|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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](#_Toc109752344)

[Caracterización del Objeto de Investigación 4](#_Toc109752345)

[Justificación 5](#_Toc109752346)

[Objetivo General 6](#_Toc109752347)

[Objetivos Específicos 6](#_Toc109752348)

[Ejecución del Proyecto 7](#_Toc109752349)

[1. Entendimiento del Negocio y Acercamiento Analítico 8](#_Toc109752350)

[1.1. Entendimiento del Negocio 8](#_Toc109752351)

[1.2. Acercamiento Analítico 9](#_Toc109752352)

[1.3. Herramientas Utilizadas 9](#_Toc109752353)

[2. Obtención de los Datos 9](#_Toc109752354)

[3. Análisis de los Datos 9](#_Toc109752355)

[4. Preprocesamiento de Datos 10](#_Toc109752356)

[4.1. Limpieza de Datos 10](#_Toc109752357)

[4.2. Creación de Variables 10](#_Toc109752358)

[Impacto y Resultado 12](#_Toc109752359)

[5. Modelamiento 12](#_Toc109752360)

[6. Análisis de Resultados 12](#_Toc109752361)

[Análisis de errores 13](#_Toc109752362)

[Líneas de Continuidad 16](#_Toc109752363)

[Conclusiones 17](#_Toc109752364)

[Anexos 18](#_Toc109752365)

|  |  |  |  |
| --- | --- | --- | --- |
|  | |  | |
| 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.  Es la empresa ClientMetrica 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 **Metodológía del Ciclo de Data Science** 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 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 Licitaciones**: 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.

### Herramientas Utilizadas

## Obtención de los Datos

Mediante la Plataforma del Mercado Público, se extrajeron los detalles de las licitaciones públicas de Chile y los productos contenidos en cada licitación, por otro lado, de obtuvieron los datos de un proveedor, su catálogo de productos y sus participaciones en las licitaciones. **COMPLETAR,** se aplicaron métodos de SQL para la fusión de los datoscreando así, un Datasetcon 51.000 registros para entrenar con los campos que se detallan a continuación. **COMPLETAR** (Ver figura 2)

* Tender\_id
* Item\_Key
* Rubros
* Nombre linea adquisición
* Descripcion línea adq.
* Label o marca que identifica si la licitación es o no negocio

## Análisis de los Datos

En la primera iteración se realizó un análisis superficial de los datos donde se comparó la distribución de labels de la empresa seleccionada según Rubro 1, obteniendo como resultado que en el Rubro que más licitaciones publica es el Rubro donde la empresa objeto de estudio no ha participado o no ha obtenido licitaciones. (Ver figura 3)

Asimismo, las muestras contienen un número importante de casos en los cuales no participa el cliente, representados con el valor “0”.

Por otro lado, se pudo observar que se manifiesta un desbalance significativo en la variable “label”, presentándose en la categoría “0” un total del 90% de la muestra y en la “1” solo el 10%.

## 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 columnas 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 detalle el proceso de creación de cada una.

#### Categorías Licitaciones

##### Dummies

##### Catboost para selección de dummies

#### 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.

Se analizaron las descripciones del Dataset y se pudo notar que se presentan muchas irregularidades, textos en mayúsculas, minúsculas, muchos símbolos y caracteres especiales, tildes y palabras vacías, por lo que se procedió a eliminar estas últimas y convertir todo el texto en minúscula.

En el primer ciclo transformamos las palabras de las descripciones a su raíz aplicando la técnica de “Stemming” utilizando SnowballStemmer de NLTK, luego, empleando 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

Modelos seleccionados

Estrategia de entrenamiento

Métricas seleccionadas para su evaluación

* AUROC
* Matriz de confusión
* Precisión 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, 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 a las variables las palabras contenidas en las descripciones de las licitaciones, 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 |

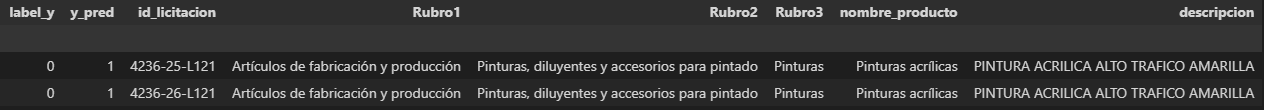
Los resultados para los parámetros utilizados arrojaron buenas predicciones, el balance de muestra mejora significativamente en la sensibilidad del modelo, lamentablemente, se cometió un error al aplicar la técnica de “Smote” en el entrenamiento, por lo que se realizó la corrección eliminando el balanceo de muestras del test, por otro lado, aplicamos “GridSearch” para encontrar los mejores parámetros y un Stratified Kfolds de 2 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 |  |  |  |  |  |  |  |  |
| XGBoost |  |  |  |  |  |  |  |  |

### Análisis de errores

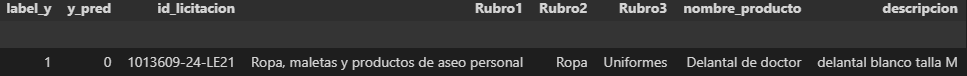
Se analizaron los errores arrojados por el modelo mediante una matriz de confusión, donde se determinó que tan solo el x% de los resultados son falsos positivos y el y% falsos negativos concluyéndose lo siguiente:

Para los casos en los que el modelo realiza una predicción positiva cuando debería ser negativa, es decir, para los “falsos positivos”, se pudo observar que, 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.



Propuesta soluciones

Diccionario para “normalizar” las palabras que no es más que un traductor de sinónimos, una palabra tiene varias acepciones

Puntaje en base a palabras encontradas en el campo descripción que sean parte del catálogo de productos de la empresa.

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

# Líneas de Continuidad

Sería ideal, a partir del modelo creado, 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.

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

# Conclusiones

En el procesamiento de texto, la palabra “NO” tiene gran valor ya que podría cambiar por completo un producto, además, existen muchos términos que son siglas, las cuales pueden representar la marca o el modelo de un producto, por ende, son palabras relevantes que se consideran como variables.

Con respecto a la técnica empleada de reducción de dimensionalidad, el PCA, reduce de forma considerable la cantidad de variables y no afecta tanto en las métricas de evaluación.

Se hace de suma importancia aplicar el modelo con distintos parámetros, esto con la finalidad de poder corroborar que el modelo funciona bien bajo cualquier circunstancia.

# Anexos

**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.

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

Chart

Description automatically generated

**Figura 1**

Interfaz de usuario gráfica

Descripción generada automáticamente

Figura 2

Texto

Descripción generada automáticamente

Figura 3

Texto

Descripción generada automáticamente

Figura 4

Gráfico

Descripción generada automáticamente con confianza media