

## Entendimiento del Negocio y Acercamiento Analítico

### Entendimiento del Negocio

**Mercado Público** al ser una **plataforma transaccional de compras del Estado chileno**, representa una oportunidad de negocios para todos aquellas empresas o personas puedan ofrecer sus productos o servicios. Su objetivo es generar transparencia sobre el proceso de oferta y adjudicación de licitaciones a proveedores privados, además de lograr hacer un buen uso de recursos dado su modelo basado en competencia por precio (asumiendo que se cumplen los requerimientos especificados por el solicitante en la publicación de la licitación).

Todos los años el gasto del estado tiene un crecimiento real lo cual hace aún más atractivo para las empresas poder participar y lograr una participación dentro de las adjudicaciones. El **total gastado durante el 2021 llegó a los 15 millones de USD** lo cual resulta un segmento atractivo para ser abordado por una empresa. En el Gráfico 1 (ver Anexo) se puede apreciar la tendencia en el aumento del total de gasto realizado por el Estado chileno en compras de productos y contratación de servicios.

El total de participantes se separa en **demandantes** de productos y servicios (850 Organismos del estado chileno que representan cerca de 15 mil compradores individuales) y en **oferentes** (91 mil proveedores que pueden ser grandes empresas, MiPymes -micro, pequeñas, o medianas empresas-, o personas naturales). En el gráfico 2 (ver Anexo), se puede apreciar mayor detalle de la distribución de compras (en monto USD de gasto) en los distintos tipos de organismos del estado.

Analizamos los distintos **tipos de mecanismos de compra** que Mercado Público habilita en su portal identificando que las **licitaciones** son el tipo de compra **más utilizado** por los distintos órganos del estado (representa sobre un **60%** **del gasto total** realizado durante el 2021). Para mayor detalle de la distribución del gasto en el resto de los distintos mecanismos referirse al gráfico 3 en los anexos. Es en este tipo de compra donde se centrará la solución para detectar oportunidades.

Para este primer modelo se seleccionó a **uno de los clientes de ClientMetrica** (el cual no se identificará por efectos de acuerdos de confidencialidad) el cual tiene historia participando en Mercado Público realizando ofertas en distintas licitaciones con un mix de productos extenso (cerca de **13 mil SKU’s**) centrado principalmente en **productos y materiales de seguridad**.

### Acercamiento Analítico

El problema que se plantea tiene relación con la dificultad en identificar licitaciones donde se requieran productos o servicios de la empresa en cuestión. Por lo que se definió resolverlo mediante un **modelo supervisado de clasificación**. Tal definición se basa en lo siguiente:

* **Variable Objetivo**: {Si o No} la licitación clasifica para poder ofertar productos de la empresa
* **Historia de participación en Licitacione**s: contamos con datos de licitaciones en los cuales la empresa (uno de los tantos clientes de ClientMetrica) ha participado y en cuáles no. Así poder enseñarle al robot de forma supervisado a discriminar entre distintas licitaciones.

Cabe mencionar que se discutieron distintos enfoques y objetivos posibles de abordar, tales como una vez determinada la licitación en la cual participar sugerir el producto adecuado, o predecir la probabilidad de éxito de ganarse la licitación basándose en resultados anteriores. Para acotar el alcance del problema y simplificar la solución, se decidió que tales funcionalidades serán abordadas como mejoras para nuevas versiones de Juanito a futuro dado que el objetivo es sólo identificar licitaciones en la cuales participar.

### Herramientas Utilizadas

Utilizamos distintas herramientas que nos facilitaran tanto la ideación y organización del proyecto como herramientas técnicas para la construcción del modelo de machine learning. A continuación se lista las soluciones relevantes que se utilizaron durante el proyecto, para mayor detalle referirse al anexo 1 para mayor detalle.

#### Ideación y Organización

* Miro, Mural (mapas mentales, diagramas, diseño de flujos)
* Trello (Canvas: organización y coordinación distribución de tareas, planificación diaria y semanal)

#### Tecnología

* Lenguaje: Python
* IDE: VSCode (programación código, ejecución de archivos .py, .ipynb)
* Administrador de librerías: Anaconda
* BBDD: Azure SQL Database
* Control y versionamiento: GIT (repositorio alojado en Github https://github.com/metalkutz/juanito\_v2.0)
* Librerias
  + pickle
  + NumPy
  + Pandas
  + Matplotlib
  + Seaborn
  + Sklearn (model\_selection, metrics, linear\_model, naive\_bayes, ensemble)
  + XGBoost
  + NLTK

## Obtención de los Datos

Existen 2 fuentes de información relevante para el desarrollo del proyecto:

* Mercado Público: detalle licitaciones y participantes
* ClientMetrica (cliente interno): catálogo de productos, y marcas adicionales sobre licitaciones en las cuales no ha participado históricamente pero que si tiene productos que puede ofrecer y le permitiría participar (así aumentar el número de casos sobre los cuales entrenar).

### Datos Mercado Público

Existen tres formas para acceder a los datos de las licitaciones de Mercado Público:

1. Portal transaccional web [Mercado Público - La plataforma de compras públicas y oportunidades de negocio del Estado de Chile (mercadopublico.cl)](https://www.mercadopublico.cl/Home) (requiere usuario y contraseña)
2. Descarga de archivo planos [Mercado Público - Datos abiertos (chilecompra.cl)](https://datos-abiertos.chilecompra.cl/descargas/ordenes-y-licitaciones) (no requiere usuario ni contraseña)
3. Mediante API [Desarrolladores MercadoPublico.cl](https://desarrolladores.mercadopublico.cl/docs/services/5a9ed02f718ed712f4403e75/operations/5a9ed1fc718ed712f4403e78) (requiere de token de acceso asociado a empresa o persona natural)

Cualquiera de las tres formas de entrega de datos mantiene la misma estructura de datos la cual se compone de licitaciones (Tenders definición según estándares OCDE) con productos (o Items) asociados. Es decir, una licitación puede tener muchos productos. Y cada licitación como resultado tiene 1 ganador (a veces varios o incluso ninguno si nadie oferta o cumple con los requerimientos). Para efectos del entrenamiento la data histórica utilizada se basa solo en licitaciones en estado adjudicadas (es decir existe 1 ganador) y por ende están cerradas. Esto nos permite levantar casuística de participación histórica del cliente en licitaciones anteriores.

Actualmente tales datos se encuentran descargados en una base de datos SQL alojado en Azure (MS) de ClienteMetrica. En el anexo 2 se esboza un diagrama simple de tablas y campos (relación mediante campo ID de la licitación/tender) datos sobre los cuales se trabajó durante el proyecto.

Como el objetivo de Juanito es detectar productos del cliente presentes en licitaciones públicas es que utilizaremos la fuente de Items/Productos para la construcción del dataset. En éste se encuentra el detalle que permitirá identificar el producto, esta tabla se encuentra filtrada (mediante query SQL utilizando el RUT o código fiscal chileno de la empresa) sólo por aquellas licitaciones que aplican al cliente (donde puede ofrecer al menos 1 producto de su catálogo). A continuación se describen los campos seleccionados:

* Item Key: primary key del registro
* Tender ID: foreign\_key en base al primary\_key de la tabla Tender (licitación).
* Nombre producto: nombre categórico del producto (definida por Mercado Público).
* Rubros 1, 2, 3: categorías jerárquicas sobre los rubros en los cuales Mercado Público permite a los oferentes clasificar las licitaciones a ser publicadas.
* Descripción: texto libre el cual detalla el producto o servicio demandado.

### Datos ClientMetrica (cliente)

ClientMetrica almacena la información del cliente en el mismo servidor de base de datos SQL en Azure. En éste se guarda tanto el listado del catálogo de productos como las marcas adicionales sobre licitaciones en las que pudo haber participado. Con esta tabla simple (ID\_Item, label) es que se construye la variable target para el modelo, definiendo 1 como licitación que “si aplica” y 0 como licitación que “no aplica”.

De esta forma se construye el data set el cual es una combinación de tablas [ITEM] filtrado por licitaciones en las cuales participó el cliente y la tabla [FEEDBACK\_MP] con marcas adicionales donde pudo haber participado y marcas sobre licitaciones en las cuales no debe participar.

Es importante destacar que para la construcción de las marcas no se pudo basar en todas las licitaciones donde el cliente no participó, por un lado para acotar el total de registros dismiuyendo el desbalance que se produce en la muestra (el cliente no participa de todas las licitaciones de mercado público). De esta forma, aseguramos que Juanito aprenda sobre casos donde realmente “no aplica” y queden casos marcados de forma incorrecta. Sumado a esto, se hizo un esfuerzo para aumentar el total de registros de “si aplica” dada la baja participación del cliente en licitaciones en comparación al total de registros históricos. Acá se descubre que muchas oportunidades se han perdido producto del ineficiente proceso de búsqueda humana. De esta forma, además de aumentar la cantidad de “si aplica” también agregamos casuística para que Juanito aprenda de casos donde la empresa si pudo participar y no lo ha hecho. Así Juanito aprenda a discriminar de forma correcta las licitaciones donde el cliente pueda o no participar.

El resultado de la query entrega un Datasetcon 54.962 registros sobre los cuales entrenar, datos extraídos a un archivo plano formato csv.

## Análisis de los Datos

Como primer análisis, los datos obtenidos presentan un total de 54.962 filas y 8 columnas, donde se pudo notar la ausencia de valores en registros únicamente de la variable “Descripción”, sumando un total de 113 registros.

Se analizaron las palabras del dataset y se pudo notar que se presentan muchas irregularidades, textos en mayúsculas, minúsculas, caracteres especiales, tildes y palabras vacías

Por otro lado, se pudo observar que la variable del resultado de las licitaciones de la empresa(label), conformada únicamente por unos y ceros, donde los unos (1) se refieren a las licitaciones ganadas y los ceros (0) a las licitaciones donde la empresa no participó o no obtuvo la licitación, se manifiesta un desbalance significativo, presentándose en los registros iguales a “0” un total del 90% de la muestra y en los “1” solo el 10%.

Se comparó la distribución del resultado obtenido de las licitaciones de la empresa seleccionada según Rubro 1, llegando a la conclusión de que el Rubro que más licitaciones publica es el Rubro de “Medicamentos y productos farmacéuticos”, donde la empresa no ha participado ya que no cuenta con los productos solicitados en esta categoría (Ver Gráfico 4).

Para las licitaciones ganadas, se pudo apreciar que el área más favorable para la empresa es la de “Equipos y suministros de defensa, orden público, protección y seguridad”, lo cual es coherente comparándolo con el catálogo de productos de la empresa (Ver Gráfico 5).

Asimismo, se pudo evaluar la frecuencia de las palabras encontradas en las licitaciones ganadas por la empresa, predominando las palabras “acero”, “soldadura”, entre otros. (Ver Figura 3)

## Preprocesamiento de Datos

### Limpieza de Datos

Una vez realizado el análisis de los datos, se procedió a efectuar algunas modificaciones para mejorar su lectura y manejo, como lo fue el cambio de nombre de las columnas por unos de fácil entendimiento para el usuario, el reemplazo del index por el ID de los productos (Item\_Key), eliminación de 113 filas cuyas descripciones eran valores “NaNs” y retiro de aproximadamente 3.000 registros duplicados.

### Creación de Variables

El proceso de creación de variables se abordó de manera distinta según el tipo de variable; categórica y campo descriptivo (Descripción Licitación). A continuación, se detalla el proceso de creación de cada una.

#### Categorías Licitaciones

##### Dummies

De las variables categóricas del dataframe, es decir, los distintos rubros y nombre del producto, aplicamos la función get\_dummies para obtener los indicadores ficticios, generando así un dataframe de variables categóricas de 54.849 x 6365.

##### Catboost para selección de variables

Se deseaba reducir la cantidad de variables categóricas eligiendo solo aquellas que tuviesen alguna importancia, por lo que se seleccionaron todas las variables categóricas convertidas en dummies y se les aplicó el modelo Catboost empleando previamente el método Gridsearch para encontrar los mejores parámetros, obteniendo así, la importancia de cada variable que se puede detallar en el gráfico 6, luego, se seleccionaron aquellas variables cuya importancia fuese mayor o igual a 0.000000000001, eligiendo de 6.365 variables categóricas solo 110.

#### Campo Descriptivo

##### Preprocesamiento de texto

Natural Language Processing (**NLP**) es el área de estudio centrada en cómo los ordenadores entienden el lenguaje humano, lo interpretan y procesan, por este motivo, se hace necesario modificar el texto de los datos para que el ordenador sea capaz de comprenderlo.

En la variable descripción se presentan muchas irregularidades en los textos, por lo que se procedió a convertir todo el texto en minúscula, eliminar tildes, caracteres especiales y palabras vacías.

Se transformaron las palabras limpias de las descripciones a su raíz aplicando la técnica de “Stemming” utilizando SnowballStemmer de NLTK, luego, con el empleo de la matriz TF-IDF pudimos obtener la frecuencia en la que aparece cada palabra en las descripciones y al crearse la matriz, cada columna significa una palabra diferente, obteniendo de esta forma 19.535, columnas.

##### Reducción de Dimensionalidad

Al contar con gran cantidad de palabras consideramos oportuno aplicar una reducción de estas mediante PCA, que, al iterar con todas las variables, se pudo notar que con 10.000 componentes se explicaba cerca del 100% de las variables. (Ver figura 4)

|  |  |
| --- | --- |
|  |  |

# Impacto y Resultado

## Modelamiento

En esta etapa del ciclo dividimos los datos de entrenamiento de los cuales nuestro modelo debe aprender y los datos de prueba para comprobar si el modelo que hemos generado funciona, es decir, si las respuestas predichas por el modelo para un caso nuevo son acertadas o no.

Realizamos varias iteraciones en las que los datos se vieron modificados, pero la variable a predecir siempre fue la misma, el label, que es la variable que contiene el resultado positivo (1) o negativo (0) para cada producto licitado.

Decidimos entrenar distintos modelos, el Naive Bayes, XGBoost y Regresión Logística, utilizando criterios diferentes según el resultante de la búsqueda de los mejores hiperparámetros para cada uno de ellos aplicando el método GridSearch, y para evaluar su desempeño utilizamos las siguientes métricas:

* AUROC
* Matriz de confusión
* Precision y Recall (sensibilidad)
* Estabilidad del modelo (varianza de AUROC según muestra utilizando CV)

## Análisis de Resultados

En una primera iteración, se aplicó un modelo de regresión logística tomando en consideración solo las variables categóricas transformadas en dummies, es decir, los rubos y el nombre del producto, dando un total de 6,4 mil variables con un test size de 0,33, un random state de 42 y un thershold de 0,5 y estos fueron los resultados.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Modelo** | **ROC** | **Accuracy** | **Precision 0** | **Precision 1** | **Recall**  **0** | **Recall**  **1** | **F1-score**  **0** | **F1-score**  **1** |
| Regresión Logística | 0,87 | 0,89 | 0,89 | 0.87 | 1 | 0.10 | 0.94 | 0,17 |

A partir de estos datos, se decidió agregar las variables de las palabras contenidas en las descripciones de las licitaciones donde se aplicaron 3 modelos diferentes, empleando un test size de 0,33, un random state de 42 y un threshold de 0,5 en cada uno de ellos, arrojando los resultados que se muestran a continuación:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Modelo** | **ROC** | **Accuracy** | **Precision 0** | **Precision 1** | **Recall**  **0** | **Recall**  **1** | **F1-score**  **0** | **F1-score**  **1** |
| Regresión Logística | 0,97 | 0,95 | 0,96 | 0.84 | 0.98 | 0.67 | 0.97 | 0,74 |
| XGBoost | 0,96 | 0,95 | 0,95 | 0.87 | 0.99 | 0.57 | 0.97 | 0,69 |
| Naive Bayes | 0,83 | 0,70 | 1 | 0.26 | 0.67 | 0.99 | 0.80 | 0,41 |

Se puede observar que la adición de la lista de palabras fue muy positiva para el modelo, mas, logramos notar que seguían arrojando muchas predicciones erróneas, por lo que decidimos aplicar un balanceo de muestras con la técnica “Smote” sobre el Dataframe procesado con 10.000 componentes, se escogió una estrategia de oversampling de la muestra “1” de un 30% de la muestra mayoritaria (“0”), en cuanto al modelo “Naive Bayes” , decidimos descartarlo ya que los resultados eran menos asertivos, obteniendo lo siguiente:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Modelo** | **ROC** | **Accuracy** | **Precision 0** | **Precision 1** | **Recall**  **0** | **Recall**  **1** | **F1-score**  **0** | **F1-score**  **1** |
| Regresión Logística | 0,988 | 0,95 | 0,95 | 0.93 | 0.98 | 0.83 | 0.97 | 0,88 |
| XGBoost | 0,96 | 0,90 | 0,90 | 0.92 | 0.98 | 0.62 | 0.94 | 0,74 |

Los resultados para los parámetros utilizados arrojaron buenas predicciones, el balance de muestra mejora significativamente en la sensibilidad del modelo, lamentablemente, se cometió un error al aplicar la técnica de “Smote” en el entrenamiento, por lo que se realizó la corrección eliminando el balanceo de muestras del test. Por otro lado, a las variables categóricas se les aplicó la función catboost donde se seleccionaron solo las variables categóricas que tuvieran importancia, aplicamos “GridSearch” para encontrar los mejores parámetros y un Stratified Kfolds de 3 para estratificar cada una de las divisiones, obteniendo así estos datos:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Modelo** | **ROC** | **Accuracy** | **Precision 0** | **Precision 1** | **Recall**  **0** | **Recall**  **1** | **F1-score**  **0** | **F1-score**  **1** |
| Regresión Logística | 0,98 | 0,96 | 0.98 | 0.82 | 0.98 | 0.86 | 0.98 | 0.84 |
| XGBoost | 0.98 | 0.97 | 0.98 | 0.86 | 0.98 | 0.82 | 0.98 | 0.84 |

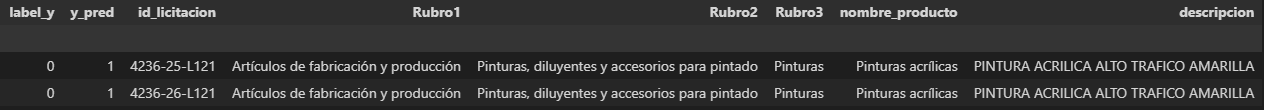
Ambos modelos arrojaron resultados favorables, por lo que decidimos aplicar la técnica de ensamble “Stacking” y aunque el AUC y el f1- score no variaron mucho, generó mayor porcentaje de falsos positivos y negativos, por lo que decidimos no considerarlo y decretamos que el mejor modelo es el de Regresión logística, utilizando como hiperparámetros los siguientes:

* solver=’lbfgs’
* c=1.00
* max\_iter=500
* multi\_class= ‘ovr’
* Penalty=’12’

Los fundamentos de la selección del modelo se deben a la estabilidad con respecto al test estandarizado que arrojó una varianza de solo 0.002062, por otro lado, es más sensible para predecir el resultado de interés (label = 1).

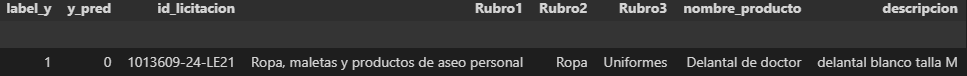
### Análisis de errores

Se analizaron los errores arrojados por el modelo final mediante una matriz de confusión, (Ver figura 4) donde pudimos observar que para los casos en los que el modelo realiza una predicción positiva cuando debería ser negativa, es decir, para los “falsos positivos”, en algunos casos el modelo arroja positivos a todos productos de una categoría específica, aun cuando la empresa no ha ganado licitaciones con tales productos o no cuenta con ellos en su catálogo.



En la imagen anterior, se puede notar que la predicción(y\_pred) para todos los casos de pinturas acrílicas genera resultados positivos, aunque la empresa no cuenta con estos productos en su catálogo.

Por otro lado, con relación a los falsos negativos, sabiendo que el sector sanitario es el sector que presenta más licitaciones, hay algunos productos con los que la empresa podría participar y se ven situaciones en las que las predicciones para estos casos son negativas, por lo que se estarían perdiendo excelentes oportunidades.



|  |  |
| --- | --- |
|  |  |

# Conclusiones

Luego de todo el proceso de trabajo realizado buscando una solución al problema de búsqueda de oportunidades de negocio que tienen las empresas, es que podemos concluir que sí es posible automatizar el proceso de búsqueda de oportunidades de negocio en licitaciones públicas mediante el uso de técnicas de machine learning.

Es posible entrenar de forma supervisada un modelo de regresión logística el cual distingue con una precisión de un 82% y sensibilidad de un 86% casos que son oportunidad de negocios para una empresa en particular. El modelo aprendió de forma particular para el cliente en base a la historia particular de licitaciones donde éste ha participado o potencialmente puede ofrecer su catálogo de productos. Además, se puede aseverar que el modelo es bastante estable para ser puesto en producción dada la baja desviación en su métrica de AUROC frente a distintas muestras aleatorias de testeo aplicadas.

Por otro lado, de los distintos mecanismos que tiene el Estado de Chile para adquirir productos y servicios, son las licitaciones públicas las más utilizadas y es factible acceder a sus datos mediante distintos mecanismos para su obtención y procesamiento. Lo que permite ampliar su alcance y aplicación a distintas empresas de distintos rubros y sectores económicos.

Pudimos comprobar que una de las variables más importantes para el modelo es el campo de descripción de la licitación (texto libre), dado que es sobre éste que es posible identificar al producto que se está licitando y el modelo clasifica de forma correcta como oportunidad de negocio. Los campos categóricos propios de la plataforma de Mercado Público (Rubros) mantienen cierta relevancia para la clasificación del modelo, pero nuestra hipótesis que la categoría *nom\_producto* pierde peso dado que no siempre guarda relación con el producto licitado (ya sea porque el producto que se licita no está disponible dentro del listado o simplemente por error humano).

Es así que podemos concluir que Juanito resuelve la problemática de la búsqueda manual de un ser humano al reducir los tiempos y costos de búsqueda, así como aumentar las oportunidades de negocios. Juanito es capaz de detectar oportunidades donde el ser humano no es capaz de encontrar (debido a restricciones propias del ser humano).

|  |  |
| --- | --- |
|  |  |

# Líneas de Continuidad

Durante el proyecto logramos poner en la práctica todo lo aprendido durante el bootcamp, el cómo resolver un desafío real con la metodología y herramientas de data science. Aplicar el pensamiento analítico y método científico para abordar el problema definiendo el enfoque y herramientas que se utilizaran para diseñar una solución final. Y no sólo aprendimos en el ámbito técnico, sino que además a como trabajar en equipo y organizar el trabajo conjunto.

De todas formas, aún cuando contamos con una solución práctica y confiable para ser puesta en producción, sabemos que aún existe espacio de mejora para Juanito a futuro. Algunos puntos identificados se detallan a continuación:

* Buscar nuevas variables que permitan que Juanito sea más genérico y permita abordar a empresas de distintos sectores.
* Se desea crear un diccionario para “normalizar” las palabras, que no es más que un traductor de sinónimos, ya que una palabra tiene varias acepciones y se presentan muchas de este tipo en los productos solicitados en las licitaciones y en las descripciones de las mismas.
* Otra propuesta sería incluir el catálogo de productos de la empresa al modelo, añadiendo un puntaje en base a palabras encontradas en el campo descripción que sean parte de los productos propios de la empresa.
* A partir del modelo creado, una opción es originar un nuevo modelo de recomendación que arroje como resultado la sugerencia de los productos con los que un proveedor pueda ofertar en las licitaciones obtenidas del modelo anterior, tomando en consideración su catálogo de productos.

# Anexos

**Figura 1:** Búsqueda de Licitaciones en el Mercado Público de Chile.

Interfaz de usuario gráfica

Descripción generada automáticamente

**Figura 3:** Análisis de palabras frecuentes en licitaciones ganadas.

**Texto

Descripción generada automáticamente**

**Figura 4:** Matriz de confusión modelo final.

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

**Gráfico 1**: Evolución del Gasto Público del Estado de Chile, por año tamaño de empresa (USD M)

Chart, bar chart

Description automatically generated

En el gráfico se puede apreciar un crecimiento constante en el gasto del Estado de Chile a través de los años. Salvo algunos años donde el gasto disminuyó, la tendencia general es al alza. Cabe mencionar que el alto volumen de compras a nivel nacional supondría una compensación en ganancias frente a una posible baja en precio (producto de la competitividad misma de la licitación).

**Gráfico 2**: Distribución de Gasto según tipo de Organismo del Estado durante el 2021

Chart, bar chart

Description automatically generated

En el gráfico se puede apreciar que gran parte del gasto fiscal se realiza en el sector de la Salud, donde es esperable dado el modelo de protección social chileno. Luego le siguen el Gobierno Central y Universidades como tipo de Organismo del Estado que más gasto genera en el año.

**Gráfico 3**: Distribución de Gasto según tipo de Mecanismo de Compra en Mercado Público durante el 2021.

Chart

Description automatically generated

**Gráfico 4**: Cantidad de licitaciones por Rubro.

Texto

Descripción generada automáticamente

**Gráfico 5**: Cantidad de licitaciones ganadas por Rubro

Gráfico

Descripción generada automáticamente

**Gráfico 6**: Importancia de variables categóricas

Gráfico, Gráfico de barras

Descripción generada automáticamente

**Gráfico 7**: Varianza total explicada por cada componente

Gráfico

Descripción generada automáticamente con confianza media

**Anexo 1**

**Ideación y Organización**

* Miro, Mural (mapas mentales, diagramas, diseño de flujos)
* Trello (Canvas: organización y coordinación distribución de tareas, planificación diaria y semanal)

**Tecnología**

* Lenguaje: Python
* IDE: VSCode (programación código, ejecución de archivos .py, .ipynb)
* Administrador de librerías: Anaconda
* BBDD: MS SQL (Management Studio)
* Versionamiento: GIT (repositorio alojado en Github)
* Librerias
  + Os
  + Pickle
  + NumPy
  + Pandas
  + Matplotlib
  + Seaborn
  + Sklearn (model\_selection, metrics, linear\_model, naive\_bayes, ensemble, feature\_extraction, decomposition)
  + XGBoost
  + Catboost
  + Collections(Counter)
  + NLTK (corpus, stem.snowball)
  + Wordcloud
  + Regex
  + Imblearn (SMOTE)

Anexo 2:

Table

Description automatically generated