Proyecto: Clasificación de tipo de fallas de productos

Equipo N° 3.

Integrantes: Stefano Canossini, Marcela Distefano, Hernán Ifrán, Damián Joglar, José Valdés.

## Alcance del proyecto

Desarrollar una solución que permita automatizar la clasificación de reclamos a partir de la información obtenida en formularios web.

## Descripción del problema

La empresa Escorial se dedica principalmente a la producción en serie de termotanques, cocinas y calefones en menor escala, siendo el foco en ofrecer el mejor precio del mercado.

Al ser tan elevada la cantidad de productos producidos, requiere de un área aparte de postventa y se le brinda la posibilidad al cliente poder mandar un formulario de contacto desde la página web https://escorial.com.ar/postventa en caso de que exista algún problema con el producto. El formulario exige datos de contacto, del producto con inconvenientes y se debe describir el problema que se tuvo para que el área pueda dar contacto y resolver. A continuación, se presenta pantallazo del formulario que se observa en el sitio web:

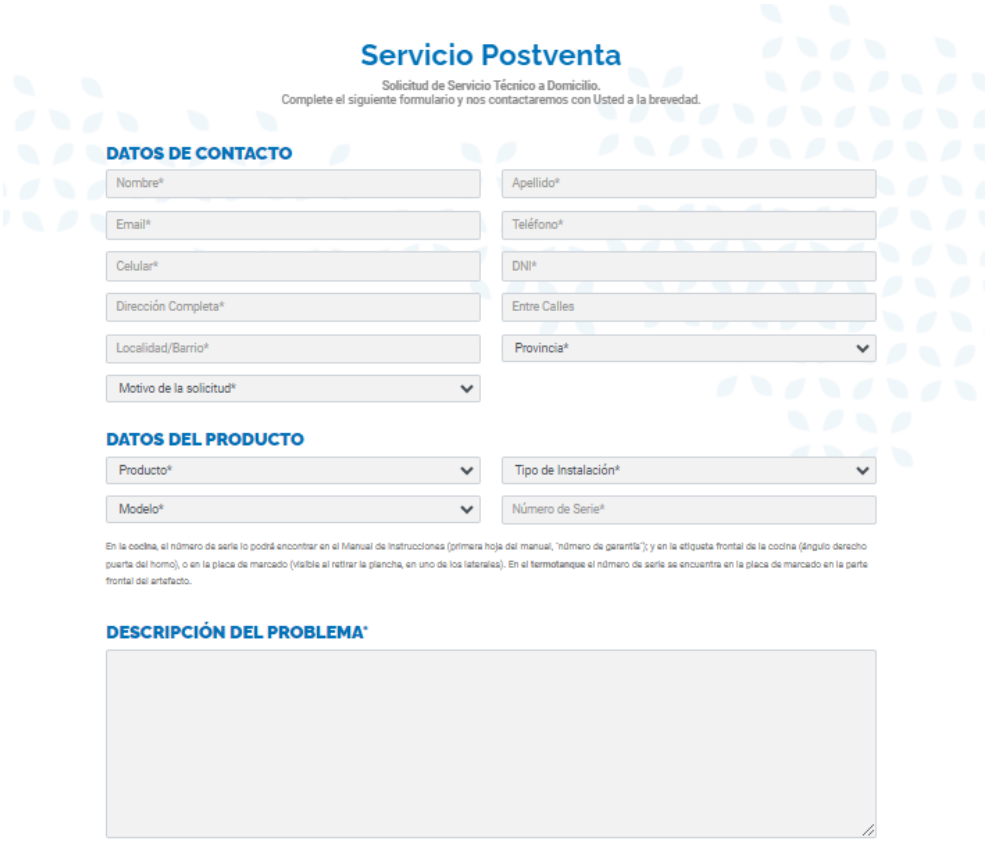


*Ilustración 1: Formulario reclamos - página web.*

En base a la observación que escriba el cliente, luego el operador debe clasificar el problema que tuvo para que luego el técnico lo resuelva realizando lo que por tabla estará tabulado. Un ejemplo a esto puede ser que el cliente escriba “Buenas tardes, mi cocina cuando el horno lo prendo, se apaga.”, el operador buscara del listado problemas de cocinas el código relacionado “C3 – Horno se apaga” que tendrá como solución sugerida cambiar la termocupla y volver a probar.

El operador cuenta con muchas tareas operativas, para agilizar esta tarea se ofrecerá realizar la clasificación del problema para que luego realice únicamente el chequeo de que esté correcto.

**Alcance**

Desarrollar una solución que permita automatizar el proceso de clasificación de un reclamo en función de la información obtenida del formulario web.

**Hoja de ruta: Pasos tentativos.**

1) Obtener los datos de la base de datos donde se aloja la página web de la empresa
2) Exportarlos como csv o documento de excel para su posterior procesamiento
3) Limpieza de los datos dejando solamente las clasificaciones realizadas por los operadores
4) Análisis de datos faltantes o incorrectos y su manipulación
5) Creación del modelo a partir de las observaciones y clasificaciones
6) Realizar el test y medir el nivel de precisión del modelo
7) Implementar el script correspondiente para que realice la clasificación correspondiente y se aplique sobre la base de datos
8) Presentación de la mejora a la gerencia

**Expectativas**

Por cada tipo de producto existe una lista de problemas específicos y estimamos que vamos a tener complicaciones en designar la clasificación correcta sin mezclar entre tipo de producto.

Al procesar las observaciones muy probablemente tengamos problemas al clasificarlos en caso de que el cliente escriba de una manera incorrecta.

Un punto importante para considerar es la inexperiencia del equipo en el procesamiento de texto lo cual la va a provocar que el tiempo del proyecto se extienda mas allá del tiempo normal de cualquier proyecto de esta índole hasta lograr obtener el modelo correcto.

**Potenciales errores**

Para el desarrollo del modelo predictivo basado en text mining en la empresa en la que se propone solventar la problemática descripta se pueden observar que los posibles errores en el desarrollo del proyecto pueden ser lo siguientes:

1. Preprocesamiento Incompleto o Incorrecto: En esta etapa se elimina el ruido que pueden traer los datos antes de iniciar el análisis, el desarrollar esta etapa de forma incorrecta puede repercutir en procesar datos que no resulten en una respuesta al problema que se tiene. Errores comunes incluyen no eliminar stopwords (palabras comunes sin mucho significado) o no normalizar el texto (convertir a minúsculas, por ejemplo).
2. Selección de Características Inadecuada: Elegir las características adecuadas del texto es esencial. Se puede utilizar técnicas como TF-IDF (Frecuencia de Término-Inverso de Documento) o word embeddings (incrustaciones de palabras) para representar las palabras, pero elegir el método incorrecto puede llevar a resultados ineficientes o inexactos.
3. Sobreajuste: Debido a la alta dimensionalidad de los datos de texto, el modelo puede sobreajustarse fácilmente a los datos de entrenamiento. Es importante utilizar técnicas de regularización y validación cruzada para evitar el sobreajuste.
4. No Considerar Contexto: El significado de una palabra puede variar según el contexto. Ignorar el contexto puede llevar a malentendidos y predicciones incorrectas.
5. Falta de Validación Cruzada: La validación cruzada es esencial para evaluar la capacidad de generalización del modelo a nuevos datos. No realizar una validación cruzada adecuada puede dar lugar a evaluaciones sesgadas de la precisión del modelo.
6. Falta de Análisis Exploratorio de Datos: No comprender completamente los datos puede llevar a interpretaciones erróneas y a la elección incorrecta de técnicas de procesamiento de texto y modelado.
7. Selección Inadecuada de Algoritmo: La elección del algoritmo de modelado (por ejemplo, clasificadores, regresión, redes neuronales) depende de la naturaleza de los datos y del problema. Elegir el algoritmo incorrecto puede afectar la precisión del modelo.
8. Evaluación Inadecuada: Evaluar el rendimiento del modelo solo en función de una métrica puede conllevar a resultados errados, es importante analizar cual o cuales métricas son relevantes en el proyecto y hacer el seguimiento para evaluar el rendimiento del algoritmo utilizado. Ejemplo de métricas para evaluar proyectos de ciencia de datos son la precisión, exhaustividad, F1-score, matriz de confusión, etc.
9. Generalización Limitada: A veces, un modelo entrenado en un dominio específico puede no generalizarse bien a otros dominios debido a diferencias en el vocabulario y la estructura del lenguaje.
10. Desafíos de Procesamiento del Lenguaje Natural (NLP): La ambigüedad, la polisemia y otros desafíos del NLP pueden afectar la precisión del modelo.

Para mitigar estos errores, es esencial tener un enfoque metodológico sólido, realizar un análisis exploratorio exhaustivo de los datos, realizar validación cruzada y estar dispuesto a ajustar y mejorar el modelo según sea necesario. La iteración y la experimentación son clave para lograr un modelo predictivo preciso y efectivo basado en Text Mining.

**Presupuesto**

El proyecto tendrá una duración estimada de 1 mes siendo la fecha límite de finalización el 19 de agosto de 2023, siendo extensible la entrega a una semana.

El proyecto tiene el costo de USD 2000 + IVA

Cotización dólar BNA tipo vendedor al momento de realizar la facturación.

Pago a 15 días de la fecha de factura.

**Versión del documento**

Versión: 1.0