|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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](#_Toc109634857)

[Caracterización del Objeto de Investigación 4](#_Toc109634858)

[Justificación 5](#_Toc109634859)

[Objetivo General 5](#_Toc109634860)

[Objetivos Específicos 5](#_Toc109634861)

[Ejecución del Proyecto 7](#_Toc109634862)

[Ciclo Metodologíco de Data Science 7](#_Toc109634863)

[Obtención de los Datos 8](#_Toc109634864)

[Análisis de los Datos 8](#_Toc109634865)

[Preprocesamiento de Datos 8](#_Toc109634866)

[Limpieza de Datos 8](#_Toc109634867)

[Procesamiento de texto 9](#_Toc109634868)

[Reducción de Dimensionalidad 9](#_Toc109634869)

[Impacto y Resultado 10](#_Toc109634870)

[Modelamiento y Análisis de Resultados 10](#_Toc109634871)

[Líneas de Continuidad 12](#_Toc109634872)

[Conclusiones 13](#_Toc109634873)

[Anexos 14](#_Toc109634874)

|  |  |  |  |
| --- | --- | --- | --- |
|  | |  | |
| 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.  En resumen, este problema es fácilmente abordable con una máquina la cual pueda resolver estas dificultados sin problema. Es así como nuestro trabajo se enfoca 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. Así les permite enfocar esfuerzos y recursos en ganar las licitaciones identificadas.  CLIENTMETRICA (CLIENTE) Justificación Participar en licitaciones públicas suele generar muchas dudas e incertidumbres sobre los requisitos que se deben cumplir para ganarlas, encontrar la licitación que se adapte al proveedor según producto y cantidades, estar en constante proceso de búsqueda de nuevas licitaciones y una de las más importantes, no perder oportunidades para licitar.  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 que identifique las licitaciones públicas en Chile en las que el proveedor interesado pueda participar según 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. * Automatizar la búsqueda de productos o servicios en licitaciones públicas. | | | |
| Lupa con relleno sólido | Ojo con relleno sólido | | Libros con relleno sólido |
| Filtrado  No se puede filtrar por descripción de la licitación. | Reconocimiento  Es difícil identificar a simple vista las licitaciones en las que un proveedor pueda optar. | | Alto volumen  Diariamente se ingresa un alto volumen de licitaciones en la plataforma. |

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

# Ejecución del Proyecto

Esta etapa tuvo como finalidad describir y definir el modelo siguiendo el proceso completo del ciclo de vida del Machine Learning comenzando por la obtención de los datos, seguido por su análisis, preparación, modelamiento, evaluación, producción y finalizando con un feedback, a partir de este último, según las conclusiones, se corrigen los errores y se repite el ciclo hasta tener un modelo que pueda generar resultados verídicos y deseables.

## Ciclo Metodologíco de Data Science

Diagram

Description automatically generated

## Entendimiento del Negocio y Acercamiento Analítico

El problema es de clasificación

Para resolver el problema nos enfocamos en

* Buscar datos

En este caso, se iteró el ciclo **n** veces, realizando ajustes, modificando variables, experimentando con distintos parámetros, entre otros, donde a continuación se pueden observar las técnicas y procedimientos implementados:

## 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.

### Procesamiento 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 |

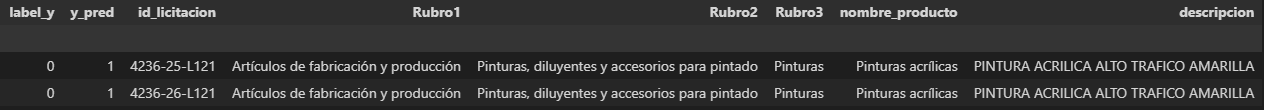
Los resultados para los parámetros utilizados arrojaron buenas predicciones, el balance de muestra mejora significativamente en la sensibilidad del modelo, lamentablemente, se cometió un error al aplicar la técnica de “Smote” en el entrenamiento, por lo que se realizó la corrección eliminando el balanceo 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

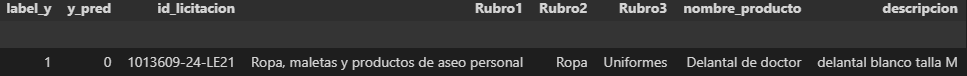
Se analizaron los errores arrojados por el modelo mediante una matriz de confusión, donde se determinó que tan solo el x% de los resultados son falsos positivos y el y% falsos negativos concluyéndose lo siguiente:

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

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