



Tecnológico de Monterrey

INSTITUTO TECNOLÓGICO DE ESTUDIOS SUPERIORES DE MONTERREY

Escuela de Ingeniería y Ciencias - Ingeniería en Ciencia de Datos y Matemáticas

ServicIA

**Análisis de portafolio de servicios mediante el análisis de sentimiento y
categorización de comentarios de servicios y de satisfacción de
usuarios**

Análisis de ciencia de datos (Gpo 301) - TC2004B.301

Socio Formador: Dirección de Servicio Social | Campus Monterrey | Programa de Calidad

Profesores: Angelina Alarcón Romero & Felipe Castillo Rendón

Equipo 2

| | | |
|---------------------|------------------------------|-----------|
| PM | Ángel Azahel Ramírez Cabello | A01383328 |
| Diseñadora UX/UI | Annette Pamela Ruiz Abreu | A01423595 |
| CDO | Luis Angel López Chávez | A01571000 |
| Científico de datos | Jorge Raúl Rocha López | A01740816 |
| Científico de datos | Franco Mendoza Muraira | A01383399 |

Monterrey, Nuevo León.

12 de junio de 2023

Índice

| | |
|---------------------------------|-----------|
| Resumen | 04 |
| Marco teórico | 05 |
| Comprensión del negocio | 10 |
| Objetivo del negocio | 10 |
| Problema | 10 |
| Solución | 11 |
| Hipótesis | 11 |
| Objetivo del proyecto | 11 |
| Justificación | 12 |
| Mercado potencial | 13 |
| Clientes | 14 |
| Plan de actividades | 15 |
| Comprensión de los datos | 16 |
| Descripción | 16 |
| Exploración | 20 |
| Calidad | 25 |
| Preparación de los datos | 27 |
| Selección | 28 |
| Limpieza | 28 |
| Transformación | 29 |
| Variables nuevas | 31 |
| Técnicas de Modelación | 33 |
| Extracción de características | 33 |
| Definiciones | 33 |

| | |
|--|-----------|
| Metodología | 36 |
| Modelos de aprendizaje supervisado | 38 |
| Redes neuronales | 38 |
| Máquina de vectores de soporte (SVM) | 40 |
| Árboles de decisión | 42 |
| Naïve Bayes | 44 |
| Regresión lineal múltiple | 48 |
| Modelos de aprendizaje no supervisado | 50 |
| Kmeans | 50 |
| Clustering jerárquico aglomerativo | 54 |
| PCA | 55 |
| Discusión de validez | 56 |
| Ajustes | 57 |
| Selección del modelo | 59 |
| Evaluación | 61 |
| Resultados | 61 |
| Proceso | 63 |
| Impacto social | 66 |
| Impacto ODS | 66 |
| Despliegue | 68 |
| Prototipo | 68 |
| Recomendaciones | 70 |
| Recomendaciones al negocio | 70 |
| Recomendaciones para mejorar la satisfacción | 70 |
| Recomendaciones para mejorar los datos | 71 |

| | |
|--------------------------|-----------|
| Recomendaciones técnicas | 72 |
| Siguientes pasos | 74 |
| Bibliografía | 76 |

Resumen

El presente informe académico presenta un estudio detallado sobre el análisis de las respuestas obtenidas en la encuesta de conclusión del servicio social realizada en el Instituto Tecnológico y de Estudios Superiores de Monterrey (TEC de Monterrey). El objetivo principal de este trabajo es utilizar modelos de aprendizaje supervisado y no supervisado para extraer información valiosa y significativa a partir de los datos recopilados en la encuesta.

Para empezar, se emplearon modelos de aprendizaje supervisado como clasificación, árboles de decisión y *SVM* para analizar la relación entre las variables independientes y las variables dependientes que escogimos para representar la satisfacción del usuario. A través de estos modelos se pudo determinar qué factores influyeron significativamente en las respuestas de los estudiantes y cómo se relacionaron entre sí. Además, se llevó a cabo un proceso de preprocesamiento de los datos para resolver posibles problemas de calidad de datos. Se aplicaron técnicas de aprendizaje no supervisado como el modelo *K Means* y *PCA* con el fin de identificar patrones y agrupaciones naturales en las respuestas. Esto permitió una comprensión más profunda de las opiniones y experiencias de los estudiantes durante su servicio social.

Los resultados obtenidos revelaron patrones interesantes y tendencias significativas en las respuestas de los estudiantes. Se identificaron grupos de organizaciones con características similares y se encontraron factores clave que influyen en la satisfacción general. Finalmente, con base en los resultados obtenidos, se presentan recomendaciones para la organización formadora para mejorar las experiencias y aumentar la satisfacción de los estudiantes.

Marco teórico

Los modelos de clasificación son modelos que dadas variables independientes este puede predecir el valor de una variable dependiente categórica. Esta técnica emplea un árbol de decisiones o los algoritmos de clasificación basados en redes neuronales. Para crear un modelo de clasificación se requiere de la fase de preparación, de aprendizaje y de evaluación y clasificación. En la fase de preparación se separan los datos en datos de aprendizaje (70 %) y en datos de prueba (30 %). En la fase de aprendizaje se crea el modelo con los datos de aprendizaje. Finalmente, en la fase de clasificación se pone a prueba el modelo usando los datos de prueba y se evalúa. (Zárate-Valderrama, Bedregal-Alpaca, & Cornejo-Aparicio, 2021)

La generación automática de etiquetas para permitir una abstracción de información de manera más rápida. Esto es vital cuando no se pueden entrenar a la perfección los algoritmos de extracción de información debido a una falta de etiquetado o clasificación por parte de los investigadores o recopiladores de información. La generación de grupos por medio de aprendizaje no supervisado permite dividir a un conjunto de datos sin requerir tanto tiempo de búsqueda o comparación. (Yan, R., Jiang, X., Wang, W. et al. 2022)

El procesamiento de lenguaje natural (NLP) es una tecnología de machine learning que ayuda a las computadoras a interpretar, manipular y comprender el lenguaje humano. (aws, s.f.) Esta tecnología es usada por chatbots, máquinas de traducción, asistentes tecnológicos y para encontrar patrones. Una técnica nueva que se implementó en el NLP fue el modelo Transformer el cual funciona con capas recurrentes llamadas capas de atención. Estas capas de atención codifican cada palabra de una frase en función del resto de la secuencia. Esto permite que la máquina tenga una representación matemática del contexto de la oración. Esta técnica para entender el significado de las palabras y frases se llama embedding. (Vaca, s.f.) Al igual que las redes recurrentes, los transformadores también tienen

dos bloques principales: codificador y decodificador, cada uno con un mecanismo de autoatención. (Keita, 2022). Ejemplos de uso de la tecnología NLP se pueden encontrar en cualquier plataforma de servicio que desea mejorar su calidad mediante la retroalimentación activa de sus usuarios, así como lo realizaron un grupo de científicos de datos que analizaron 515,000 comentarios de clientes de hoteles de lujo en Europa para encontrar los aspectos más importantes para ellos al momento de agendar una estancia en cierta residencia (Bernardes, 2023).

El modelado de temas es un método para la clasificación no supervisada de documentos, similar a la agrupación en clústeres de datos numéricos, que encuentra algunos grupos naturales de elementos (temas) incluso cuando no estamos seguros de lo que estamos buscando. Los métodos de modelado de temas son técnicas inteligentes que se aplican ampliamente en el procesamiento del lenguaje natural para el descubrimiento de temas y la minería semántica de documentos no ordenados. La asignación de Dirichlet latente (LDA) es uno de los métodos de modelado de temas más populares. Cada documento se compone de varias palabras, y cada tema también tiene varias palabras que le pertenecen. El objetivo de LDA es encontrar temas a los que pertenece un documento, en función de las palabras que contiene. Los métodos de modelado de temas basados en LDA se aplican al procesamiento del lenguaje natural, la minería de textos y el análisis de las redes sociales, la recuperación de información. Con LDA podemos clasificar documentos y comentarios y encontrar patrones o temas. (Great Learning Team, 2022) Al tener columnas con textos largos con comentarios y recomendaciones, se puede usar NLP y LDA para identificar palabras clave y determinar las mejoras que se les pueden hacer a los proyectos.

El modelo LDA descubre nuevos temas en documentos que no han sido encontrados a base de palabras. Lo que busca hacer es hacer una lista de temas, la cantidad de temas elegidas por el usuario, clasificando palabras leídas a estos temas. El algoritmo ve un

documento como una combinación de temas, los cuales son una combinación de palabras; el algoritmo empieza asignando temas aleatorios a cada palabra, y mejora la asignación de estos temas mientras las iteraciones aumentan a través de Gibbs sampling. (Great Learning Team, 2022)

Un ejemplo del uso de diferentes técnicas para conseguir información relevante a partir de un análisis de sentimientos fue el estudio realizado por Saragih y Girsang (2017) en donde se analizaron los comentarios de Facebook y Twitter de tres compañías de transporte en Indonesia y se clasificaron en 3 categorías, positivo, negativos y neutrales. Esto a partir de una biblioteca de palabras de sentimiento y se creó un sistema de puntuación para medir qué tantos comentarios de cada tipo había. Posteriormente los resultados fueron comparados con el número de seguidores de cada una de las páginas de redes sociales y se midió la correlación entre los clientes con cada empresa. (Ahmed, ElKorany, & ElSayed, 2022)

Para realizar el análisis de lenguaje natural (NLP) se requiere de un preprocesamiento. Este proceso empieza con la segmentación o tokenización de los comentarios, el cual consiste en cortar todos los comentarios en palabras individuales eliminando todos los signos de puntuación. Al realizar esto se remueven todas las palabras vacías, es decir, aquellas palabras que aportan por su cuenta nuevo significado o significan ruido en las oraciones a las que pertenecen, como algo, aquí, cuál, etc. Para esta sección ya existen diccionarios que contienen las diferentes palabras vacías en diferentes idiomas como aquellos encontrados en librerías de Python como NLTK. El siguiente paso consiste en pasar todas las palabras relevantes o que aportan al entendimiento de cada uno de los comentarios a sus formas raíz para tener todas las palabras normalizadas, pues palabras como disgusto y disgusta aunque sean diferentes se da a entender el mismo mensaje. Todo esto se puede realizar con la librería mencionada de NLTK. Al lograr procesar los comentarios de los clientes de esta manera, se encuentra la relevancia de cada uno de los términos al ser comparado con el total de

comentarios. Esto se puede realizar usando técnicas como Frecuencia de términos – Frecuencia inversa del documento, o TF – IDF por sus siglas en inglés, explicada en el artículo. Para el componente TF, a cada una de las palabras se calcula su frecuencia absoluta de la palabra w_i en el documento d_j . Este componente es multiplicada por la componente IDF que se refiere a la frecuencia inversa del documento (en este caso comentarios), indicada por las siguientes fórmulas:

$TF(w_i, d_j)$: frecuencia de la palabra w_i en el documento d_j

$IDF(w_i)$: frecuencia inversa del documento de la palabra w_i en el ensamble de comentarios

$$TFIDF(w_i, d_j) = TF(w_i, d_j) \cdot IDF(w_i)$$

$$TFIDF(w_i, d_j) = TF(w_i, d_j) \cdot \log\left(\frac{D}{n_i}\right)$$

n_i : Cantidad de comentarios que incluye la palabra w_i

D : Cantidad total de documentos o comentarios

Esta métrica indica qué tan relevante es una palabra en un ensamble de comentarios, dando a entender cuales son las palabras clave al momento de los clientes dar sus comentarios. Todo este proceso se vuelve relevante y directamente aplicable a nuestro proyecto pues consiste en el procesamiento a los que se debe someter los comentarios para posteriormente categorizarlos y realizar un análisis de sentimiento mediante, por ejemplo, métodos de aprendizaje no supervisado para clusterizar aquellos comentarios con mayor similitud una vez teniendo los valores de TF – IDF para cada uno de las palabras en sus respectivos comentarios. (Piris, Y., & Gay, A. C., 2021)

Además, es importante conocer estrategias para lidiar con valores vacíos en las bases de datos. Una de las formas de mitigar el efecto de estos casos es hacer primero un análisis exhaustivo para feature selection. (Himmi, A. et al, 2023) Primero se eligen cuidadosamente las variables a usar para el entrenamiento de un modelo teniendo en cuenta el comportamiento de sus datos. Por ejemplo, si una variable parece ser relevante para la

construcción del modelo y consiste casi en su totalidad de valores vacíos, entonces seguramente no es suficiente o apropiado para usar en el modelo. En otros casos es indispensable utilizarlas, por lo que técnicas de interpolación son necesarias.

Después de crear los modelos, es muy importante saber cómo evaluarlos. Uno de los mejores métodos de evaluación para los métodos de aprendizaje no supervisados es el silhouette score, el cual tiene un dominio de [-1,1] que indica qué tanto los registros de un mismo grupo o cluster se encuentran cercanos entre sí mientras se alejan de los clusters más lejanos; es decir, mide que la agrupación es coherente. (Zimmermann, A, 2020)

Comprensión del negocio

Objetivo del negocio

El programa de calidad de la Dirección de Servicio Social del campus Monterrey es el departamento de la coordinación de servicio social que se encarga de supervisar la calidad de los proyectos que ofrece el campus Monterrey. Tiene como objetivo garantizar la satisfacción de los alumnos que realizan los proyectos solidarios y de las organizaciones socio formadoras que trabajan con la universidad. Para lograrlo, este departamento realiza una encuesta a los estudiantes cada vez que acaban un proyecto. Esta encuesta contiene diferentes preguntas relacionadas a la satisfacción del usuario, a las características de la organización y a los comentarios que los alumnos quieran compartir con la institución. Aprovechar la información de esta encuesta permitirá que el departamento mejore los proyectos existentes y busquen proyectos con las características más deseadas por los estudiantes, lo cual mejorará el programa.

Problema

En el Instituto Tecnológico y de Estudios Superiores de Monterrey, es necesario cumplir con ciertas horas de servicio social para poder graduarse. La universidad tiene una gran oferta de proyectos solidarios diversos y emocionantes que se pueden inscribir. Una de las herramientas que utiliza la escuela para medir y controlar la calidad de los proyectos que se ofrecen es la implementación de encuestas para conocer las opiniones y perspectivas de los estudiantes sobre la experiencia que acaban de vivir.

Los datos recopilados en esta encuesta no están siendo aprovechados. Aumentar la satisfacción de los alumnos en sus proyectos solidarios no solo los beneficia a ellos, sino también beneficia a las organizaciones socio formadoras; ya que los alumnos trabajarán con entusiasmo, entregarán mejores trabajos y demostrarán un alto compromiso en los proyectos.

Además, garantizar la satisfacción hará que los alumnos realmente pongan en práctica los principios y valores que se quieren promover. Para aprovechar la información recopilada, se requiere categorizar los comentarios y determinar el nivel de satisfacción a través de variables numéricas y cualitativas de los usuarios en cuanto a las experiencias de los servicios sociales promovidas por la Dirección de Servicio Social en Campus Monterrey.

Solución

Hacer un análisis exhaustivo de los resultados de la encuesta usando estadística descriptiva, modelos de aprendizaje supervisado, modelos de aprendizaje no supervisado y procesamiento de lenguaje natural (NLP) para hacer un análisis de sentimiento de los comentarios y garantizar la satisfacción del usuario.

Hipótesis

Existen tendencias de respuestas y sentimientos de los alumnos hacia las organizaciones socio formadoras en los comentarios que responden en la encuesta. Mediante la implementación de modelos de aprendizaje supervisado y aprendizaje no supervisado, se pueden hallar áreas de oportunidad en los proyectos de Servicio Social para mejorar la experiencia general.

Objetivo del proyecto

Crear un modelo de análisis de respuestas, comentarios y sentimiento para poder calificar diferentes proyectos y mejorar la calidad del servicio social.

- a) Identificar posibles mejoras para cada servicio.
- b) Identificar las variables más importantes para la satisfacción de un estudiante.

¿Qué conforma un buen servicio social?

- c) Hacer un análisis de sentimiento basado en los comentarios.
- d) Crear un modelo de clasificación para predecir si un servicio será satisfactorio o no dadas ciertas variables o palabras claves.
- e) Crear un modelo de aprendizaje no supervisado para agrupar organizaciones socio formadores que tienen características en común y analizarlas.

Justificación

En el Instituto Tecnológico y de Estudios Superiores de Monterrey es obligatorio cumplir con 480 horas de servicio social para poder graduarse, pero es muy importante analizar la satisfacción de los estudiantes en cada proyecto solidario para mejorar este servicio y asegurar que los alumnos hagan todas sus tareas con entusiasmo, compromiso y que entreguen resultados útiles y buenos. Para lograr esto se requiere el uso de modelos de aprendizaje supervisado y no supervisado porque estos permiten mejorar la comprensión de los datos al identificar patrones, tendencias y relaciones ocultas entre las variables presentes en la encuesta. Además, mediante técnicas de clustering es posible segmentar a las organizaciones en grupos con características y necesidades similares, lo que facilita la personalización de futuras experiencias y la mejora de la satisfacción general. Asimismo, los modelos de aprendizaje supervisado, como la regresión o clasificación, brindan la capacidad de predecir la satisfacción en base a variables predictoras, identificando los factores más influyentes en la percepción positiva o negativa de los proyectos. Estos modelos también ayudan a identificar los factores clave que afectan la satisfacción, permitiendo a la institución educativa enfocar sus esfuerzos en áreas específicas de mejora. Por último, los resultados obtenidos a través de estos modelos respaldan la toma de decisiones basada en datos, proporcionando una guía objetiva para implementar estrategias efectivas y evaluar su impacto en la satisfacción de los participantes.

Mercado potencial

El mercado potencial de un proyecto que entrena modelos de aprendizaje supervisado y no supervisado para el análisis satisfacción es muy amplio; ya que es de utilidad para cualquier empresa que quiera tomar en cuenta los comentarios de sus clientes e implementar las mejoras necesarias para garantizar la satisfacción de ellos.

En primer lugar, las instituciones educativas están cada vez más interesadas en comprender y mejorar la calidad de los programas o cursos que ofrecen a sus estudiantes. Este proyecto les sería de gran utilidad, ya que es una herramienta poderosa para extraer información valiosa de los datos recopilados y tomar decisiones informadas sobre cómo mejorar la experiencia de los participantes. El mercado actual demanda enfoques innovadores y basados en datos para la toma de decisiones. Los modelos de aprendizaje supervisado y no supervisado se alinean perfectamente con esta necesidad, ya que permiten un análisis objetivo y riguroso de los datos de la encuesta. Al utilizar estas técnicas, las instituciones educativas pueden identificar patrones, tendencias y relaciones ocultas que de otra manera podrían pasar desapercibidos. Además de las instituciones educativas, las organizaciones dentro de la industria de servicios y organización de eventos pueden aprovechar esta tecnología y aplicar el modelo creado para entender la perspectiva y las opiniones de los asistentes o clientes. Esto los ayudará a identificar las áreas de oportunidad de los servicios ofrecidos y a observar cuáles servicios son los más queridos por los asistentes.

Finalmente, el uso de este proyecto sería un gran diferenciador competitivo para cualquier organización, especialmente una universidad; ya que podrá demostrar su compromiso con la mejora continua y la atención a las necesidades de sus estudiantes o clientes.

Clientes

Algunos de los posibles clientes para este proyecto son:

- Khan Academy: Una organización sin fines de lucro que se dedica a proporcionar educación en línea de calidad de forma gratuita. Khan Academy podría usar este proyecto para evaluar la efectividad de sus programas y mejorar la experiencia de sus usuarios.
- Google for Education: Al igual que Khan Academy, la división educativa de Google podría estar interesada en utilizar estos modelos para mejorar sus herramientas de análisis de datos y proporcionar información más detallada sobre la satisfacción de los estudiantes.
- Nielsen: Nielsen es una empresa líder en investigación de mercado y medición de audiencias que podría utilizar estos modelos para analizar la satisfacción de los participantes en proyectos y ayudar a sus clientes a comprender mejor las preferencias y necesidades de los consumidores.
- AccorHotels: Esta cadena hotelera internacional que se enfoca en la calidad de la experiencia del cliente podría estar interesada en utilizar estos modelos para analizar la satisfacción de los huéspedes en proyectos de turismo sostenible y responsabilidad social, con el fin de mejorar sus servicios y personalizar la experiencia del cliente.
- Amazon: La empresa estadounidense podría estar interesada en utilizar este proyecto para analizar la satisfacción de los clientes en sus compras e identificar mejoras específicas para cada grupo de vendedores.

Plan de actividades

Como se ve en la Figura 1, para lograr los objetivos planteados, se seguirá la metodología de proceso intersectorial estándar para la extracción de datos, también conocida como CRISP-DM por sus siglas en inglés.

| 22 de mayo | 29 de mayo | 5 de junio | 12 de junio | 16 de junio |
|-------------------------|----------------------|--------------------------|----------------|--------------------------------|
| Comprensión del Negocio | Preparación de Datos | Modelación Evaluación | Implementación | Presentación al socio formador |
| Comprensión de Datos | | | | |



Figura 1: Línea de tiempo del plan de actividades

Compreensión de los datos

Descripción

Para la creación del modelo se utilizaron cinco bases de datos proporcionadas por la organización socio formadora. Como se ve en la Tabla 1, al unir los cinco archivos que obtuvimos el siguiente dataset:

- Filas: 10499
- Columnas: 43

| Nombre | Descripción | Tipo de dato | Posibles valores | Valores nulos |
|------------------------|--|--------------------------|--|---------------|
| Fecha de inicio | Fecha y hora en la cual el estudiante empezó la encuesta | numérico (time stamp) | [2021-04-22 13:24:00, 2023-05-13 01:06:00] | 0 |
| Fecha final | Fecha y hora en la cual el estudiante terminó la encuesta | numérico (time stamp) | [2021-04-22 13:26:00, 2023-05-13 01:08:00] | 0 |
| Tipo respuesta | Desde dónde contestó la encuesta | categórico | 'Survey Preview', 'IP Address', 'Spam' | 0 |
| Dirección IP | La dirección IP del dispositivo del estudiante | categórico | Hay múltiples. Ejemplo: 189.219.40.197 | 3 |
| Progreso | El porcentaje de encuesta que el estudiante terminó | numérico | 100, 88 | 0 |
| Duración (en segundos) | Contiene cuánto tiempo se tardó el estudiante en completar la encuesta | numérico | [25, 1298823] | 0 |
| Finalizado | Valor booleano que determina si el estudiante acabó la encuesta | categórico | True, False | 0 |
| Fecha registrada | Fecha y hora en la que se registró el envío de la respuesta | numérico (time stamp) | [2021-04-22 13:26:38.693000, 2023-05-13] | 0 |

| | | | | |
|---|--|------------|---|------------------------|
| | | | 01:08:29.755000] | |
| ID de respuesta | Identificación de la encuesta | categórico | Ejemplo: R_12tdDn1LDz4Bm zD | 0 |
| Apellido del destinatario | Apellido del estudiante | categórico | Hay múltiples. Ejemplo: Toca Balderas | 7283 |
| ID | Segunda identificación de la encuesta | categórico | Ejemplo: @00003 | 0 |
| Datos de referencia externos | La columna está vacía | categórico | NaN | 10499 |
| Latitud de ubicación | Latitud de la ubicación del dispositivo del estudiante | numérico | [-37.828, 59.955] | 4 |
| Longitud de ubicación | Longitud de la ubicación del dispositivo del estudiante | numérico | [-123.1337, 144.9669] | 4 |
| Canal de la distribución | Cómo se contestó la encuesta | categórico | 'preview', 'email', 'anonymous', 'gl' | 1520 |
| Idioma del usuario | Idioma de la encuesta | categórico | 'ES-ES' | 0 |
| Preguntas de satisfacción (8) | Preguntas sobre la satisfacción en diferentes aspectos como responsabilidad, valores, etc. | categórico | '\n5 Muy Satisfecho', '4','3','2', '\n1 Nada Satisfecho' | cada pregunta tiene: 2 |
| Preguntas de sí o no (2) | Si la organización ofreció retroalimentación y si fue interesante | categórico | 'Sí', 'No' | cada pregunta tiene: 2 |
| Comentario para compartir con la organización | Comentario del estudiante para la organización | categórico | "..." | 161 |
| Preguntas de opinión (4) | Preguntas sobre si el estudiante está de acuerdo con algunos enunciados sobre los | categórico | '\n5 Totalmente de Acuerdo', ' 4', ' 3', ' 2', '\n1 Nada de Acuerdo' | cada pregunta tiene: 2 |

| | | | | |
|--------------------------------------|---|------------|--|-------|
| | valores y la ética | | | |
| Comentario para compartir con el Tec | Comentario del estudiante para el Tec sobre la experiencia | categórico | “...” | 1429 |
| Tipo de comentario que hiciste | Categoría del comentario hecho sobre la experiencia | categórico | 'Reconocimiento', 'Área de oportunidad' | 1435 |
| OSF | Nombre de la organización socio formadora | categórico | Hay múltiples. Ejemplo: EAAD - Impulso Urbano | 4695 |
| CRN | Identificación del proyecto de servicio social | categórico | [341, 43269] | 4695 |
| Nombre del destinatario | La columna está vacía | categórico | NaN | 10499 |
| Correo electrónico del destinatario | La columna está vacía | categórico | NaN | 10499 |
| OSF y nombre del proyecto | Nombre de la organización socio formadora y el proyecto en el que participó el estudiante | categórico | Hay múltiples. Ejemplo: PRODAN, Prodefensa Animal, A.C._Préstame tu Casa | 6260 |
| Nombre de Experiencia | Nombre del proyecto en el que participó el estudiante | categórico | Ejemplo: 'Cybersecurity App' | 8707 |
| Periodo | Periodo en el que se realizó el proyecto | categórico | 44593, 2, 202211, 3 | 8707 |
| Tipo de Formato | Formato en el que se realizó el proyecto | categórico | 'Regular' | 8707 |
| Matrícula | Matrícula del estudiante | categórico | Hay múltiples. Ejemplo: A01097190 | 7735 |
| Semana | Modalidad en la que se realizó el proyecto | categórico | '2', '1-2', '1-3', '2-3', '3' | 8720 |

Tabla 1: Descripción de la base de datos unida

Las preguntas de satisfacción del alumno son las siguientes:

1. Evalúa tu nivel de satisfacción en los siguientes aspectos:
 - 1.1. Al concluir este Proyecto Solidario.
 - 1.2. Vivir la experiencia de aprendizaje relacionada con un Objetivo de Desarrollo Sostenible.
 - 1.3. Nivel de valor aportado a la organización socio formadora a través de tus entregables.
 - 1.4. Momentos de interacción y escucha con los beneficiarios / destinatarios del proyecto.
 - 1.5. Herramientas que aplicaste como las actividades, reportes, "quizzes", dentro de la plataforma de CANVAS en el desarrollo de la experiencia de Servicio Social.
 - 1.6. Experiencia de colaboración con la organización socio formadora.
 - 1.7. Seguimiento y liderazgo de la organización socio formadora.
 - 1.8. Atención y servicio del área que administra el Servicio Social en el campus (asesoría y orientación, información puntual, atención de dudas, seguimiento de incidentes).
2. La Organización Socio Formadora ¿ofreció retroalimentación sobre el desarrollo del Proyecto Solidario y tu desempeño?
3. ¿Consideras interesante la causa social del socio formador?
4. Escribe algún comentario que te interese compartir con la organización socio formadora.
5. Expresa tu opinión en los siguientes enunciados, en esta experiencia de servicio social tuve la oportunidad de:

- 5.1. Ser sensible ante la vulnerabilidad, el dolor y el sufrimiento del otro y actuar con el fin de eliminarlo, aliviarlo o evitarlo, a través de acciones justas alejadas de la pasión egoísta y/o de sentimientos de superioridad.
- 5.2. Actuar con responsabilidad, con el fin de asegurar el bienestar de la colectividad, a través de acciones que garantizan el acceso a los derechos humanos, el empoderamiento de los ciudadanos y de las comunidades, así como el cuidado, mantenimiento y uso sostenible de los recursos y bienes comunes.
- 5.3. Actuar con respeto ante la diversidad de género, sexual, étnica, cultural, de capacidades, generacional, religiosa y socioeconómica mostrando una cordial aceptación de las diferencias y la capacidad para gestionar de manera razonable los conflictos.
- 5.4. Promover soluciones cooperativas en problemas o coordinar acciones colectivas con el fin de mejorar la calidad de vida de la sociedad, fomentando la cultura de la legalidad, los derechos humanos y/o el fortalecimiento de la democracia.

Exploración

Como se ve en las Figuras 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 para entender la base de datos realizamos una exploración exhaustiva de ellos.

En la Figura 2 podemos observar que la gran mayoría de los alumnos completan la encuesta en su totalidad, aunque hay algunas excepciones. Además, podemos observar algunas anomalías o peculiaridades en la columna de duración. Esta dice que los encuestados se tardan en promedio aproximadamente 8580 segundos, lo cual equivale a 2.38 horas. Al ser una encuesta muy corta podemos concluir que esta columna puede tener algunos valores

atípicos que se deben al hecho de que los alumnos pueden dejar la encuesta abierta y reanudarla en cualquier momento, habilitando la posibilidad de completar la encuesta días después de empezar.

| | Progreso | Duración (en seg) | Latitud | Longitud |
|--------------|-----------------|--------------------------|----------------|-----------------|
| count | 10499.00 | 10499.00 | 10495.00 | 10495.00 |
| mean | 100.00 | 8582.59 | 25.00 | -99.21 |
| std | 0.23 | 52148.45 | 4.07 | 10.76 |
| min | 88.00 | 25.00 | -37.83 | -123.13 |
| 25% | 100.00 | 119.00 | 25.64 | -100.33 |
| 50% | 100.00 | 190.00 | 25.65 | -100.31 |
| 75% | 100.00 | 361.00 | 25.68 | -100.26 |
| max | 100.00 | 1298823.00 | 59.96 | 144.97 |

Figura 2: Medidas estadísticas descriptivas de las variables cuantitativas

Para comprobar la existencia de valores atípicos en la duración, se crea un diagrama de caja y bigotes y un histograma. En la Figura 3 podemos ver que existe una gran cantidad de valores atípicos (puntos rojos) que están muy alejados de la media. Se puede observar que la gran mayoría de los datos están dentro del rango de 100 y 1000 segundos. La gran cantidad de datos atípicos puede ser explicada por el hecho de que la encuesta se puede reanudar en un tiempo después de empezar, por lo cual las personas pudieron haberse olvidado de terminar la encuesta. Gracias a estos valores atípicos, la columna de duración no es muy confiable y será eliminada posteriormente.

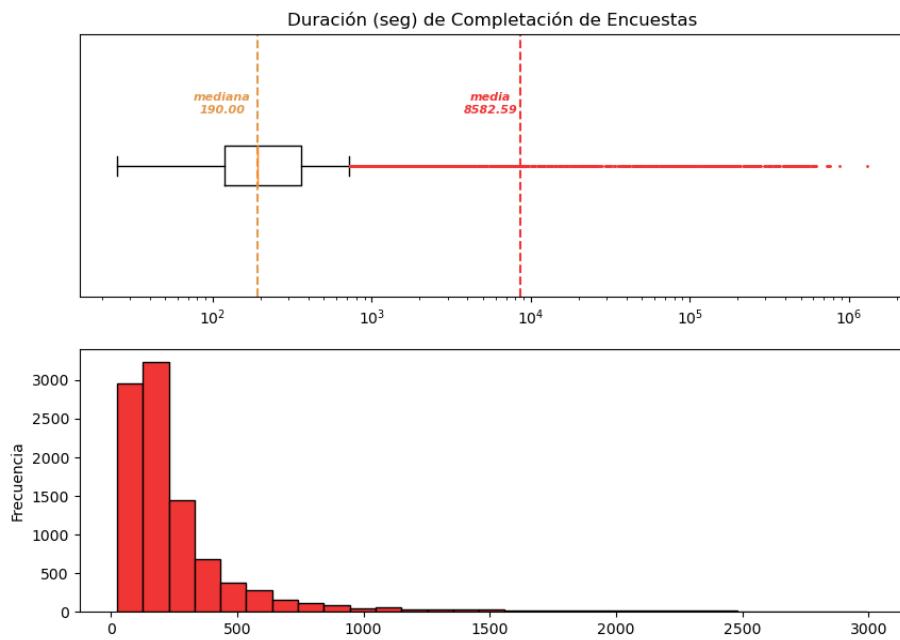


Figura 3: Medidas estadísticas descriptivas de las variables cuantitativas

En el mapa presentado en la Figura 4 se puede observar que la gran mayoría de las encuestas fueron realizadas en México e incluso Estados Unidos. Sin embargo, también existen casos donde se completan hasta en otros continentes como Europa.



Figura 4: Medidas estadísticas descriptivas de las variables cuantitativas

En la Figura 5 podemos observar la moda de las variables cuantitativas y cualitativas. Viendo la columna de duración, vemos que la mayoría de los estudiantes completan la encuesta en 104 segundos (menos de 2 minutos). Asimismo, en todas las preguntas de satisfacción de la encuesta, la respuesta más popular es la de máxima satisfacción con un

valor de 5. Esto puede indicar un sesgo de los participantes para dar una buena respuesta, lo cual explica el hecho de que la mayoría de los encuestados terminan la encuesta de manera rápida.

| | mode | | |
|---------------------------|----------------------------|----------------------------|---------------------------|
| Fecha de inicio | 2021-05-29 10:19:02 | P1.5 | \n5 Muy Satisfecho |
| Fecha final | 2022-11-28 21:45:52 | P1.6 | \n5 Muy Satisfecho |
| Tipo respuesta | IP Address | P1.7 | \n5 Muy Satisfecho |
| IP | 131.178.200.61 | P1.8 | \n5 Muy Satisfecho |
| Progreso | 100 | P2 | Sí |
| Duración | 104 | P3 | Sí |
| Finalizado | True | Comentario para OSF | |
| Fecha registrada | 2021-04-22 13:26:38.693000 | P5.1 | \n5 Totalmente de Acuerdo |
| ID respuesta | R_12tdDn1LDz4BmzD | P5.2 | \n5 Totalmente de Acuerdo |
| Apellido | NaN | P5.3 | \n5 Totalmente de Acuerdo |
| ID | @00001 | P5.4 | \n5 Totalmente de Acuerdo |
| Datos referencia externos | NaN | Comentario general | |
| Latitud | 25.6449 | Tipo comentario | |
| Longitud | -100.311 | NomOSF_Experiencia | NaN |
| Canal de distribución | anonymous | OSF | 444.0 |
| Idioma | ES-ES | CRN | NaN |
| P1.1 | \n5 Muy Satisfecho | Nombre destinatario | NaN |
| P1.2 | \n5 Muy Satisfecho | Correo | NaN |
| P1.3 | \n5 Muy Satisfecho | OSF y nombre proyecto | NaN |
| P1.4 | \n5 Muy Satisfecho | Nombre experiencia | 3.0 |
| | | Periodo | NaN |
| | | Formato | NaN |
| | | Matrícula | NaN |

Figura 5: Moda de las variables cuantitativas y cualitativas

En la tabla de la Figura 6 podemos observar una descripción estadística de las preguntas de satisfacción. Como se muestra en la Figura 7, la pregunta con la media más baja es la P1.5; la cual hace referencia a la satisfacción del encuestado en relación a los quizzes y a las actividades de Canvas. Gracias a esto y a la distribución presentada en la Figura 8, podemos concluir que los alumnos no aprecian los exámenes obligatorios que aparecen en Canvas; ya que en comparación con las otras preguntas, esta es la que más respuestas de 1 y 2 tiene. Finalmente, parece que la mayoría de los encuestados se encuentran satisfechos con el servicio; sin embargo, es importante mencionar que algunos alumnos no contestan la encuesta honestamente; ya que ponen la calificación más alta, pero en los comentarios mencionan que no disfrutaron de la experiencia. Esto se explorará al hacer un análisis de sentimiento.

| | P1.1 | P1.2 | P1.3 | P1.4 | P1.5 | P1.6 | P1.7 | P1.8 | P5.1 | P5.2 | P5.3 | P5.4 |
|------|------|------|------|------|------|------|------|------|------|------|------|------|
| mean | 4.53 | 4.59 | 4.56 | 4.49 | 4.30 | 4.49 | 4.47 | 4.46 | 4.6 | 4.72 | 4.72 | 4.71 |
| std | 0.74 | 0.70 | 0.72 | 0.84 | 0.95 | 0.84 | 0.87 | 0.84 | 0.7 | 0.57 | 0.58 | 0.59 |
| min | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 1.00 | 1.00 | 1.00 |
| 25% | 4.00 | 4.00 | 4.00 | 4.00 | 4.00 | 4.00 | 4.00 | 4.00 | 4.0 | 5.00 | 5.00 | 5.00 |
| 50% | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.0 | 5.00 | 5.00 | 5.00 |
| 75% | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.0 | 5.00 | 5.00 | 5.00 |
| max | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.0 | 5.00 | 5.00 | 5.00 |

Figura 6: Medidas estadísticas descriptivas de las columnas de preguntas de satisfacción

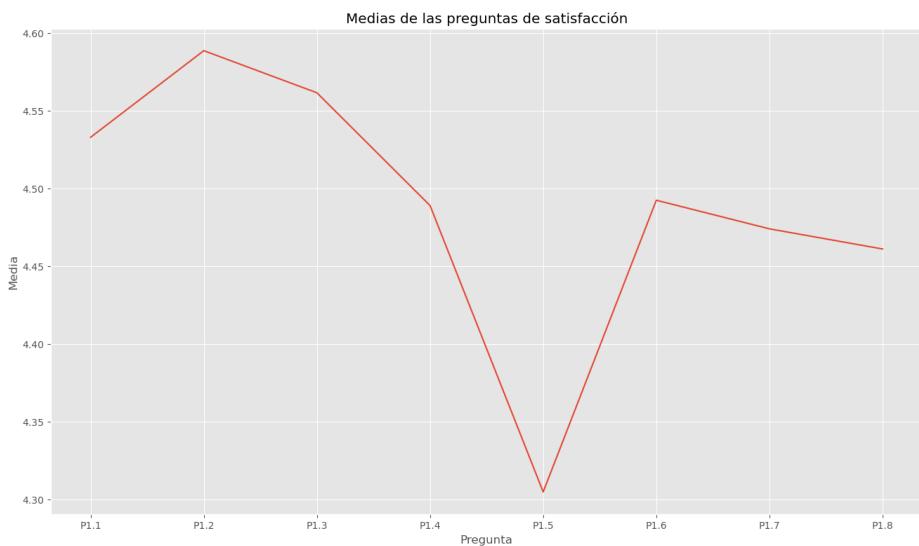


Figura 7: Representación de las medias de las columnas de preguntas de satisfacción

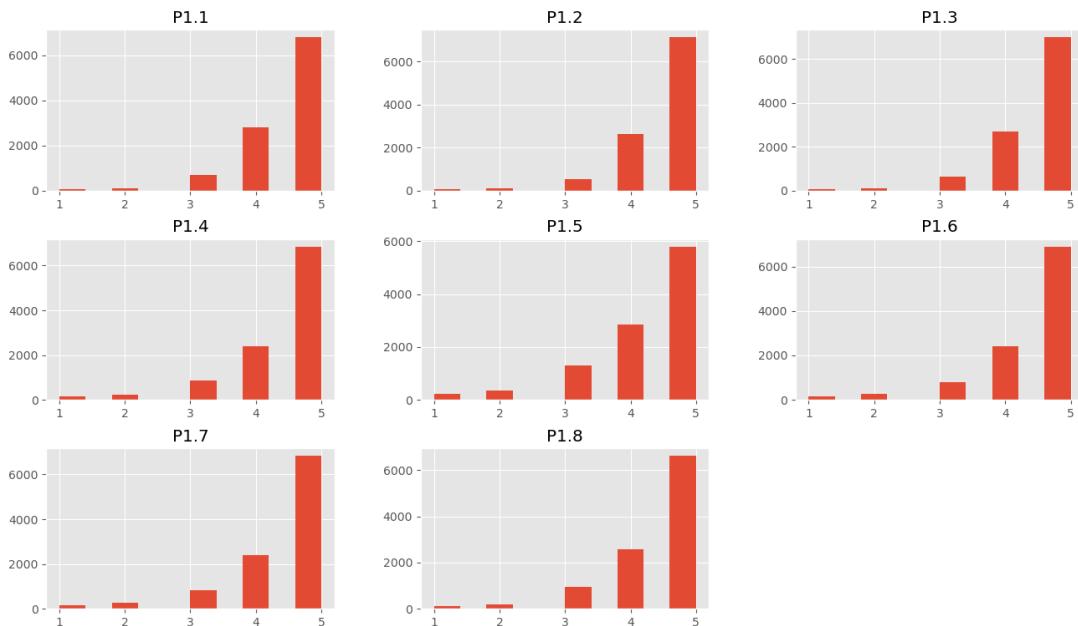


Figura 8: Histograma de las respuestas de las preguntas de satisfacción

En la Figura 9 podemos observar que a la mayoría de los encuestados les parece interesante las causas de los servicios sociales; sin embargo, no todos los encuestados recibieron retroalimentación en sus trabajos. De igual manera, en la Figura 10 podemos observar que la mayoría de los comentarios son positivos, pero hay más de 2000 que son de áreas de oportunidad. Finalmente, la Figura 11 muestra la correlación entre las preguntas de satisfacción. Aquí podemos ver que la pregunta de satisfacción (P1.1) tiene una alta correlación con la pregunta relacionada con un Objetivo de Desarrollo Sostenible (P1.2) y con la pregunta sobre la experiencia colaborativa con el socio formador.

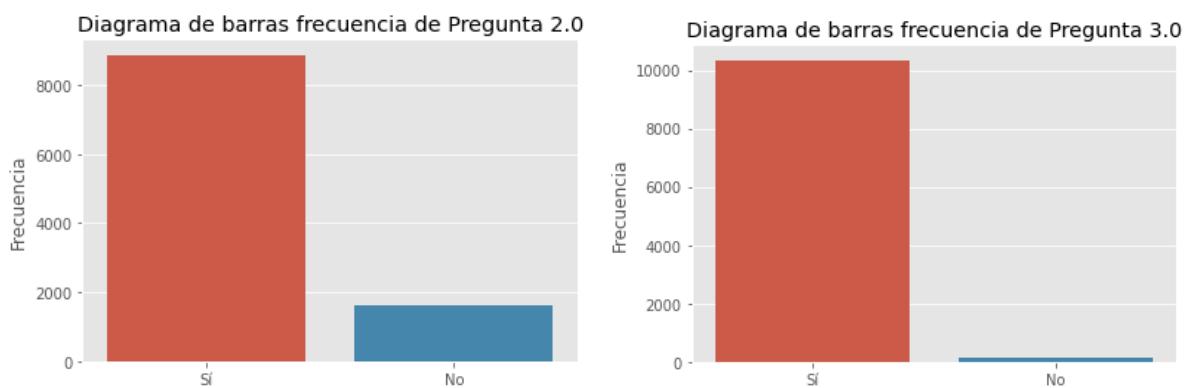


Figura 9: Representación de las preguntas cerradas de tipo “Sí” y “No”

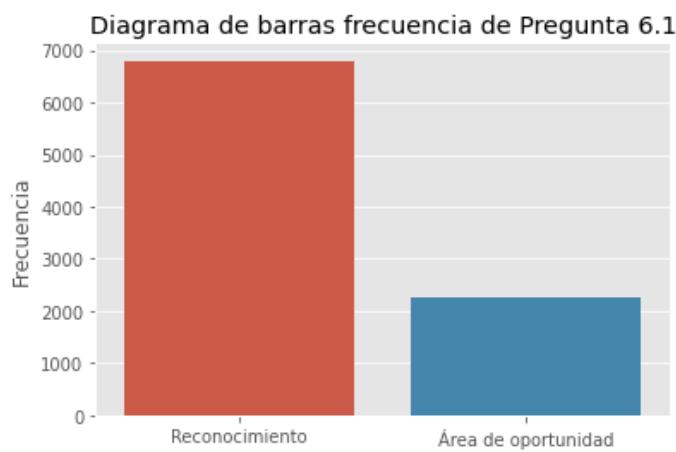


Figura 10: Representación de los tipos de comentarios

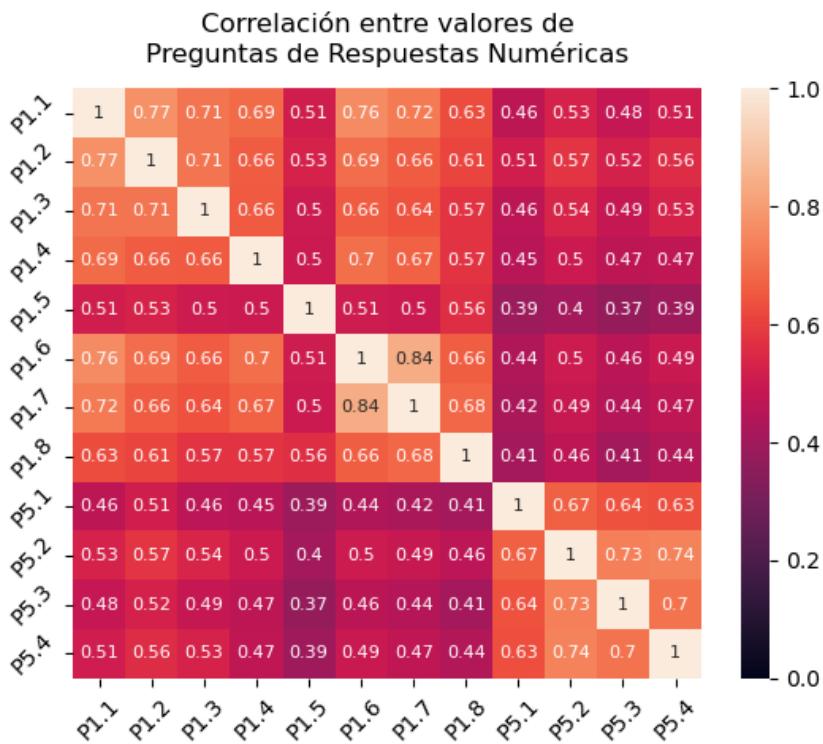


Figura 11: Gráfica de correlación entre las preguntas de satisfacción

Calidad

En un proyecto de análisis de datos es crucial asegurarse de que la fuente de los datos sea confiable, que la muestra sea lo suficientemente grande y diversa y que los comentarios sean claros y comprensibles. También es importante verificar la veracidad de los datos y la cobertura de categorías relevantes. Al incluir esta sección, se garantiza que los datos sean confiables y representativos, fortaleciendo la validez de los resultados y proporcionando una base sólida para la toma de decisiones informadas.

En primer lugar, sabemos que podemos confiar en la fuente de los datos; ya que son primarios y se recopilaron directamente de una encuesta. Con respecto a la calidad de los comentarios, en la Figura 11 podemos observar los 5 comentarios más populares. Evidentemente, estos comentarios no son de utilidad y tendrán que ser ignorados al momento de hacer un análisis exhaustivo de los comentarios.

Diagrama de barras frecuencia de Pregunta 6.0 (Comentarios más comunes)

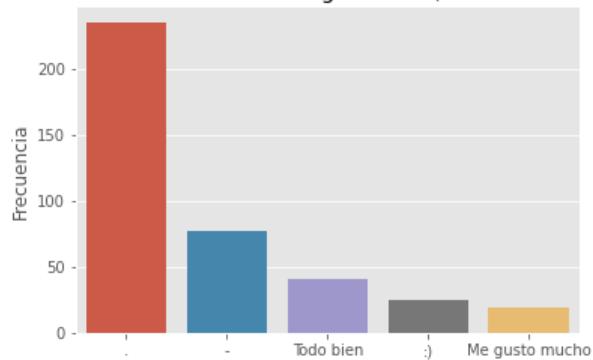


Figura 12: Comentarios más comunes

Preparación de los datos

Selección

Como se mencionó en la descripción de los datos, el dataset tenía 43 columnas; sin embargo no todas tenían información relevante para el análisis y la modelación. Las columnas de interés son las siguientes: “Fecha de inicio”, “Duración (en segundos)”, Preguntas de satisfacción (8), Preguntas de sí o no (2), “Comentario para compartir con la organización”, Preguntas de opinión (4), “Comentario para compartir con el Tec”, “Tipo de comentario que hiciste”, “OSF”, “OSF y nombre del proyecto”. Estas columnas se eligieron porque son las que más se alinean con el objetivo del proyecto y las que tienen menos registros vacíos. Las demás columnas como “Fecha final”, “Tipo de respuesta”, “Dirección IP”, “Finalizado”, “Fecha registrada”, “ID de respuesta”, “Apellido del destinatario”, “ID”, “Datos de referencia externos”, “Latitud de ubicación”, “Longitud de ubicación”, “Canal de la distribución”, “Idioma del usuario”, “CRN”, “Nombre del destinatario”, “Correo electrónico del destinatario”, “Nombre de Experiencia”, “Periodo”, “Tipo de Formato”, “Matrícula” y “Semana” no se eligieron porque no eran relevantes para el análisis y no ayudarían a entrenar los modelos.

Limpieza

La primera parte de limpieza se hizo desde el inicio del trabajo cuando se juntaron los cinco archivos con las respuestas de la encuesta.

Al elegir las columnas con las que queríamos trabajar, eliminamos las demás que no eran relevantes para el proyecto. Además de estas columnas, eliminamos los registros duplicados y los que tenían valores vacíos en las preguntas de satisfacción. También separamos la columna “Fecha de inicio” en “Año” y “Mes” y la columna “OSF y nombre del proyecto” en “OSF” y eliminamos el nombre del proyecto.

Con respecto a las filas, existían dos instancias de personas que no habían contestado del todo la encuesta, faltaban sus respuestas de las preguntas P1.1 a P6.1, por lo que se decidió eliminarlas. Una vez que se tenían los datos y las columnas que se seleccionaron como importantes, se reinició el índice de las instancias, ya que en algunos casos se repetían. Decidimos no eliminar los registros que no tenían comentarios porque sus respuestas de satisfacción ayudarán a entrenar los modelos de aprendizaje.

Transformación

Para facilitar la programación y la visualización de los datos renombramos las columnas para que el dataset final tuviera la la siguiente forma: “Año”, “Mes”, “Progreso”, “Duración”, “P1.1”, “P1.2”, “P1.3”, “P1.4”, “P1.5”, “P1.6”, “P1.7”, “P1.8”, “P2”, “P3”, “Comentario para OSF”, “P5.1”, “P5.2”, “P5.3”, “P5.4”, “Comentario general”, “Tipo comentario”, “OSF” y “Comentario traducido”. Otra modificación importante fue el cambio del tipo de dato de las columnas de preguntas de satisfacción. Antes eran columnas con texto, así que eliminamos el texto agregado y transformamos los datos en números enteros.

Una de las transformaciones más importantes en el proyecto fue la traducción de comentarios. Es necesario traducir los comentarios de español a inglés para realizar procesamiento de lenguaje natural (NLP) en Python por varias razones. En primer lugar, el inglés es el idioma dominante en el ámbito de la programación y la informática, por lo que la mayoría de las herramientas y bibliotecas de NLP están optimizadas para trabajar con texto en inglés. Al traducir los comentarios de español a inglés, se abre la puerta a utilizar estas poderosas herramientas y aprovechar al máximo su funcionalidad. Otro motivo importante es la interoperabilidad y la reutilización del código. Al trabajar con comentarios en inglés, se facilita la integración de bibliotecas y modelos pre entrenados desarrollados por la comunidad de NLP en Python. Muchos de estos recursos están específicamente diseñados para funcionar

con texto en inglés, por lo que al traducir los comentarios se asegura una mayor coherencia y compatibilidad en todo el flujo de trabajo de NLP.

Al hacer esta transformación pudimos analizar los comentarios clasificados como áreas de oportunidad para identificar estas áreas de oportunidad para poder lograr el objetivo y ayudar a la organización socio formadora. En la Figura 13 podemos observar la nube de palabras de los 2263 comentarios que son áreas de oportunidad.

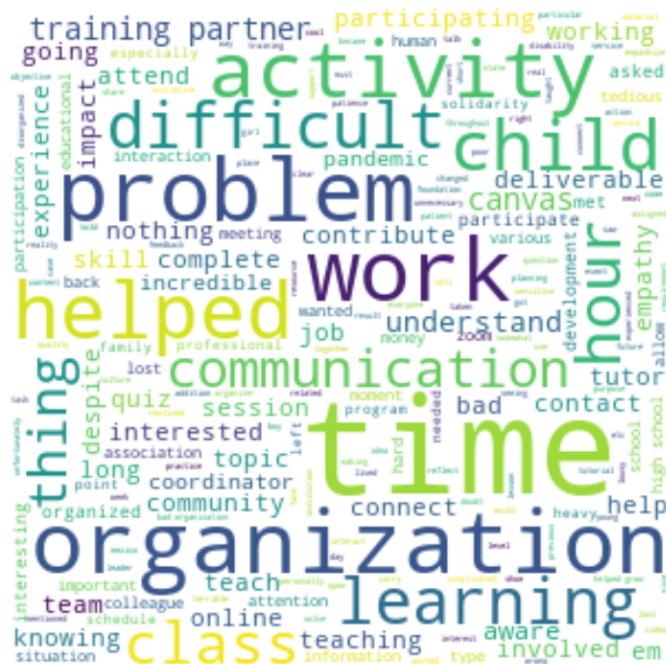


Figura 13: Palabras que más se repiten en los comentarios negativos (con sentimiento menor que 0)

Al analizar la nube de palabras pudimos identificar cuatro categorías de mejora: tiempo, impacto, quizzes y actitud.

| Categoría | Cantidad | Comentarios negativos | Promedio de satisfacción de los negativos (P1.1) | Conclusiones |
|-----------|----------|-----------------------|--|--|
| tiempo | 701 | 265 | 3.8 | Más del 35 % de los alumnos sienten que les encargan demasiado en muy poco tiempo. Estos comentarios sugieren que las organizaciones organicen |

| | | | | |
|---------|-----|-----|-----|--|
| | | | | mejor las actividades o que pidan menos entregables. |
| impacto | 431 | 47 | 3.8 | Solo el 10 % de los encuestados sienten que no causaron un impacto o que su trabajo no sirvió; sin embargo, al no sentir que aportan a la organización, su satisfacción disminuye. |
| quizzes | 504 | 196 | 4.0 | De los 500 comentarios que hablan sobre las actividades, el 40 % cree que esto es un área de mejora. Muchos consideran que son una pérdida de tiempo, repetitivos, innecesarios y que ponen en peligro las horas de trabajo empleadas. |
| actitud | 35 | 12 | 3.1 | Aunque no muchos alumnos hablan sobre la actitud o el comportamiento del socio formador, el 30 % de los comentarios sobre la actitud de las personas encargadas del proyecto son negativas. Además, esta es la categoría que tiene el promedio de satisfacción más bajo. Las relaciones con las personas encargadas y las organizaciones son vitales para la satisfacción de los alumnos. Algunos mencionaron que eran groseros, que los trataban como empleados y que eran irrespetuosos. |

Variables nuevas

Para aprovechar los comentarios traducidos de los encuestados se realizó un análisis de sentimiento. Para ello se utilizó la librería “*SentimentIntensityAnalyzer*” y se hizo un preprocesamiento de los comentarios con la librería “*NLTK*”. En primer lugar, decidimos tokenizar los comentarios. La tokenización consiste en dividir un texto en entidades más pequeñas llamadas tokens. Luego se eliminaron los números y las palabras con menos de dos

letras. Después pasamos todas las palabras a minúsculas y eliminamos las palabras vacías o inútiles; es decir, eliminamos las preposiciones y las palabras de relleno. Finalmente, aplicamos el modelo de análisis de sentimiento de la librería y añadimos la puntuación a una columna nueva llamada “sentimiento”. Este modelo regresa un valor que está en el rango de [-1, 1]; -1 representa un comentario completamente negativo, 1 representa un comentario completamente positivo y 0 representa un comentario neutro. Los comentarios vacíos fueron clasificados como neutros. En la Figura 14 se ve la frecuencia de sentimiento de los comentarios generales y podemos observar que el promedio de sentimiento es de 0.43; es decir, positivo.

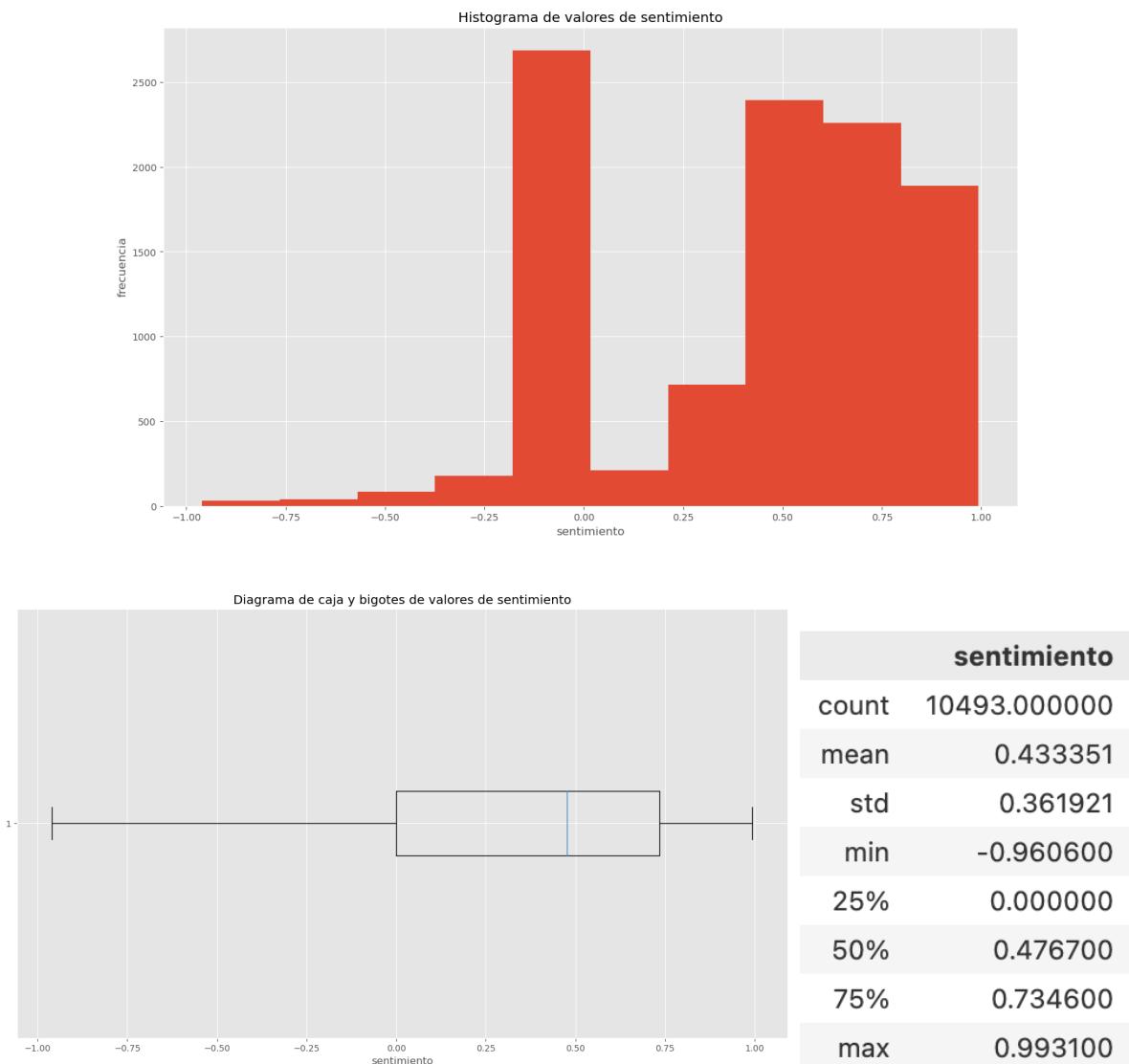


Figura 14: Análisis de la columna de sentimiento de los comentarios generales

Técnicas de Modelación

Extracción de características

Para la realización de este trabajo se hace uso de modelos de aprendizaje automático, por lo que establecer las variables a usar para lograr cualquier objetivo es primordial. Tomando en cuenta que en este proyecto se define como uno de los objetivos lograr ordenar las diferentes organizaciones socio formadoras (OSF) según la satisfacción de sus participantes, las primeras variables consideradas fueron aquellas preguntas en la encuesta relacionadas a la opinión de los encuestados hacia ciertos aspectos específicos de los proyectos encontrados en la Dirección de Servicio Social. Estas preguntas fueron sobre la calidad de atención y servicio del área administrativa, si la OSF ofreció retroalimentación sobre el desarrollo de los proyectos, y si se consideró interesante sus causas sociales.

Por otro lado, se crearon nuevas variables a partir de datos ya existentes. Como se mencionó, se aplicaron técnicas de análisis de sentimiento en los comentarios de los encuestados para encontrar la actitud o tono de los comentarios de los usuarios hacia la OSF para determinar si se trata de uno positivo o negativo.

Definiciones

Una vez teniendo establecido las variables dependientes e independientes a utilizar, se debe encontrar el mejor modelo que capture adecuadamente la tendencia general de los datos. En este caso se utilizan ambos tipos de modelos de aprendizaje automático, aprendizaje supervisado y no supervisado.

Para el caso de modelos supervisados, donde se hace uso de variables independientes, se intenta predecir una variable objetivo usando como entrenamiento ejemplos donde se conocen las etiquetas explicadas por las variables independientes visto como un vector de característica numéricas. Uno de los ejemplos más simples de este tipo de modelos son los de

clasificación como el método *K-Nearest Neighbors*, donde se sitúan los datos de entrenamiento en un espacio según los valores de sus características y se crean regiones delimitadas según las K cantidad de instancias más cercanas según sus etiquetas (Wu, Ianakiev, Govindaraju, 2002).

Por otro lado, los modelos no supervisados consisten principalmente en la agrupación o creación de *clusters* donde se encuentren elementos similares entre sí, pues se desconocen las etiquetas que describen los vectores de datos. Uno de los ejemplos más comunes de este tipo de modelos es el *K-Means*. En este método se definen K número de centroides en cualquier posición inicial en el espacio de característica de los datos de entrenamiento donde cada uno de ellos se le asigna al *cluster* más cercano y en cada iteración la posición de los centroides se actualizan según el promedio de las características de los datos en cada uno de las categorías hasta que su posición se mantenga constante (Pham, Dimov, Nguyen, 2005).

Una vez conociendo los conceptos generales de los modelos de aprendizaje automático, se entiende que existen variaciones de un mismo modelo según se definan parámetros que ajustan al momento de ser entrenados con datos. Por ejemplo, las K cantidades de puntos más cercanos para clasificar una instancia en el modelo de *K-Nearest Neighbors*, e incluso la K cantidad de clusters a crear en el modelo de *K-Means*. Estos valores a definir son conocidos como los hiperparámetros de cualquier modelo de aprendizaje automático. Otros posibles casos consisten en las diferentes definiciones de distancias entre las instancias, pesos dados a las clases de los datos para favorecer una clase en específico al momento de entrenar, o el número de divisiones a realizar en un modelo de árboles de decisión. Los resultados de cualquier modelo son afectados por el valor de sus parámetros, por lo que encontrar el conjunto óptimo es de suma importancia sin caer en *overfitting* o *underfitting*, casos donde el modelo se ajusta de más o de menos respectivamente a los datos

de entrenamiento fallando en capturar el comportamiento general de los datos a presentarse en un espacio exterior de evaluación.

Al evaluar modelos de aprendizaje, se utilizan diversas métricas para medir su rendimiento y efectividad. Algunas de las métricas más comunes incluyen la precisión, la sensibilidad, la especificidad y el valor F1. La precisión es la proporción de instancias correctamente clasificadas sobre el total de instancias clasificadas. La sensibilidad, también conocida como tasa de verdaderos positivos o recall, es la proporción de instancias positivas correctamente identificadas. La especificidad mide la capacidad del modelo para identificar correctamente las instancias negativas. El valor F1 combina la precisión y la sensibilidad en una sola métrica que proporciona una medida general del rendimiento del modelo. Estas métricas permiten evaluar aspectos como la capacidad de predicción correcta de clases positivas y negativas, así como el equilibrio entre falsos positivos y falsos negativos. Es importante seleccionar las métricas apropiadas según el problema y el contexto específico, ya que cada métrica proporciona una perspectiva única del rendimiento del modelo. (Mora, 2022)

Metodología

Con el objetivo de encontrar y seleccionar los mejores modelos se pone en práctica lo anteriormente descrito siguiendo la metodología que está a continuación para asegurar encontrar los mejores modelos posibles. Las variables a utilizar para el entrenamiento de modelos supervisados son todas las preguntas con respuestas numéricas indicando la satisfacción del encuestado en cuanto su experiencia y la columna nueva de sentimiento, mientras la variable objetivo es la pregunta 1 (P1.1), la cual consiste en la satisfacción general del usuario y tiene un valor entero entre 1 y 5. Para el caso de los modelos no supervisados utilizados, se hace uso principalmente de los comentarios hacia la OSF por parte

de los encuestados y los comentarios generales de los usuarios en cuanto a su experiencia con los proyectos ofrecidos. Debido a la escasez de herramientas rigurosas para el análisis de comentarios y texto en español primero se traducen los comentarios a analizar al inglés mediante el uso de *Google Translate*, así habilitando el uso de un análisis más exhaustivo. Sin embargo, es importante mencionar que cierta precisión del significado de los textos se pierde debido al cambio de idioma.

Para cada uno de los modelos a evaluar se usan las mismas variables de entrenamiento para poder comparar los resultados de manera justa, esto aplica para los modelos supervisados y no supervisados respectivamente. Se eligen una variedad de modelos candidatos para ambos tipos de aprendizajes intentando ser lo más diferentes entre ellos para obtener una mayor perspectiva al compararlos. Igualmente, para el entrenamiento específico se usaron diferentes semillas al momento de realizar la separación de los datos de entrenamiento y de evaluación para obtener una mayor perspectiva del desempeño de cada modelo. Asimismo, cada uno de los modelos diferentes serán utilizados variando los valores de sus hiperparámetros, donde dependiendo de los resultados se concluye entre cuál combinación de valores se obtienen los mejores resultados para cada modelo.

Las métricas utilizadas para la evaluación de los modelos supervisados fueron la precisión, exactitud, sensibilidad, y el puntaje *f1* para evaluar la capacidad del modelo para asignar correctamente la satisfacción general del encuestado dependiendo de ciertas variables independientes. Por otro lado, para medir la calidad de los *clusters* generados se usarán valores como el *silhouette score* para entender qué tan bien se categorizan y segmentan los comentarios de los encuestados. Al realizar la evaluación de los modelos, se eligen aquellos con los mejores desempeños al momento de capturar el comportamiento general de los datos.

Para crear los conjuntos de entrenamiento y de prueba se utilizó el método de validación *Holdout*; ya que teníamos suficientes datos. Este enfoque implica la separación de

los datos utilizando la función *train_test_split* de la biblioteca *sklearn*. Esta función realiza un muestreo aleatorio de los datos y divide la información relevante para nuestro modelo en cuatro variables distintas: variables de entrenamiento (variables independientes y variable dependiente) y variables para evaluar el rendimiento del modelo (variables independientes y variable dependiente). El conjunto de entrenamiento se utiliza para alimentar los modelos y permitirles aprender los patrones y relaciones en los datos. Por otro lado, el conjunto de prueba se utiliza para evaluar qué tan bien se comporta el modelo al predecir valores desconocidos. Al tener datos no vistos previamente, podemos obtener una estimación del rendimiento del modelo en situaciones reales. Al usar este método de validación, 80 % de los datos se usaron para entrenar a los modelos y 20 % se usaron para evaluarlo y hacer las predicciones.

Modelos de aprendizaje supervisado

Redes neuronales

Uno de los modelos de aprendizaje supervisado que se usó fue redes neuronales para predecir las clasificaciones de satisfacción de los usuarios. Se trata de un tipo de proceso el cual utiliza nodos o neuronas que están interconectados a través de una estructura de capas. Crean un sistema adaptable, el cual permite que las computadoras puedan utilizar para aprender de sus errores y mejorar continuamente. Pueden ser utilizadas tanto para clasificación como regresión.(aws, s.f) El modelo fue creado a partir de la librería *tensorflow* con la API *keras* y también se utilizó la librería *scikit-learn* para hacer una separación de los datos. El modelo se creó siguiendo los siguientes parámetros:

```
model = Sequential([
    Dense(units=192, input_shape=[15,]),
    Dropout(0.3),
```

Dense(units=160, activation='sigmoid'),

Dropout(0.2),

Dense(units=64, activation='sigmoid'),

Dropout(0.2),

Dense(units=5, activation='softmax')

)

Hiperparámetros

Primera capa: *Dense(units=192, input_shape=[15,])*

Segunda capa: *Dense(units=160, activation='sigmoid')*

Tercera capa: *Dense(units=64, activation='sigmoid')*

Cuarta capa: *Dense(units=5, activation='softmax')*

A la hora de entrenar al modelo:

epochs = 50

optimizer = Adam(0.001)

loss='categorical_crossentropy'

metrics=['accuracy']

Evaluación

exactitud = 83.35 %

sensibilidad = 62%

F1 = 65 %

Máquina de vectores de soporte (SVM)

Uno de los modelos de aprendizaje supervisado utilizados fue una máquina de vectores de soporte (SVM) para hacer clasificaciones. En este método se intenta crear un hiperplano que intente separar los datos en sus clasificación de la mejor manera posible, donde uno de sus hiperparámetros es el kernel usado para transformar los datos y llevarlos a un espacio donde el hiperplano pueda separar linealmente los datos. Asimismo el parámetro de regularización indica qué tan estricto se intenta lograr la separación. En este caso se definió la variable objetivo de manera binaria; es decir, se estableció que todo aquel registro con la calificación máxima de satisfacción general (5) de la experiencia es la clase positiva a predecir demostrando que estuvo satisfecho, de lo contrario se le asigna la etiqueta negativa. Se creó un nuevo campo donde se aplica esta definición de satisfacción binaria usando como variables independientes las demás respuestas a las preguntas de satisfacción de los demás aspectos de la experiencia junto con el puntaje de sentimiento en los comentarios realizados por los encuestados. Debido a que aún se encuentra un cierto desbalance en las etiquetas de la variable objetivo donde se encuentra una mayor cantidad de casos positivos por una razón aproximada de 2:1, se le asigna en pruebas posteriores un mayor peso a la clase negativa para compensar este desbalance. Al realizar de esta manera el modelo, con un parámetro de regularización de $C = 1$ y con un kernel *RBF* o función de base radial se obtienen buenos resultados. El modelo se creó con el siguiente objeto de *scikit-learn*, haciendo uso de *train_test_split* con una semilla para tener resultados reproducibles.

SVC(class_weight={0:1},C = 1, kernel = "rbf")

Hiperparámetros

$C = 1$

kernel: RBF

Peso clase negativa: 1

Evaluación

exactitud = 89.3%

sensibilidad = 93.3%

F1 = 92.5%

Al evaluar el modelo, se tomó en cuenta el buen puntaje de sensibilidad indicando que para todos los casos verdaderamente positivos, se captó un 93.6% de los casos de manera correcta. No solo eso, sino en las demás métricas se obtuvieron resultados congruentes. Sin embargo, no se tomó en cuenta el desbalance de las clases, por lo que se asigna un mayor peso a la clase negativa siendo de 1.5, concordando con la proporción de ambas clases.

Hiperparámetros

$C = 1$

kernel: RBF

Peso clase negativa: 1.5

Evaluación

exactitud = 89.9%

sensibilidad = 91.8%

F1 = 92.3%

Como se puede observar se obtienen resultados muy similares aun intentando compensar por el desbalance de clases. Al mover los parámetros de manera diferente no se logró obtener resultados significativamente diferentes, únicamente se encontraron peores resultados como con la siguiente prueba.

Hiperparámetros

$C = 0.5$

kernel: Sigmoid

Peso clase negativa: 1.5

Evaluación

exactitud = 65.9%

sensibilidad = 100%

F1 = 79.5%

Además de usar un modelo SVM para predecir la satisfacción, se utilizó para intentar predecir el tipo de comentario que haría un estudiante con base en sus respuestas en las preguntas de satisfacción. Esto se hizo con el propósito de tener un sistema que pueda identificar las organizaciones o las experiencias que necesitan mejoras.

Hiperparámetros

Peso de clases = 'balanced'

Evaluación

exactitud = 68.36%

sensibilidad = 65 %

F1 = 61%

Árboles de decisión

Otro de los principales modelos utilizados fue el de árboles de decisión. A diferencia del modelo anterior, este puede llegar a capturar comportamientos de datos más complejos

debido a la gran cantidad de hiperparámetros que lo vuelven altamente versátil. En este método se intentan crear construcciones lógicas para poder separar (*samples_split*) los datos lo más posible en posibles caminos hasta llegar a una de sus *hojas* (*samples_leaf*) definiendo así su clase. Para este modelo, se aprovecha su habilidad para clasificación multiclas y obtener las probabilidades para cada registro de caer en las clasificaciones. En este caso se siguió el mismo proceso de entrenamiento de datos para predecir los distintos valores que se pueden tomar en la respuesta de satisfacción general. Se utilizaron los siguientes parámetros en algunos casos tomando en cuenta el desbalance de las clases siendo una mayoría de valores de 5. La representación del árbol se puede ver en la Figura 15; sin embargo, es imposible apreciar los criterios que usa para hacer las divisiones ya que crea demasiadas ramas.

```
DecisionTreeClassifier(criterion='entropy',  
min_samples_split=20,  
min_samples_leaf=5,  
max_depth = 15)
```

Evaluación

exactitud = 81.16%

sensibilidad = 60.0%

F1 = 62.2%

(Para las evaluaciones se toma el puntaje promedio por cada clase)

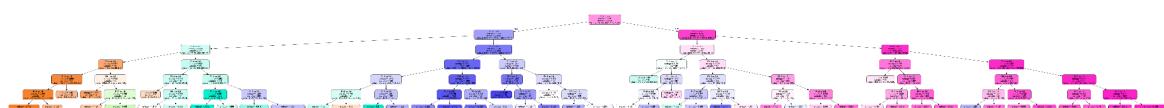


Figura 15: Árbol de decisión

Aquí se puede observar un puntaje medio o bajo para las métricas de sensibilidad y *F1*, aunque se obtiene un mejor valor en el área de exactitud. En el siguiente modelo se

intenta ajustar de mejor manera el modelo añadiendo pesos adicionales a las repuestas diferentes de 5 y aumentando el mínimo de hojas y máximo de profundidad del árbol.

Hiperparámetros

```
criterion = 'entropy'  
min_samples_split = 20  
min_samples_leaf = 5  
max_depth = 15  
class_weight = {1: 3.5, 2: 2.5, 3: 1.5, 4: 1.25}
```

Evaluación

```
exactitud = 81.96%  
sensibilidad = 58.8%  
F1 = 57.2%
```

Se observa nuevamente un comportamiento similar, por lo que encontrar la combinación óptima de hiperparámetros es necesario. Estos son los principales modelos usados con los mejores desempeños y ejemplos de las variaciones que se realizaron.

Naïve Bayes

Por otro lado, se hace también uso del modelo Naïve Bayes para predecir la satisfacción general del usuario. En este modelo se hace uso de distribuciones de los datos para encontrar la probabilidad de una de las clases dado un vector de datos. Esto provoca que sea muy sensible a la información usada para su entrenamiento. Para crear el objeto del modelo se usaron distintos parámetros.

GaussianNB(priors=None, var_smoothing=1e-09)

Hiperparámetros

priors = None

var_smoothing = 1e-09

Evaluación

exactitud entrenamiento = 76%

exactitud evaluación = 75%

sensibilidad = 49.8%

sensibilidad clase 5 = 87%

F1 = 75%

Se obtienen una exactitud con los datos de entrenamiento de 76.0% y con los datos de evaluación un 75.0%. Esto quiere decir que en todos los casos para los datos de evaluación se obtuvieron 1572 de 2099 posibles. Asimismo, la sensibilidad del modelo resultó en un promedio de 49.8% en promedio para todas las clases, donde la sensibilidad para la clase 5 resultó en 87.0%. Por otro lado, la precisión promedio resultó en 77.0% mientras que el puntaje *F1* se obtuvo un 75% ponderando según los valores de satisfacción de las clases. En general se obtuvieron resultados moderados al hacer uso de este modelo a pesar de hacer cambios en los hiperparámetros.

Posteriormente se decidió hacer la prueba de combinar el poder de reducción de componentes PCA, que será descrito más adelante en este documento al ser un modelo de aprendizaje no supervisado, para realizar la misma operación de clasificación usando este mismo método usando los siguientes datos para cada uno de los algoritmos usados en este proceso de mejora:

Hiperparámetros(PCA)

PCA(n_components=8)

Hiperparámetros(Naïve Bayes)

priors = None

var_smoothing = 1e-09

Evaluación

varianza acumulada = 72%

exactitud entrenamiento = 80%

exactitud evaluación = 79%

sensibilidad = 53%

sensibilidad clase 5 = 91 %

F1 = 79%

Al haber obtenido resultados relativamente mejores a los obtenidos por los parámetros anteriores con una varianza resumida de la matriz original de 72%, la importancia de este número se refleja en la figura 16, donde se puede ver la magnitud de la correlación de las ocho nuevos componentes con los originales, apunta a que este modelo quizás sea el más adecuado para presentar a la Organización Socio Formadora como el mejor para predecir la satisfacción del alumnado al final del semestre de acuerdo con las variables predictoras seleccionadas, la representación gráfica de la eficacia de este modelo para predecir se puede visualizar en las figuras 17 y 18, ambas se logran usando un PCA con reducción a solamente dos componentes. Evidentemente este hiperparámetro de PCA no sería lo suficientemente alto para acumular un porcentaje significativo de la varianza total de la información, sin embargo, es una forma eficaz de mostrar gráficamente cómo es que el modelo decide predecir según su información de entrenamiento.

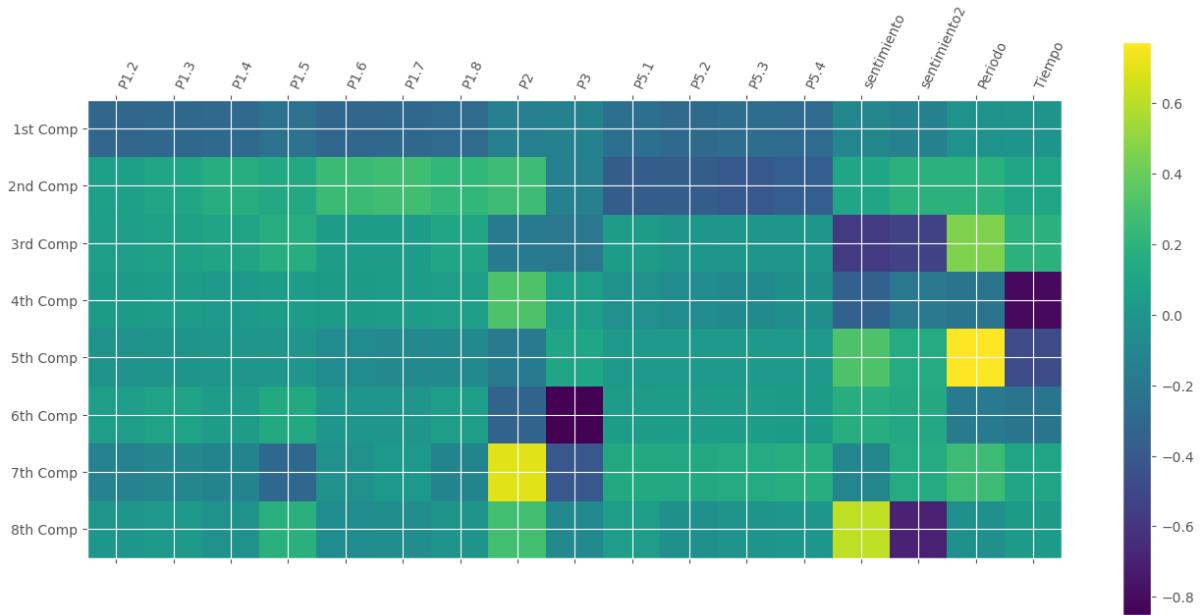


Figura 16: Mapa de calor de correlación entre nuevos componentes de PCA y originales.

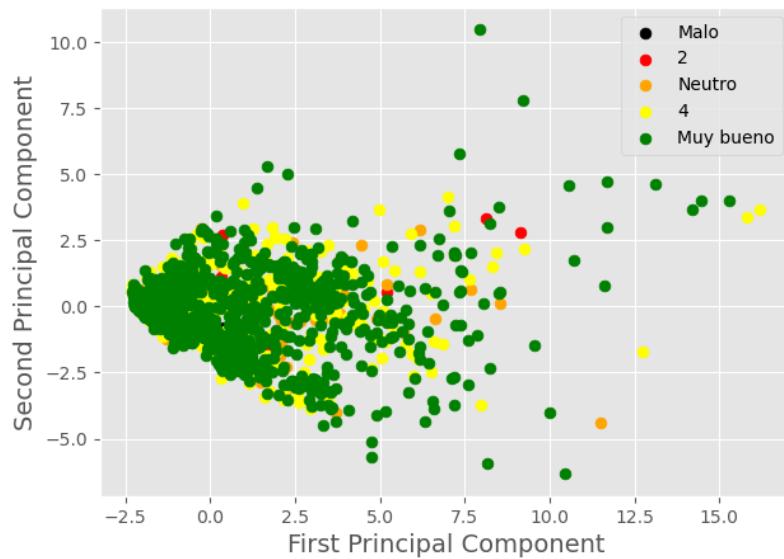


Figura 17: Representación gráfica de la distribución de las categorías de P1.1 en los datos de validación de acuerdo con un PCA de dos componentes.

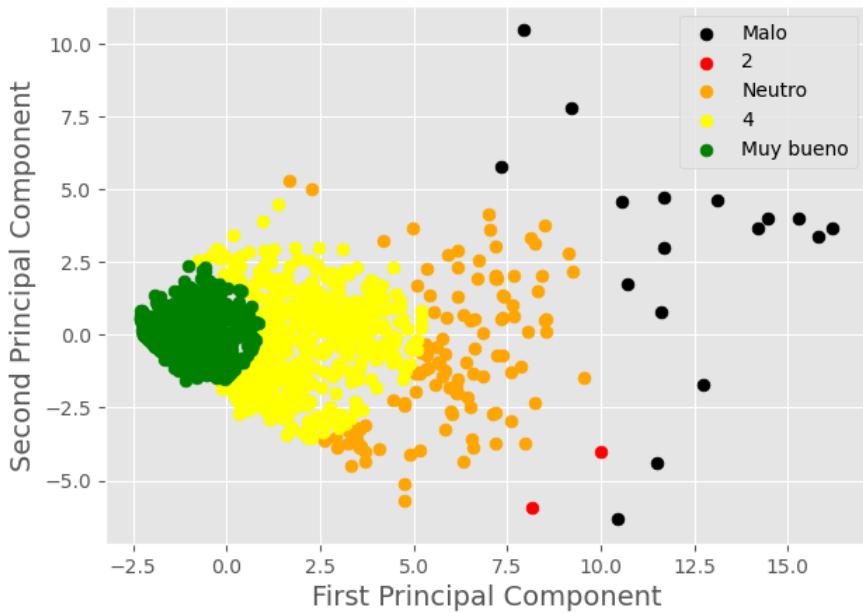


Figura 18: Representación gráfica de la distribución de las categorías de P1.I en los datos predecidos por el modelo de acuerdo con un PCA de dos componentes.

Regresión lineal múltiple

Finalmente, se hace uso de la regresión lineal con el principal objetivo de poder obtener de manera ordenada las mejores y peores organizaciones u OSF para las experiencias de servicio social. En este modelo se intenta ajustar una recta lineal intentando minimizar el cuadrado de los errores entre las predicciones realizadas y los datos reales otorgándole pesos a las variables independientes utilizadas. Para este modelo se usa un regresor de mínimos cuadrados ordinarios donde no se aplican medidas de regularización. En este caso se tomó la variable a predecir como un valor continuo de entre 1 y 5 indicando la satisfacción del encuestado. Al realizar esto se obtuvieron buenos resultados observando un coeficiente de determinación de 72.6%, indicando que una gran parte de la variación en la satisfacción del encuestado es explicado por este modelo. Por otro lado se obtuvo una raíz de error medio cuadrado de 0.384, indicando que el modelo en promedio se acerca grandemente a las satisfacciones reales. Para usar este modelo para el *ranking* de las OSF, se obtuvieron los coeficientes de cada una de las variables independientes utilizadas, incluyendo el sentimiento

de los comentarios del encuestado y se calcularon los pesos de todas las variables. Con estos pesos se formuló una ecuación lineal de cada uno de los aspectos de satisfacción del encuestado hacia una OSF. Esta ecuación describe la satisfacción del encuestado i hacia la OSF j , donde C se refiere a los coeficientes dados por la regresión lineal y las variables independientes se refieren al valor de satisfacción en un campo del encuestado en cuanto a la OSF:

$$(P1.1) SATISFACCIÓN_{ij} = (P1.2_{ij}) \cdot C_{P1.2} + (P1.3_{ij}) \cdot C_{P1.3} + \dots$$

Con esto se estableció un valor de satisfacción de cada uno de los encuestados hacia las organizaciones. Para cada una de las OSF se calcula su media o valor esperado según todas los puntajes de satisfacción dirigidos a ellas. Ordenando de mayor a menor se encuentran las OSF con mayor puntuación de satisfacción logrando rankear las organizaciones. Las mejores y peores organizaciones se presentarán en el prototipo. Los pesos usados en este modelo se pueden observar en la Figura 19. Con estos podemos concluir qué factores son los que más afectan la satisfacción del usuario en un proyecto de servicio social.

| Peso | |
|-------------|---------------|
| P2 | 8.732259e-01 |
| P5.1 | 8.455487e-01 |
| P5.4 | 6.243197e-01 |
| P5.3 | 4.752365e-01 |
| P5.2 | 4.238399e-01 |
| P3 | 2.556342e-01 |
| P1.5 | 6.081360e-02 |
| const | 1.199266e-04 |
| P1.8 | 1.314730e-08 |
| P1.7 | 7.287333e-15 |
| P1.4 | 1.246765e-30 |
| P1.3 | 4.575751e-57 |
| P1.6 | 2.707175e-97 |
| P1.2 | 1.527719e-216 |

Figura 19: Pesos del modelo de regresión lineal

Con base en los pesos determinamos que los factores que más afectan ordenados de más importante a menos importante son: retroalimentación, valores, soluciones, diversidad, responsabilidad, interés y actividades.

Modelos de aprendizaje no supervisado

K-Means

En cuanto a modelos no supervisados, uno de los modelos utilizados fue el modelo K-Means para la categorización de los comentarios de los encuestados. La manera en que funciona el algoritmo K-Means es que se quieren añadir k agrupaciones para los datos. Para cada una de las agrupaciones se les asigna un centroide, el cual se estará moviendo intentando centrarse en cada uno de los clusters. Una vez que los centroides dejan de moverse, el algoritmo se detiene. (Al-Masri, 2022)

La clave de este algoritmo es encontrar la cantidad óptima de clusters. Para determinar eso se usa la curva del codo. La curva del codo es una herramienta utilizada en el

algoritmo de agrupamiento K-means para determinar el número óptimo de clusters en un conjunto de datos. La curva del codo representa la relación entre la cantidad de grupos y la variación total dentro de los grupos. Se calcula para diferentes valores de K , donde K representa el número de grupos. La curva del codo muestra cómo disminuye la variación total a medida que aumenta K . El punto en la curva donde se observa un cambio significativo en la pendiente se considera el "codo". Este punto sugiere el número óptimo de grupos, ya que representa un equilibrio entre una buena división de los datos y evitar una excesiva fragmentación. (Jarroba, 2017)

Para la primera agrupación usando *K Means* solo se consideraron las variables numéricas; es decir, las preguntas de satisfacción y el sentimiento. Como se ve en la Figura 20, el número de clusters óptimo para esta agrupación sería tres.

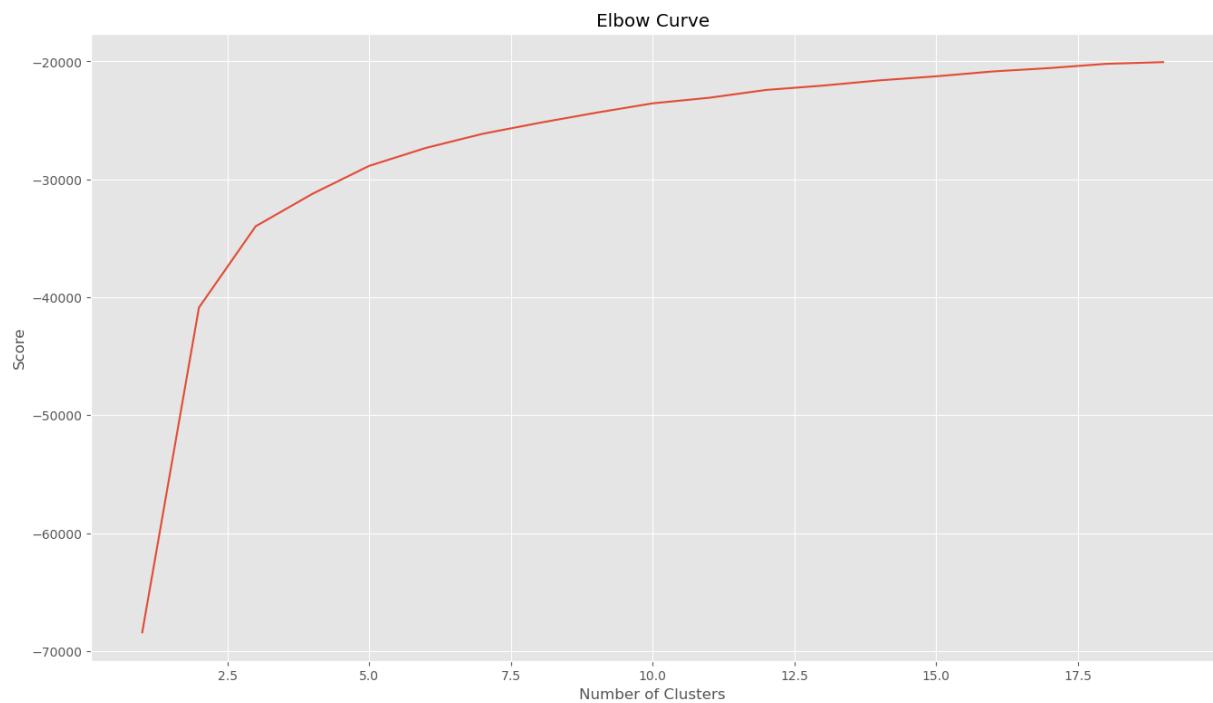


Figura 20: Curva del codo para elegir el número de clusters óptimo

Con base en la curva del codo, se creó la primera instancia del objeto del modelo especificando los hiperparámetros correspondientes.

`KMeans(n_clusters=3, max_iter=10000, tol=0.0001)`

Hiperparámetros

n_clusters = 3

max_iter = 10000

tol = 0.0001

Este modelo obtuvo una puntuación de silueta de 0.43, por ende podemos concluir que los puntos dentro de cada grupo están relativamente cerca entre sí y algo separados de los puntos de otros grupos. Sin embargo, podría existir cierto solapamiento o falta de distinción entre algunos puntos. Como se puede observar en la Figura 21, este modelo agrupó los registros en tres grupos diferentes. El representante (centroide) del grupo uno fue San Pedro Parques, el representante del grupo dos fue Dirección de Servicio Social, Fundación Enrique Yturria García, A.B.P. Al revisar las respuestas a las preguntas de estas organizaciones como se ve en la Figura 22 podemos observar que el grupo uno representa a las organizaciones que tienen 5 de calificación, el segundo grupo representa a las organizaciones que tienen 4 de calificación y el grupo tres representa a los que tienen calificaciones iguales o menores que 3. Con esta agrupación podemos concluir que hay alrededor de 698 organizaciones o servicios que requieren de mejoras.

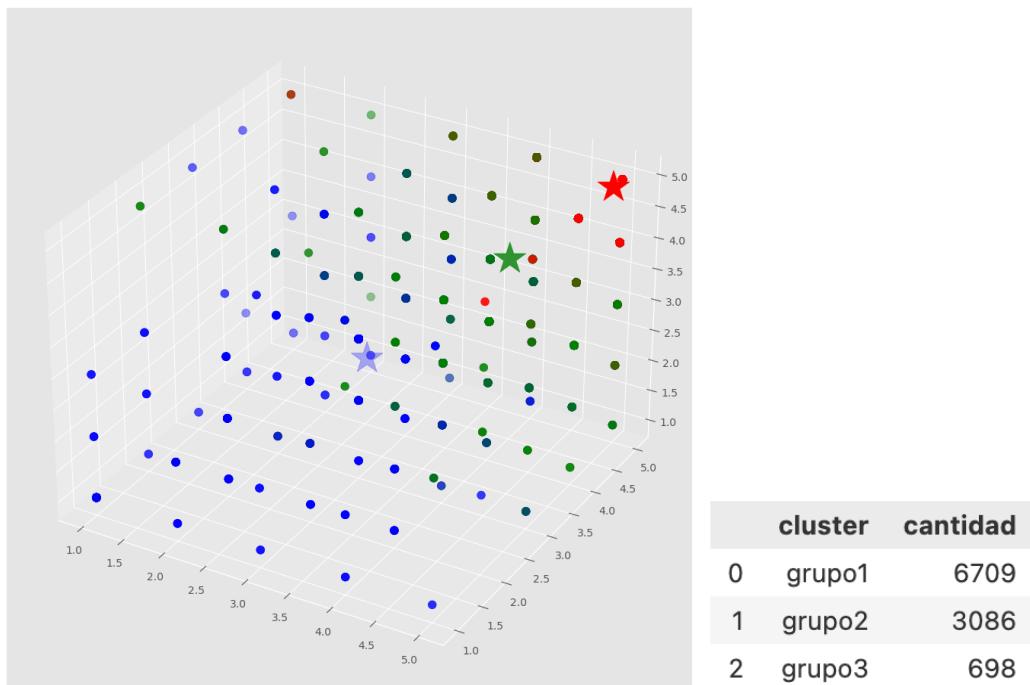


Figura 21: Agrupaciones hechas por el modelo Kmeans

| | P1.1 | P1.2 | P1.3 | P1.4 | P1.5 | P1.6 | P1.7 | P1.8 | P2 | P3 | P5.1 | P5.2 | P5.3 | P5.4 | OSF | sentimiento | cluster |
|---|------|------|------|------|------|------|------|------|----|----|------|------|------|------|--------|-------------|---------|
| 0 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 1 | 1 | 1 | 5 | 5 | 0.4779 | 0 | |
| 1 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 1 | 1 | 4 | 4 | 4 | 0.4019 | 1 | |
| 2 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 3 | 3 | 1 | 1 | 4 | 4 | 4 | 0.0000 | 2 | |

Figura 22: Datos de los representantes (centroídes) de cada grupo

Además, se intentaron generar más clusters para ver si esto mejoraba la categorización. Uno de los casos utilizados fue donde se generaron 6 clusters para categorizarlos siguiendo la siguiente creación del modelo.

KMeans(n_clusters=6, init='k-means++', n_init=10, max_iter=300, tol=0.0001)

Hiperparámetros

n_clusters = 6

init = 'kmeans ++ '

n_init = 10

max_iter = 300

tol = 0.0001

Con ese modelo se obtuvo un puntaje *silhouette* de 0.38. Tomando en cuenta que el dominio de esta métrica es de -1 a 1, este puntaje quiere decir un leve grado de desempeño al momento de crear las categorías, indicando que puede haber errores. Al probar diferentes hiperparámetros la puntuación quedaba igual o empeoraba.

Clustering jerárquico aglomerativo

Otro de los modelos no supervisados que se utilizó fue el clustering jerárquico aglomerativo con el objetivo de poder categorizar y encontrar patrones en los comentarios de los estudiantes. La manera en que funciona este algoritmo es que divide a la población en clusters de forma que los datos que pertenecen al mismo cluster son más similares entre sí y tiene un enfoque ascendente, lo que quiere decir que al inicio cada uno de los puntos pertenece a un cluster diferente; sin embargo eventualmente los pares de clusters se van fusionando a medida que van ascendiendo en la jerarquía . (Kumar, 2021)

En uno de los casos que se utilizó se crearon 2 clusters, siguiendo la siguiente creación del modelo:

$$\text{AgglomerativeClustering}(n_clusters=2)$$

Con ese modelo se obtuvo un *silhouette* de 0.54, interpretando este resultado se puede decir que los clusters tienen una separación moderada, por lo general un *silhouette_score* arriba de 0.5 se considera razonablemente bueno.

Hiperparámetros

$$n_clusters = 2$$

En algunas de las pruebas que se hicieron se fue modificando el parámetro *n_clusters* pero cada vez que se ponía otro valor diferente de 2, arrojaba valores más bajos en el *silhouette*.

PCA

Otro modelo que se usó fue el *Principal Component Analysis* (PCA), el cual es un método que permite reducir la dimensionalidad de grandes conjuntos de datos, transformando un gran conjunto de variables en uno más pequeño que siga manteniendo la mayor parte de la varianza de la información del conjunto grande. El reducir el número de las variables de un conjunto de datos deteriora la precisión, sin embargo es un cambio por simplicidad. La manera en cómo es que funciona, es que estandariza los datos, después se saca la matriz de covarianzas y después se sacan los eigenvectores y eigenvalores de la matriz de covarianza y en base a eso se construyen los componentes principales (Jaadi, 2021).

Para hacer el modelo de PCA, se tuvo que primero estandarizar los datos, por lo que se utilizó la función *StandardScaler()* de la librería *scikit-learn* en las variables independientes. Posteriormente se hizo el modelo con la siguiente configuración, de acuerdo con lo obtenido por la figura 20, que representa el porcentaje de varianza explicada cumulativa que reserva la nueva matriz reducida al usar x cantidad de componentes en el algoritmo no supervisado:

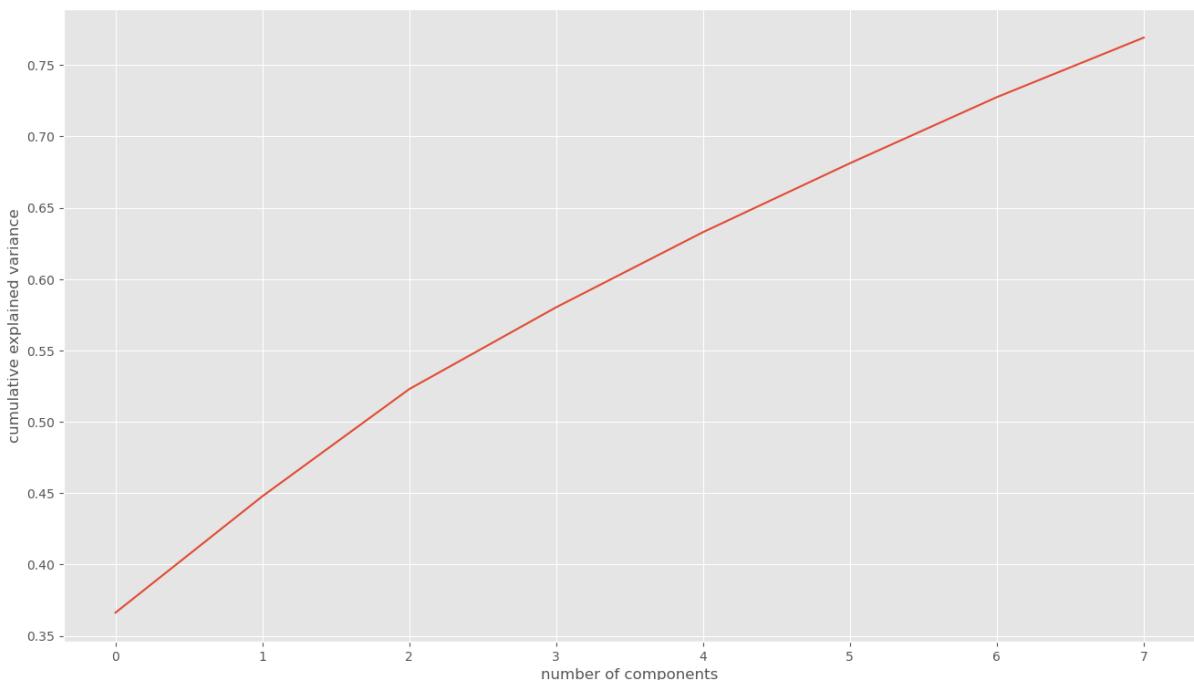


Figura 23: Gráfica de la varianza acumulada usando x cantidad de componentes en el PCA

Hiperparámetros

PCA(*n_components*=5)

El resultado de simplificar la información original en solamente ocho componentes resume 77.7% de la varianza total del modelo, lo cual ayuda a un posterior modelo de aprendizaje supervisado para un procedimiento más rápido que llegue a los mismos resultados, la correlación de los nuevos componentes con los originales se puede observar en la figura 24.

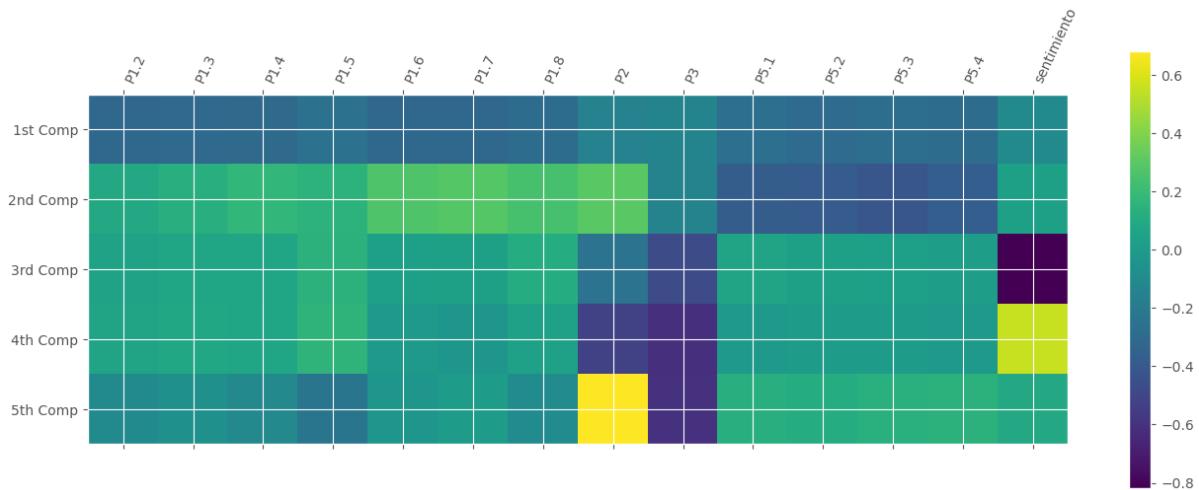


Figura 24: Correlación de componentes nuevos con los originales de la matriz de datos

Discusión de validez

El principal elemento a considerar cuando se trata de la validez de los modelos es el manejo del desbalance de clases o valores de la variable a predecir, teniendo un sesgo negativo donde la mayorías de las respuestas de satisfacción se encuentran en los valores altos. Es por esto que para los modelos de clasificación de SVM y en otros se añadió un mayor peso al momento del entrenamiento. Es también por esto que se transformó en ocasiones la variable a predecir a una variable binaria donde las calificaciones del 1-4 se

agruparon mientras que los valores de 5 se mantuvieron en un mismo grupo. Por otro lado, en los modelos de árboles de decisión al igual que en los demás se les asignó pesos mayores a las clases menos representadas en los datos. Asimismo para la validación de los modelos se tomaron en cuenta varias métricas como la sensibilidad y precisión de las clases para determinar su desempeño en las clases menos representadas.

Para el caso de los modelos no supervisados, se realizó el preprocesamiento adecuado para los comentarios, siendo la tokenización, eliminación de palabras vacías, y encontrando la relevancia de cada una de las palabras utilizando el vectorizador *TF-IDF*, apoyando la validación de los modelos. Teniendo en cuenta que algunos encuestados no contestan las encuestas con sinceridad o que hay sesgo en los datos, las puntuaciones de silueta de los modelos no supervisados son razonables y relativamente buenos considerando que las puntuaciones arriba de cero reflejan un agrupamiento distinguible entre los datos. Evidentemente hay espacio para mejorar este agrupamiento usando modelos más complejos o más procesamiento de comentarios.

Ajustes

Para mejorar el rendimiento de los modelos y encontrar la combinación óptima en modelos no supervisados se hace uso de técnicas como el método del codo y el cambio de hiperparámetros. En el modelo de K-Means se intenta maximizar la compactación de los clústeres y la separación entre ellos graficando el desempeño de varios modelos generados con diferente cantidad de clusters (Jarroba, 2017). Mientras se establecen más clusters se obtienen mejores resultados debido a que tienden a agruparse individualmente en este punto, es por eso que se busca el punto donde la mejoría de resultados ya no es significativa teniendo el mínimo número de clusters posibles usando una misma métrica siendo *silhouette*.

Viendo el resultado, se observa que el número óptimo de clusters a generar es de tres. Por lo tanto, este es el principal parámetro usado para este modelo para agrupar las OSF en diferentes categorías según sus niveles de satisfacción.

Para los ajustes de los hiperparámetros, se utilizaron técnicas y herramientas como la validación cruzada, donde se dividen los datos en varios conjuntos más pequeños y realizan validaciones entre ellos iterando por diferentes particiones cada iteración. Esta técnica se aplica en el modelo de árboles de decisión y se realiza para entender de mejor manera el desempeño del modelo como se ve en la Figura 25 en donde se muestran los puntajes *F1* al usar cada pliegue, en total 5.

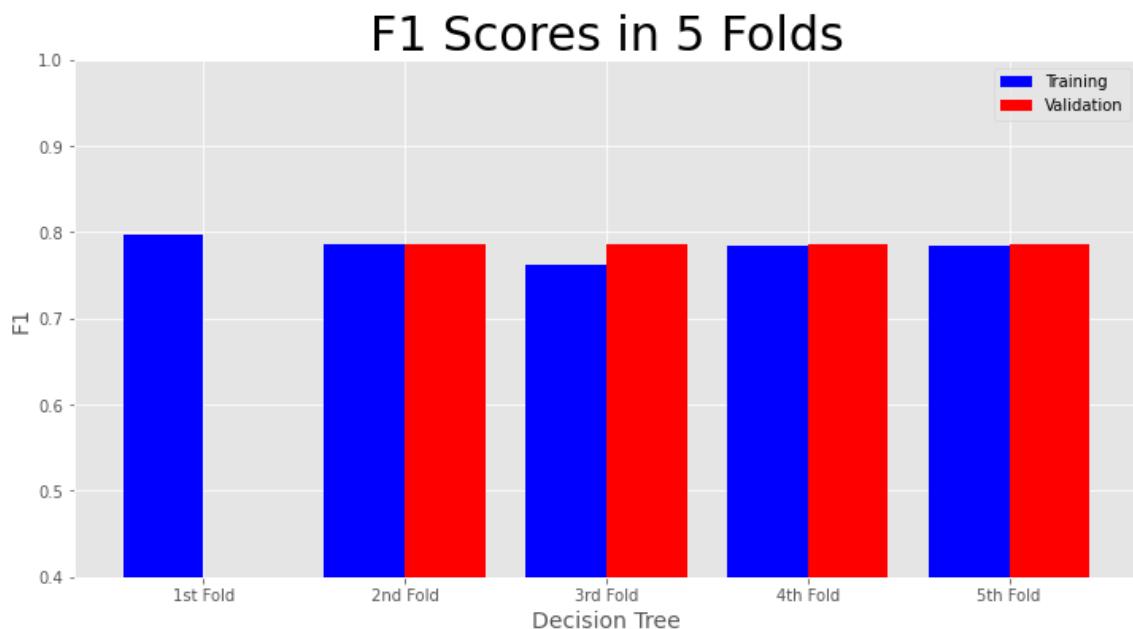


Figura 25: Validación cruzada de modelo de árboles de decisión

Como se puede observar, el modelo general obtiene resultados buenos rondando un 80% en esta métrica, la cual consiste de la media armónica entre la sensibilidad y exactitud. Por lo tanto, este modelo sí parece ser suficientemente eficiente para ser candidato a ser el modelo finalmente seleccionado.

Los otros hiperparámetros que se modificaron fueron la cantidad de iteraciones, el error aceptado, el tipo de kernel a utilizar y la profundidad máxima. Todo esto con la

intención de mejorar el rendimiento de los modelos y el resultado de las predicciones o agrupaciones.

Selección del modelo

Los modelos elegidos para la realización final del proyecto consisten de la regresión lineal como método supervisado para el *ranking* de las OSF, el árbol de decisión para predecir la satisfacción del usuario y el modelo K-Means como el modelo no supervisado. Estos modelos se escogieron por los puntajes altos que obtuvieron en comparación con los demás y por lo útiles que pueden ser. Estos modelos tienen los siguientes hiperparámetros.

Regresión Lineal Múltiple

Incluir intersección: *Verdadero*

Coeficientes positivos: *Falso*

Paralelizar Proceso: *Falso*

Copiar Variables Indep: *Verdadero*

Árboles de decisión

Criterio: *entropy*

Mínimo split: *20*

Mínimo hojas: *5*

Profundidad máxima: *15*

KMeans

Número de Clusters: *3*

Inicio de Clusters: *k-means++*

Iteraciones algoritmo: *10*

Máx. Iteraciones: *300*

tolerancia: *0.0001*

semilla: 100

algoritmo: 'lloyd'

Para el modelo no supervisado de K-Means se elige el número óptimo de clusters encontrado al aplicar el método del codo siendo tres clusters. Se aplica una configuración *k-means++* donde la posición inicial de los centroides no se realiza de manera aleatoria sino que se sitúan primero en lugares cercanos a los registros según su distribución para intentar aproximarse lo más posible a las posiciones finales de los centroides de los clusters desde un principio. El algoritmo se repite 10 veces con diferentes posiciones iniciales de los centroides. En cada iteración, se actualiza un máximo de 300 veces la posición de los centros de los clusters. Si al actualizar los clusters no hay una diferencia de posición de centroides mayor a 0.0001, se para la iteración. Se usa la semilla 100 para tener resultados reproducibles. Finalmente se usa el algoritmo *lloyd* el cual es el más común siguiendo el mismo algoritmo anteriormente explicado. Para la regresión lineal no se sigue un método de regularización y se incluyen todos los coeficientes reales, sin convertirlos a positivos, se incluye la intersección o sesgo del modelo y no se alteran los datos de entrenamiento. Es importante destacar que se eligió el modelo K-Means a pesar de que no haya sido el de mayor puntaje de silueta porque pudimos identificar qué tenían en común estas agrupaciones mientras que en los otros modelos agrupados no se logró.

Evaluación

Resultados

Con base en el modelo de regresión, usando los coeficientes obtenidos por columna, se pudo crear una columna de ranking en la cual se da una calificación a cada registro basado en las respuestas dadas. Posteriormente se hace un groupby, donde conseguimos el ranking general de las OSF usando la media de su puntaje.

El orden de importancia de las variables numéricas para la satisfacción de los encuestados es el siguiente:

1. Ofrecer retroalimentación
2. Ser sensible al dolor
3. Promover soluciones
4. Respetar la diversidad
5. Ser interesante
6. Herramientas de Canvas
7. Atención del departamento de servicio social
8. Seguimiento y liderazgo de la organización
9. Interacción con los líderes de proyecto
10. Aportación a la organización
11. Colaboración
12. Relacionado con las ODS

El modelo hecho con árboles de decisión puede ser utilizado por la OSF para predecir la satisfacción general de un alumno dadas las variables independientes. Podría usarse para experimentar con los valores y anticipar el resultado de una organización.

El modelo K-Means arrojó tres grupos de organizaciones: excelentes, geniales y malas. Para ayudar a mejorar los proyectos de servicio social, se decidió hacer un análisis de

los comentarios del tercer grupo del modelo de aprendizaje no supervisado. Los resultados de este análisis se muestran en la Figura 26. Como se puede observar los resultados son casi los mismos que se encontraron anteriormente analizando los comentarios relacionados a las áreas de mejora. Es decir, las áreas de mejora más importantes son de organización, actividades, actitud e impacto. Al tener esta agrupación el Departamento de Servicio Social podrá identificar los proyectos que no son muy buenos e implementar las mejoras que se requieren.



Figura 26: Nube de palabras de los comentarios del tercer grupo de K-Means

Mediante el uso de la regresión lineal múltiple pudimos rankear las organizaciones socio formadoras como se muestra en la Figura 27. Esto ayudará a premiar a las mejores organizaciones y a mejorar las peores.

| Ranking | OSF |
|---|-----------|
| META, Impulso Deportivo y Académico, S.C. | 98.429529 |
| Escalando Fronteras, A.C. | 98.429529 |
| Fundación Chávez Ramos, A.C | 98.429529 |
| Dirección de carreta Ing. Mecánica | 98.429529 |
| Secretaría de Medio Ambiente de Nuevo León | 98.429528 |
| Instituto de Bienestar Integral (IBI) | 98.429528 |
| Fundación FEMSA , A.C. | 98.429528 |
| Emprendedores Líderes Generando la Evolución por México A.C. (Elige México) | 98.429527 |
| GeoStats | 98.429527 |
| Conexión TEC | 98.429527 |

Figura 27: Ranking de las OSF

Los resultados obtenidos serán de gran utilidad y causarán un impacto significativo en la calidad del servicio social si el Departamento de Servicio Social decide implementar las mejoras sugeridas al final del documento. En primer lugar, al utilizar modelos de aprendizaje supervisado y no supervisado, hemos logrado extraer información valiosa y significativa de los datos que nos permite comprender mejor las experiencias y opiniones de los estudiantes durante su servicio social.

Nuestro análisis utilizando modelos de aprendizaje supervisado reveló la relación entre las variables independientes y las variables dependientes que representan la satisfacción del usuario. Esto nos ha permitido identificar los factores que influyen significativamente en las respuestas de los estudiantes y cómo se relacionan entre sí. Estos hallazgos permitirán a las organizaciones formadoras tomar medidas concretas respecto a las variables más significativas para mejorar la calidad de la experiencia del estudiante. Además, mediante el uso de técnicas de aprendizaje no supervisado, como el modelo K Means y PCA, hemos identificado patrones y agrupaciones naturales en las respuestas de los estudiantes. Al identificar grupos de organizaciones con características similares, podemos proporcionar

recomendaciones más específicas y personalizadas para cada grupo, lo que aumentará la eficacia de las medidas de mejora.

Las recomendaciones basadas en nuestros hallazgos permitirán mejorar las experiencias y aumentar la satisfacción de los estudiantes durante su servicio social. Esto, a su vez, tendrá un impacto positivo en el bienestar y el desarrollo académico de los estudiantes, así como en la reputación y la calidad general del programa de servicio social.

Proceso

Realizar este proyecto ha sido un proceso completo y minucioso que ha arrojado resultados significativos. En primer lugar, se estudió y analizó el negocio y los datos proporcionados por el Departamento de Servicio Social del Tecnológico de Monterrey. Después se realizó un preprocesamiento exhaustivo de los datos, especialmente de la columna de los comentarios para poder extraer información útil. Durante el estudio, se emplearon diversos modelos de aprendizaje supervisado y no supervisado para extraer información valiosa de los datos recopilados en la encuesta.

Uno de los principales desafíos que se presentaron durante el proyecto fue el proceso de selección y modificación de los modelos adecuados para lograr mejores puntuaciones. Dado que el objetivo era analizar y extraer información valiosa de los datos recopilados en la encuesta, era crucial elegir los modelos de aprendizaje supervisado y no supervisado más apropiados para abordar las características específicas del conjunto de datos. Esto implicaba considerar el tamaño y la estructura de los datos, así como la naturaleza de las variables involucradas. El equipo se enfrentó a la tarea de evaluar una variedad de modelos y técnicas, ajustando sus parámetros y realizando pruebas exhaustivas para encontrar la combinación óptima que produjera los mejores resultados. Se realizaron comparaciones detalladas de desempeño y se realizaron modificaciones en los modelos seleccionados para mejorar su

capacidad de predicción y su capacidad para revelar patrones y tendencias. Este proceso de refinamiento y optimización fue fundamental para garantizar que los modelos utilizados fueran capaces de brindar información valiosa y significativa a partir de los datos de la encuesta, lo que a su vez contribuyó al éxito general del proyecto. Otro reto importante fue identificar los factores clave que influyen en la satisfacción general de los estudiantes. Para ello, se utilizaron modelos de aprendizaje supervisado como la clasificación, los árboles de decisión y el SVM. Estos modelos permitieron analizar la relación entre las variables independientes y las variables dependientes seleccionadas para representar la satisfacción del usuario. A través de este análisis, se logró determinar qué factores ejercieron una influencia significativa en las respuestas de los estudiantes y cómo se relacionaron entre sí.

Además, se aplicaron técnicas de aprendizaje no supervisado, como el modelo K Means y el Análisis de Componentes Principales (PCA), para identificar patrones y agrupaciones naturales en las respuestas. Estas técnicas permitieron una comprensión más profunda de las opiniones y experiencias de los estudiantes durante su servicio social. Se identificaron grupos de organizaciones con características similares y se encontraron tendencias interesantes en las respuestas de los estudiantes.

Los resultados obtenidos revelaron información valiosa para la organización formadora. Se identificaron aspectos específicos que impactan significativamente en la satisfacción de los estudiantes durante su servicio social. Estos hallazgos permitieron identificar áreas de mejora y proporcionaron una base sólida para la formulación de recomendaciones. En resumen, el proyecto de análisis de las respuestas de la encuesta de conclusión del servicio social en el TEC fue un proceso riguroso que utilizó modelos de aprendizaje supervisado y no supervisado para extraer información valiosa de los datos. A través de este análisis, se identificaron los factores clave para la satisfacción de los estudiantes y se formularon recomendaciones específicas para mejorar la experiencia del

servicio social. Estas mejoras potenciales tienen como objetivo elevar la calidad del servicio social y aumentar la satisfacción de los estudiantes en el TEC de Monterrey.

Finalmente, para mejorar estos resultados se podría recopilar una muestra más amplia y representativa de respuestas de estudiantes en futuras encuestas y se podrían considerar técnicas avanzadas de procesamiento del lenguaje natural (NLP) para mejorar el análisis de las respuestas abiertas en la encuesta e intentar usar algoritmos más avanzados como redes neuronales.

Impacto social

El impacto social principal buscado en este proyecto es brindarle propuestas relevantes para el mejoramiento del ciclo de vida de un proyecto de servicio social o solidario. Al proponer mejoras y presentar la satisfacción de los involucrados en los proyectos, los servicios sociales tendrán un claro camino para mejorar las experiencias dadas e incluso llegarán a ampliar sus alcance al poder llevar a cabo los proyectos de manera más eficiente. Además de ayudar al Departamento de Servicio Social y a las organizaciones socio formadoras con las que trabaja, también ayudará a los alumnos del Instituto Tecnológico y de Estudios Superiores de Monterrey a disfrutar de los proyectos en los que participan y a aprovechar el aprendizaje técnico y emocional que estas experiencias les pueden brindar.

Impacto ODS

Este proyecto impactará los objetivos de desarrollo sostenible 4 (educación de calidad), 9 (industria, innovación e infraestructura) y 17 (alianzas para lograr objetivos). El impacto que nuestra solución brindará a estos ODS va de la mano con muchos de los objetivos de las Organizaciones Socio Formadoras (OSF) con las que trabaja el Departamento de Servicio Social.

- *Educación de calidad #4:* Aumentar la satisfacción del estudiante que realiza el servicio no solo mejorará su propia educación y experiencia, sino también mejorará la educación de los miles de niños y niñas que son parte de varios proyectos de servicio social en donde los alumnos del TEC les dan clases o asesorías; ya que, al mejorar la satisfacción del estudiante es más probable que más personas se unan al servicio y que realicen sus tareas en tiempo y forma y con mucho entusiasmo.
- *Industria, innovación e infraestructura #9:* Muchos proyectos de índole social buscan innovar y mejorar las estructuras existentes de muchos socios formadores creando aplicaciones, páginas web y otras herramientas útiles. Mejorar la satisfacción del estudiante hará que este participe activamente en el servicio social y ayudará a que se cumpla una de las metas de esta ODS, la cual es apoyar el desarrollo de tecnologías, la investigación y la innovación.
- *Alianzas para lograr los objetivos #17:* Asegurar la satisfacción del alumno permitirá que la alianza entre el TEC y todas las demás organizaciones socio formadores siguiera siendo fuerte. Esta alianza es de vital importancia; ya que, nos permite seguir trabajando con varias organizaciones que se enfocan en diversas ODS.

Despliegue

Prototipo

El prototipo consta de una página web, en donde se puede ver una tabla con el ranking que se hizo a partir de cada Organización Socio Formadora por período en la página de inicio, después el usuario puede navegar por otras páginas en donde se puede ver una tabla del ranking OSF general en base al modelo de regresión lineal, se puede ver una nube de comentarios con la que el usuario puede interactuar poniendo un resumen de los comentarios con el top 10 de las Organizaciones Socio Formadoras. En la página se le da la opción al usuario de buscar las mejores OSF según el periodo, donde puede determinar cuales con las mejores puntuadas consecutivamente y realizar un decisión con una mejor idea. Este prototipo se muestra en la Figura 28 y 29.



Figura 28: Página de inicio del prototipo

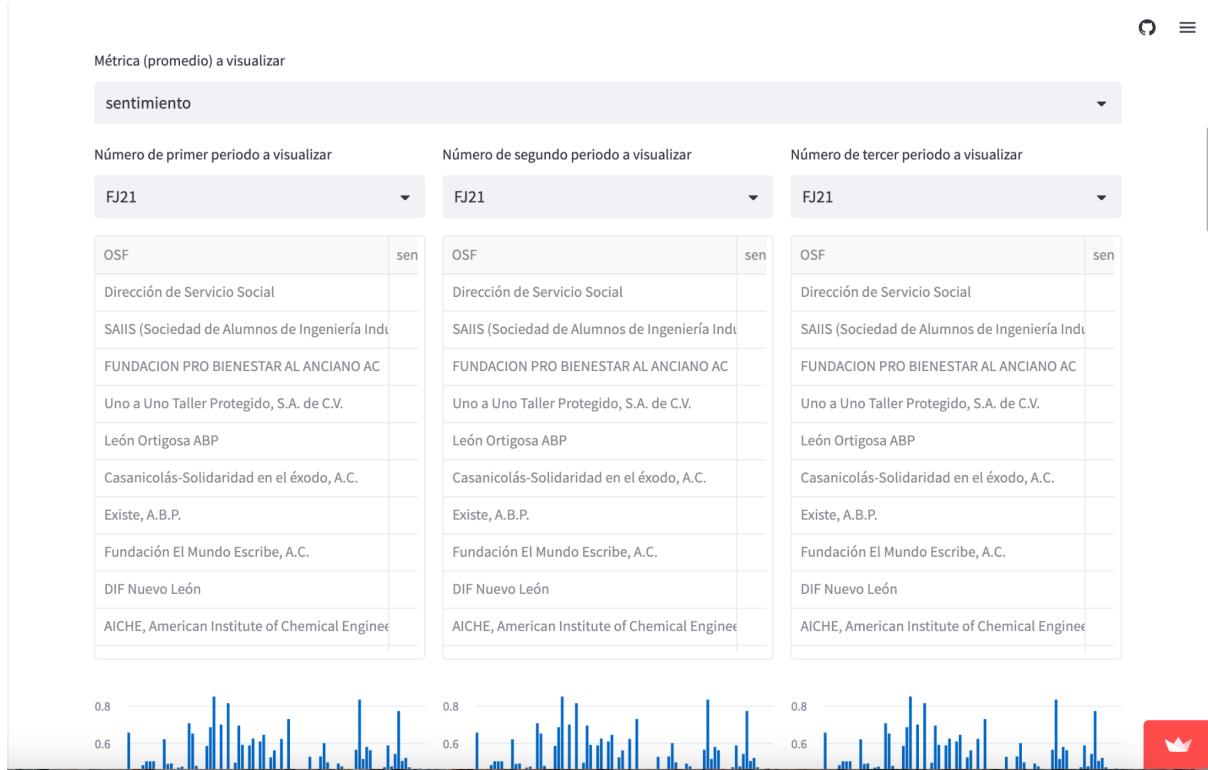


Figura 29: Segunda página del prototipo

En las Figuras 30 y 31 podemos observar las siguientes secciones del prototipo en donde el usuario puede ver la nube de palabras de las 10 mejores y peores organizaciones. En la última sección del prototipo se aprecian las características de los clusters formados por Kmeans.

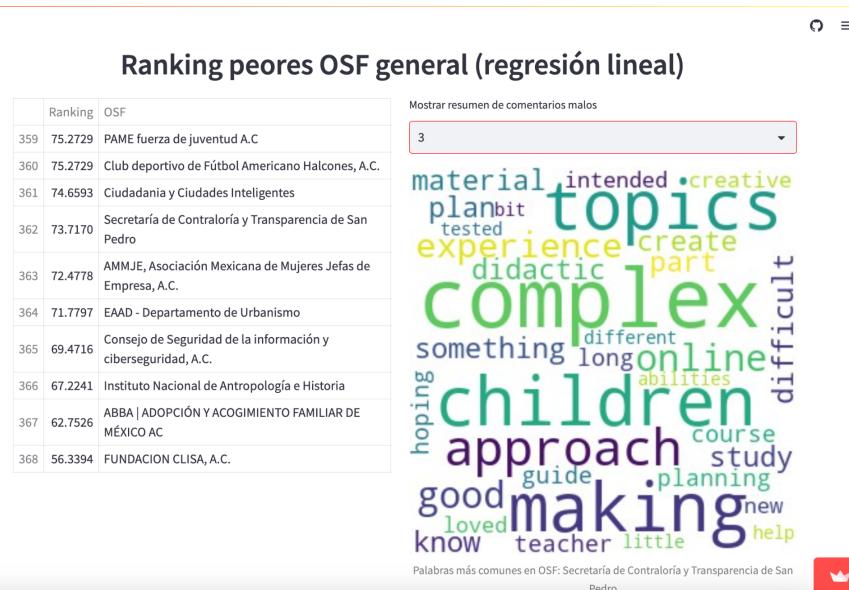


Figura 30: Ranking de las organizaciones y su nube de palabras

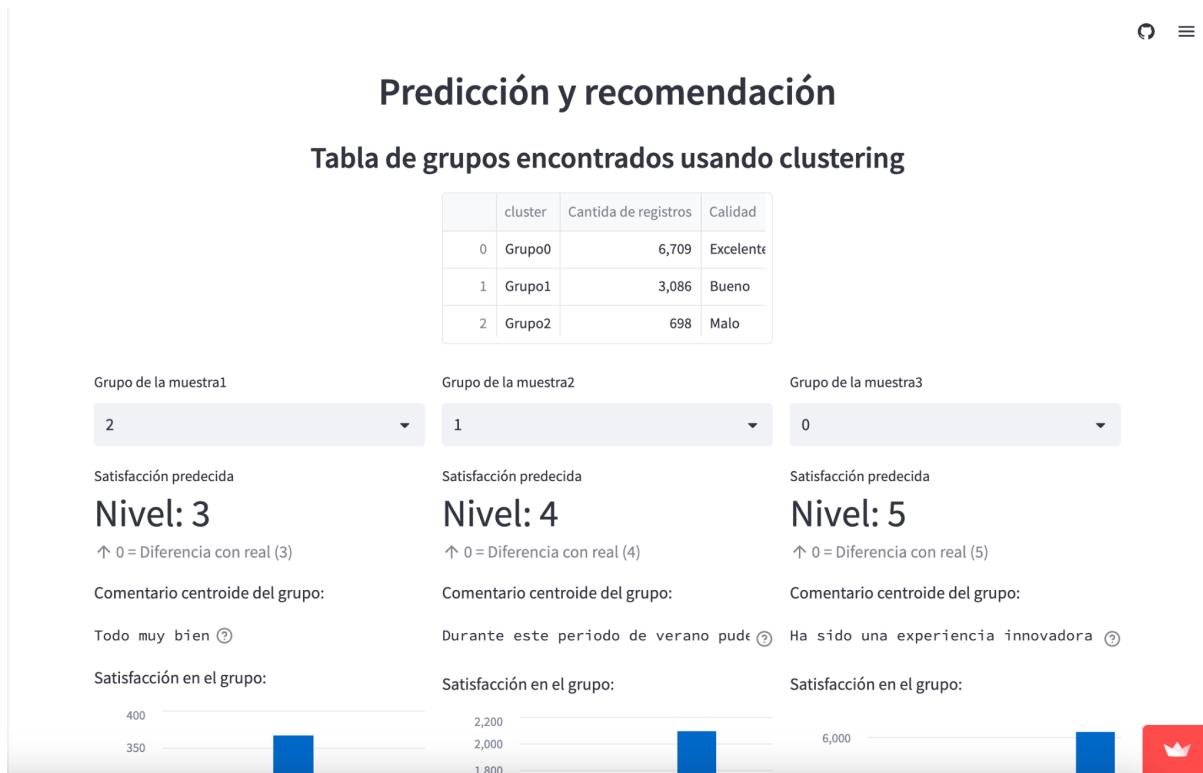


Figura 31: Muestra de las organizaciones por clusters

Para apreciar el prototipo y observar todo lo que tiene el código se encuentra en la siguiente carpeta:

<https://drive.google.com/drive/folders/1g6geNARZibhMhCtqDKkWMIIUXNgB7V05?usp=sharing>

La página web del prototipo se puede encontrar en el siguiente enlace:

<https://exampleee.streamlit.app/>

El código para realizar todo el proyecto se encuentra en la siguiente carpeta:

<https://drive.google.com/drive/folders/1z1N8WZfNqM0gGQTCFs7rfGZPXVimdou7?usp=sharing>

Recomendaciones

Recomendaciones

Recomendaciones para mejorar la satisfacción

Con base en los resultados obtenidos, el prototipo creado y los objetivos de la organización socio formadora, recomendamos hacer lo siguiente:

- Ofrecer retroalimentación: Al ser el factor que más afecta en la satisfacción de los estudiantes se sugiere crear una política en donde todas las organizaciones socio formadoras estén obligadas a ofrecer retroalimentación a los estudiantes durante la realización del proyecto.
- Mejorar la comunicación, la claridad de los objetivos y la organización: Con el fin de aumentar la satisfacción de los estudiantes, se sugiere establecer una comunicación clara y efectiva sobre los objetivos de cada servicio social, especialmente de los entregables. Esto ayudará a los estudiantes a comprender mejor las expectativas y a comprometerse de manera más significativa con las actividades propuestas. Además, para mejorar la organización se recomienda que todas las organizaciones tengan planes de trabajo con fechas y entregables requeridos.
- Revisar la carga de trabajo: Una de las quejas principales es que les piden mucho a los estudiantes en muy poco tiempo, para mejorar esto se recomienda revisar la cantidad de entregables que les pedirán a los estudiantes.
- Eliminar o cambiar las evaluaciones en la plataforma: La mayoría de los estudiantes consideran que las evaluaciones en Canvas no son de utilidad y que solo ponen en riesgo sus horas de trabajo. Se recomienda eliminar estos quizzes y cambiarlos por otras actividades que ayuden a los estudiantes a poner en práctica los valores que deben desarrollar.

- Evaluar la calidad de las organizaciones asociadas: Al tener un ranking y las agrupaciones, se recomienda analizar el grupo de las organizaciones que tienen mala puntuación y hacer cambios especializados. Para esto se recomienda realizar evaluaciones periódicas de estas organizaciones y establecer criterios de selección claros y rigurosos ayudará a mantener altos estándares de calidad.

Recomendaciones para mejorar los datos

- Fomentar la retroalimentación de los estudiantes: Las encuestas son de gran utilidad para seguir mejorando el servicio social, pero también se sugiere tener a una persona o a un departamento con el que un estudiante se pueda comunicar para reportar faltas de respeto o actitudes groseras de los encargados de los proyectos.
- Preguntas negativas y anonimato: Implementar en la encuesta más preguntas que estén enfocadas hacia los aspectos negativos de las experiencias. Otra alternativa sería hacer otra encuesta en la que se asegure el anonimato, ya que muchos estudiantes pueden estar limitando sus comentarios malos porque saben que se tiene registro de quién fue el que escribió esa respuesta.
- Tema de la organización: Agregar una pregunta en la encuesta para elegir en una lista el tema que se trabaja en la OSF, esto con el objetivo de ver cuál es el rol que desempeñan los estudiantes .
- Respuestas completas: Hacer que las preguntas que la Dirección de Servicio social considere más importantes sean obligatorias. Por ejemplo: el nombre de la OSF en la que participó el estudiante.

Recomendaciones técnicas

Como en todo proyecto, existen varias áreas de mejora. Al tener que traducir los comentarios se puede llegar a perder información, hay ocasiones en las que la traducción no es tan precisa y puede meter ruido a la hora de hacer el análisis. Para mejorar esto se tendría que hacer una investigación profunda en las herramientas disponibles para el procesamiento de texto en español o intentar entrenar modelos de aprendizaje para hacer nuestro propio análisis de sentimiento.

En la mayoría de los modelos utilizados el método de validación cruzada que se usó fue el de *Holdout*; sin embargo, esto puede llegar a perjudicar el rendimiento de algunos modelos, ya que en nuestro caso los datos de la variable objetivo no estaban balanceados y aunque en algunos casos utilizamos parámetros para intentar compensar este desbalance en algunos modelos se podía observar que había un sesgo considerable hacia la clase dominante. Para mejorar esto se recomienda implementar estrategias como hacer oversampling o undersampling. También se podría utilizar otros métodos de validación especialmente para los casos en que los datos no estén balanceados como *K-fold*.

Otra área de oportunidad es mejorar la búsqueda de parámetros para la creación de modelos. Si bien en algunos casos se realizó una búsqueda sistemática de los mejores parámetros, en otros casos los parámetros se establecieron mediante experimentación. Para lograr un enfoque más consistente y óptimo, se debería haber implementado un proceso más riguroso y sistemático para la búsqueda y selección de los parámetros adecuados. Esto habría permitido maximizar el rendimiento de los modelos y obtener resultados más confiables y generalizables.

Siguientes pasos

Una vez completado el análisis detallado de las respuestas obtenidas en la encuesta de conclusión del servicio social, el equipo y el proyecto deben seguir una serie de pasos para aprovechar los hallazgos y mejorar la experiencia de los estudiantes. En primer lugar, es importante considerar los factores identificados como influyentes en la satisfacción del usuario y determinar cómo pueden abordarse y mejorarse. Esto puede incluir cambios en las políticas o procedimientos de la organización formadora, así como la implementación de programas de apoyo adicionales para ayudar a fomentar una cultura de mejora continua y colaboración entre el Instituto Tecnológico y de Estudios Superiores de Monterrey y la organización formadora. Es crucial establecer un plan de acción basado en las recomendaciones y los hallazgos del estudio. El equipo debe trabajar en conjunto con la organización formadora para identificar las acciones específicas que se llevarán a cabo, estableciendo plazos y asignando responsabilidades claras. Es importante contar con un seguimiento regular para monitorear el progreso y realizar ajustes si es necesario.

Como en todo proyecto, siempre hay mejoras que se pueden hacer. Se sugiere realizar modelos más rigurosos y completos como una red neuronal para mejorar los resultados y desarrollar herramientas más complejas. Otra mejora que se puede hacer en un futuro es la creación de un sistema de recomendación que a partir de las respuestas de los estudiantes se les recomiende proyectos en los que deberían participar. Por ejemplo, encontrar los alumnos con más mayor similitud a un usuario, donde se puede incluir información demográfica, carrera, año de estudio, etc. para recomendar las experiencias que ya cursaron éstos. Asimismo, agregando en la interfaz creada se propone añadir un foro o chat donde los participantes puedan poner sus opiniones de manera anónima sobre sus experiencias sobre las OSF. Esto con el objetivo de dar un espacio a los usuarios donde puedan dejar comentarios que de lo contrario no plasmaron en la encuesta para no quedar mal con las organizaciones.

Es importante tomar en cuenta el aspecto moral y ético ya que al ser un lugar anónimo los usuarios pueden dejar comentarios obscenos, no apropiados, o de odio, por lo que un sistema de moderación de éstos es necesario.

Por otro lado, se pueden encontrar y generar modelos con mejores desempeños al usar unos más complejos y con un búsqueda de hiperparámetros exhaustiva. Para lograr esto se puede utilizar herramientas y equipo con el propósito de aceleración computacional mediante el uso de GPUs o tarjetas gráficas junto con librerías que habilitan su uso como CUDA desarrollada por NVIDIA. Finalmente, después de implementar las mejoras, es necesario realizar encuestas de seguimiento o estudios posteriores para evaluar el impacto de las mejoras implementadas. Esto permitirá medir el grado de satisfacción de los estudiantes después de aplicar los cambios y determinar si se han logrado los objetivos establecidos.

Bibliografía

- aws. (s. f.). *¿Qué es el procesamiento de lenguaje natural?* - Explicación del procesamiento de lenguaje natural - AWS. Amazon Web Services, Inc. <https://aws.amazon.com/es/what-is/nlp/#:~:text=La%20generaci%C3%B3n%20de%20lenguaje%20natural,personal%20de%20atenci%C3%B3n%20al%20cliente>.
- aws.(s.f.)*¿Qué es una red neuronal?* - Explicación de las redes neuronales artificiales - AWS. Amazon Web Services, Inc. <https://aws.amazon.com/es/what-is/neural-network/>
- Al-Masri, A. (2022). How Does k-Means Clustering in Machine Learning Work? *Medium*. <https://towardsdatascience.com/how-does-k-means-clustering-in-machine-learning-work-fdaaaf5acfa0>
- Ahmed, C., ElKorany, A. & ElSayed, E. Prediction of customer's perception in social networks by integrating sentiment analysis and machine learning. *J Intell Inf Syst* (2022). <https://doi.org/10.1007/s10844-022-00756-y>
- Angilella, S., Corrente, S., Greco, S., & Słowiński, R. (2014). MUSA-INT: Multicriteria customer satisfaction analysis with interacting criteria. *Omega*, 42(1), 189-200. <https://www.sciencedirect.com/science/article/pii/S0305048313000595>
- Baumann, C., Elliott, G., & Burton, S. (2012). Modeling customer satisfaction and loyalty: survey data versus data mining. *Journal of services marketing*, 26(3), 148-157. https://drive.google.com/file/d/1axJI1c5p5GifzCIWFnOi0cAoLI0ng_R/view?usp=haring
- Bernardes, V. (2023). How to analyze customer reviews with NLP: a case study. *Blog | Imaginary Cloud*. <https://www.imaginarycloud.com/blog/how-to-analyze-customer-reviews-with-nlp-case-study/>

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

<https://drive.google.com/file/d/1xF9F3NNeL-lazBkrigXWrdQYenIMJbA9/view?usp=sharing>

Doing-Harris, K., Mowery, D. L., Daniels, C., Chapman, W. W., & Conway, M. (2016). Understanding patient satisfaction with received healthcare services: a natural language processing approach. In *AMIA annual symposium proceedings* (Vol. 2016, p. 524). American Medical Informatics Association.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5333198/>

Great Learning Team. (2022). *What is Latent Dirichlet Allocation (LDA)*. Great Learning Blog: Free Resources What Matters to Shape Your Career!

<https://www.mygreatlearning.com/blog/understanding-latent-dirichlet-allocation/>

Hotz, N. (2023) What is CRISP DM? *Data Science Process Alliance*.
<https://www.datascience-pm.com/crisp-dm-2/>

Hulstaert, L. (2017). *LDA2vec: Word Embeddings in Topic Models*.
<https://www.datacamp.com/tutorial/lda2vec-topic-model>

Kumar, S. (2021). Agglomerative Clustering and Dendograms — Explained. *Medium*.
<https://towardsdatascience.com/agglomerative-clustering-and-dendograms-explained-29fc12b85f23>

Jackson, P., & Moulinier, I. (2007). Natural language processing for online applications: *Text retrieval, extraction and categorization* (Vol. 5). John Benjamins Publishing.
https://drive.google.com/file/d/1oWbhOEDfFCpI7HbND8cE0G_zCYAYWQFv/view?usp=sharing

Jarroba, R. (2017). *Selección del número óptimo de Clusters - Jarroba*. Jarroba.
<https://jarroba.com/seleccion-del-numero-optimo-clusters/>

- Jaadi, Z. (2021). A Step-by-Step Explanation of Principal Component Analysis (PCA). *Built In.*
<https://builtin.com/data-science/step-step-explanation-principal-component-analysis>
- Keita, Z. (2022). *An Introduction to Using Transformers and Hugging Face.*
<https://www.datacamp.com/tutorial/an-introduction-to-using-transformers-and-hugging-face>
- Mora, J. (2022). Proyecciones de la ciencia de datos en la cirugía cardíaca. *Revista Médica Clínica Las Condes*, 33(3), 294-306.
- Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture* (pp. 70-77). <https://dl.acm.org/doi/abs/10.1145/945645.945658>
- NVIDIA (2021) Your GPU Compute Capability. <https://developer.nvidia.com/cuda-gpus>
- Saragih, M., Girsang, A. (2017). *The data comments are classified into some categories, positive, negative.* International Conference on Sustainable Information Engineering and Technology (SIET), IEEE, pp. 24–29.
- Piris, Y., & Gay, A. C. (2021). Customer satisfaction and natural language processing. *Journal of Business Research*, 124, 264-271.
<https://www.sciencedirect.com/science/article/pii/S0148296320308249>
- Pham, D. T., Dimov, S. S., & Nguyen, C. D. (2005). Selection of K in K-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 219(1), 103-119.
- Rana, T.A., Cheah, YN. Aspect extraction in sentiment analysis: comparative analysis and survey. *Artif Intell Rev* 46, 459–483 (2016).
<https://doi.org/10.1007/s10462-016-9472-z>

UNIR México. *Árboles de decisión: qué son y cuál es su uso en Big Data.* (2023).

<https://mexico.unir.net/ingenieria/noticias/arboles-de-decision/#:~:text=Los%20%C3%A1rboles%20de%20decisi%C3%B3n%20son%20algoritmos%20estad%C3%ADsticos%20o%20t%C3%A9cnicas%20de,relaci%C3%B3n%20entre%20distintas%20variables%20para>

Vaca, A. (s. f.). *Transformers en Procesamiento del Lenguaje Natural.* iic UNAM.

<https://www.iic.uam.es/innovacion/transformers-en-procesamiento-del-lenguaje-natural/>

Yan, R., Jiang, X., Wang, W. et al. (2022). Materials information extraction via automatically generated corpus. *Sci Data* 9, 401 <https://doi.org/10.1038/s41597-022-01492-2>

Wu, Y., Ianakiev, K., & Govindaraju, V. (2002). Improved k-nearest neighbor classification. *Pattern recognition*, 35(10), 2311-2318.

Zárate-Valderrama, J., Bedregal-Alpaca, N., & Cornejo-Aparicio, V. (2021). Modelos de clasificación para reconocer patrones de deserción en estudiantes universitarios. *Ingeniare. Revista chilena de ingeniería*, 29(1), 168-177.

Himmi, A., Irurozki, E., Noiry, N., Clemenccon, S., & Colombo, P. (2023). *Towards More Robust NLP System Evaluation: Handling Missing Scores in Benchmarks.* arXiv preprint arXiv:2305.10284.

Zimmermann, A. (2020). Method evaluation, parameterization, and result validation in unsupervised data mining: A critical survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(2), e1330.