**Ingeniería de Sistemas y Computación**

**Pregrado**

**ISIS3301 – Inteligencia de Negocios**

**Semestre 2024-10**

**Proyecto 1 – Etapa 1**

**Santiago Chamie Rey - 202122182**

**Daniel Escalante Pérez – 202122384**

**Daniel Felipe Vargas Ulloa - 202123899**

Contenido

[Entendimiento del negocio y enfoque analítico 2](#_Toc1886411061)

[Entendimiento y preparación de los datos 3](#_Toc1955197390)

[Modelado y evaluación 3](#_Toc365375940)

[Naive Bayes 3](#_Toc58206416)

[Árboles de decisión 4](#_Toc621861278)

[Logistic Regression – Daniel Escalante Perez 4](#_Toc2018549400)

[Elección del mejor modelo 7](#_Toc341496806)

[Mapa de actores relacionado con el producto de datos creado 9](#_Toc378744729)

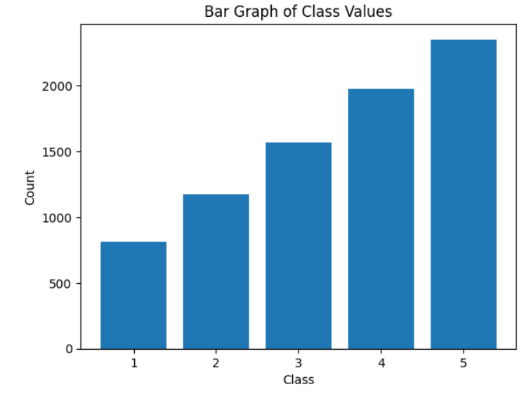
[Trabajo en equipo 10](#_Toc1498940786)

## Entendimiento del negocio y enfoque analítico

|  |  |
| --- | --- |
| **Característica** | **Respuesta** |
| **Oportunidad/problema Negocio** | El problema principal que enfrenta el negocio es la necesidad de comprender qué características específicas hacen que ciertos sitios turísticos sean atractivos para los turistas, tanto locales como internacionales. Además, existe la preocupación por identificar aquellos aspectos que están afectando negativamente la popularidad de ciertos destinos turísticos. La oportunidad radica en utilizar el aprendizaje automático para analizar grandes volúmenes de reseñas de turistas y extraer conocimientos sobre las características que influyen en la percepción de los visitantes. Esto permitirá a las organizaciones del sector turismo tomar decisiones más informadas para mejorar la calidad de los servicios, adaptar sus estrategias de marketing y, en última instancia, aumentar el atractivo de los destinos turísticos en Colombia. |
| **Enfoque analítico (Descripción del requerimiento desde el punto de vista de aprendizaje automático) e incluya las técnicas y algoritmos que propone utilizar.** | Este proyecto busca aprovechar el potencial del aprendizaje automático para analizar y comprender las reseñas de sitios turísticos, con el objetivo de identificar qué características hacen que un lugar sea atractivo para los turistas y cuáles no. La solución propuesta implica un enfoque de aprendizaje supervisado, donde se utilizarán técnicas de procesamiento de lenguaje natural (NLP) para resolver este problema de clasificación que determina la clasificación esperada de una reseña, permitiendo encontrar qué características afectan más significativamente el puntaje que recibe un hotel. Para este propósito, se aplicarán los algoritmos de clasificación de árboles de decisión, Naive Bayes classifier (Clasificador bayesiano) y regresión logística. Cada uno de estos algoritmos tiene ventajas específicas:     * Los árboles de decisión son intuitivos y pueden manejar características no lineales. * Los clasificadores Naive Bayes son simples y rápidos, lo que los hace particularmente adecuados para tareas con grandes conjuntos de datos y cuando los recursos computacionales son limitados. * La regresión logística es efectiva cuando se requiere una interpretación sencilla de los resultados. Se buscará determinar cuál de estos algoritmos ofrece la mayor precisión para este problema particular.   Para preparar los datos, se realizará un análisis exploratorio para identificar posibles problemas de calidad de los datos, como valores faltantes, duplicados o atípicos. Luego, se aplicarán técnicas de preprocesamiento de texto, como tokenización, eliminación de stopwords y lematización, para convertir las reseñas en vectores numéricos que puedan ser utilizados por los algoritmos de aprendizaje automático. Para evaluar los modelos, se utilizarán métricas como precisión, recall y F1-score, así como validación cruzada para garantizar la generalización de los resultados. Además, se realizará una comparación de los modelos para determinar cuál es el más adecuado para este problema específico |
| **Organización y rol dentro de ella que se beneficia con la oportunidad definida** | El Ministerio de Comercio, Industria y Turismo de Colombia, así como la Asociación Hotelera y Turística de Colombia – COTELCO, junto con las cadenas hoteleras de renombre como Hilton, Hoteles Estelar y Holiday Inn, son los principales beneficiarios de este proyecto. Al comprender mejor las características que influyen en la percepción de los turistas sobre los sitios turísticos, estas organizaciones podrán adaptar sus estrategias de marketing y mejorar la calidad de sus servicios para aumentar la atracción de turistas, tanto nacionales como internacionales, hacia Colombia. Esto, a su vez, impulsaría el crecimiento económico del sector turístico y contribuiría al desarrollo socioeconómico del país. |
| **Contacto con experto externo al proyecto y detalles de la planeación** | La planeación del trabajo con el grupo de estadística, conformado por Mariana Gutierrez y Emilia Hernandez se realizó a través de un acuerdo que estableció los términos de colaboración entre ambos grupos. El acuerdo definió las responsabilidades de cada grupo, incluyendo la validación del enfoque del proyecto, la definición de los impactos esperados y la planeación de la comunicación.  Se estableció como fecha de inicio de la comunicación el 1/4/2024 (fecha posterior al fin de la semana santa) utilizando el correo electrónico como canal principal. El grupo de inteligencia de negocios se comprometió a presentar el proyecto y sus objetivos, incluir las sugerencias del grupo de estadística y mantenerlos informados del progreso del proyecto. Ambas partes se comprometieron a trabajar de manera colaborativa y profesional para el éxito del proyecto |

## Entendimiento y preparación de los datos

Los datos utilizados en este proyecto son bastante diferentes a los que estamos acostumbrados a manejar, ya que tienen únicamente dos columnas: Review y Class. Review contiene una cadena de texto con la reseña de un hotel y Class contiene el valor numérico que el reseñador le asignó a su propia reseña.

Haciendo un histograma podemos ver la distribución de las calificaciones por parte de los clientes, donde las más común es 5 y la menos es 1. Esto nos sirve para considerar a futuro que un modelo de clasificación entrenado con estos datos podría asignar valores de 5 más frecuentemente que valores de 1.

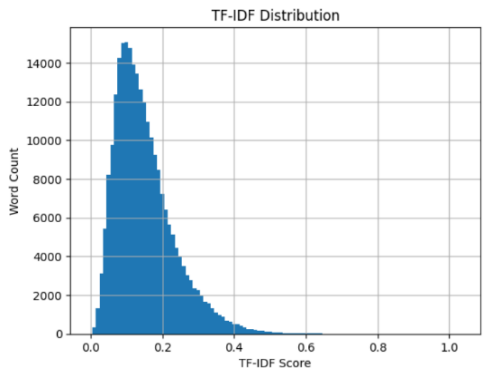
Analizamos también las reseñas de los clientes, donde pudimos evidenciar rangos de longitudes muy diversos. Sin embargo, pudimos comprobar que ninguna reseña estaba vacía y que las reseñas con valores más extremos eran coherentes con el tema en cuestión.

La preparación de los datos en este proyecto la hicimos usando la metodología bag-of-words, donde, al finalizar, cada palabra se representará como una columna de una matriz y su existencia en un documento dependerá de si este valor es 0 o no. En caso de si estar, le asignaremos su valor TF-IDF, la frecuencia del término (cantidad de veces que palabra t aparece en documento p) sobre la frecuencia inversa de documento (total de documentos sobre el total de documentos que contienen el término); el cual nos servirá para la creación de los modelos.

El primer paso de la preparación es poner cada reseña en su forma más estándar eliminando caracteres no ASCII, poniéndo las letras en minúscula, eliminando puntuación, eliminando números, eliminando tildes y eliminando stopwords (o palabras con poco valor en su significado).

Con esta transformación, hacemos un proceso de tokenización, donde creamos una nueva columna ‘words’ donde se guarda una lista de las palabras filtradas que hay en el documento. Estas listas nos permiten normalizar las palabras usando el método de lematización, dónde cada palabra se lleva a su raíz, así agrupando palabras similares como carros y carro, las cuales irían a la misma raíz.

Al normalizar las palabras, podemos unirlas nuevamente, luego de quitar los espacios creados por palabras eliminadas. Tras un análisis de las nuevas palabras podemos evidenciar que la distribución de los valores disminuye considerablemente luego de las transformaciones.

Finalmente, con las palabras transformadas podemos utilizar el TfidfVectorizer para calcular los valores tf-idf de cada palabra y luego usarlos para crear la matriz que usaremos para entrenar los modelos. Eliminamos las 71 filas repetidas creadas por la normalización para concluir la preparación de los datos

## Modelado y evaluación

### Naive Bayes

El algoritmo de Naive Bayes Classifier es un método de clasificación probabilístico basado en el teorema de Bayes con una suposición ingenua (de ahí su nombre "naive"). Este clasificador se utiliza comúnmente en tareas de clasificación, incluido el procesamiento de lenguaje natural (PLN).

**Funcionamiento del algoritmo**

El algoritmo de Naive Bayes utiliza el teorema de Bayes para calcular la probabilidad condicional de que una instancia pertenezca a una determinada clase dada la presencia de ciertas características. En nuestro caso, este algoritmo nos permitirá asignarle una clasificación a una reseña a partir de las palabras que esta contenga.

El teorema de Bayes es una regla fundamental en la teoría de la probabilidad que permite calcular la probabilidad condicional de un evento dado el conocimiento de otro evento relacionado. Formalmente, el teorema de Bayes establece que la probabilidad de que ocurra un evento A dado que ha ocurrido un evento B. Se puede calcular utilizando la fórmula: , donde:

* P(A∣B) es la probabilidad condicional de que ocurra el evento A dado que ha ocurrido el evento B.
* P(A) y P(B) son las probabilidades de que ocurran los eventos A y B respectivamente.

El algoritmo de Naive Bayes utiliza el teorema de Bayes para calcular la probabilidad condicional de que una instancia pertenezca a una determinada clase dada la presencia de ciertas características. Supongamos que tenemos un conjunto de datos con características (variables predictoras, en nuestro caso estas siendo palabras) X1,X2,...,Xn y una clase objetivo C (en nuestro caso la clase objetivo siendo el valor de la calificación del 1 al 5). El objetivo es predecir la clase de una nueva instancia utilizando las características observadas. El algoritmo calcula la probabilidad de que una instancia pertenezca a cada clase Ci utilizando la fórmula del teorema de Bayes: , donde:

* P(C\_i | X\_1, X\_2, ..., X\_n) es la probabilidad condicional de que una instancia pertenezca a la clase C\_i dado que tiene características X\_1, X\_2, ..., X\_n.
* P(X\_1, X\_2, ..., X\_n | C\_i) es la probabilidad condicional de observar las características X\_1, X\_2, ..., X\_n dado que la instancia pertenece a la clase C\_i.
* P(C\_i) es la probabilidad previa de que una instancia pertenezca a la clase C\_i.
* P(X\_1, X\_2, ..., X\_n) es la probabilidad total de observar las características X\_1, X\_2, ..., X\_n, también conocida como probabilidad marginal.

El clasificador predice la clase Ci que maximiza esta probabilidad condicional.

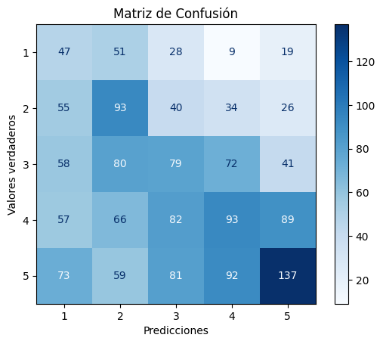
**Resultados del algoritmo y análisis.**

A pesar de los esfuerzos realizados en el modelado, los resultados obtenidos no alcanzan el nivel deseado de calidad. Si bien las métricas de precisión y recall son consistentes en su desempeño moderado, con valores oscilando entre 0.16 y 0.44 para la precisión y entre 0.24 y 0.38 para recall, respectivamente, el puntaje F1 general del modelo es de solo 0.29. Esto indica que el modelo no logra un equilibrio satisfactorio entre la precisión y el recall, lo que puede ser atribuido a una proporción desfavorable de falsos positivos y falsos negativos en las predicciones.

Un puntaje F1 de 0.29 sugiere que el modelo no está generalizando bien a través de las clases, lo que implica una capacidad limitada para clasificar correctamente las reseñas en las cinco categorías. En nuestro contexto de clasificación de reseñas, esto significa que el modelo está cometiendo errores tanto al sobrevalorar como al subvalorar las reseñas. En términos del negocio, esto podría traducirse en una experiencia insatisfactoria para aquellos interesados en el obtener información valiosa del modelo, como pueden ser las características de un hotel que más influyen en una review.

En última instancia, determinar si el modelo es adecuado para el negocio dependerá de qué tan cerca esté de calificar correctamente las reseñas en la escala de 1 a 5 (Dado que un modelo que incorrectamente asigna una calificación de 5 a una review etiquetada con un score de 1 seria completamente inútil para el negocio, mientras que un modelo que incorrectamente asigna un 4 en lugar de un 5 es un modelo que si tiene un entendimiento decente de como las características afectan a la calificación final de una review). Para evaluar esto más a fondo, sería útil construir una matriz de confusión que muestre con claridad los errores de clasificación del modelo.

A partir de la matriz de confusión, podemos observar que la mayoría de las reseñas clasificadas incorrectamente por el modelo están asignadas a valores cercanos al valor verdadero. Por ejemplo, si el valor verdadero de la reseña es 4 o 5, la mayoría de los errores del modelo están en los valores 3 o 5, relativamente cercanos al correcto. Además, son pocos los errores que están altamente mal clasificados.

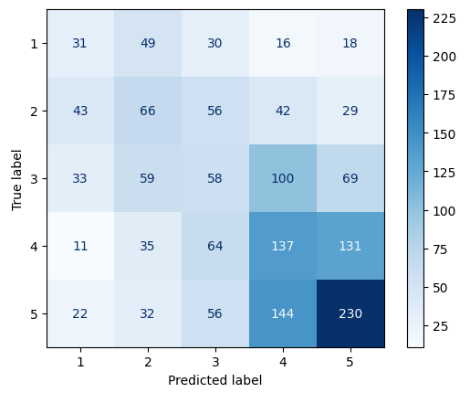
Sin embargo, existe una clara tendencia en la matriz de confusión hacia la sobrevaloración de las reseñas. Esto sugiere que el modelo tiende a identificar correctamente los elementos que contribuyen a una reseña positiva, pero tiene dificultades para identificar los aspectos negativos de una reseña. Esto es un factor crucial, ya que los stakeholders como los gerentes de hoteles están interesados en identificar las características específicas de sus hoteles que llevan a reseñas negativas para poder corregirlas.

### Árboles de decisión

El algoritmo de árbol de decisión fue el único de los algoritmos de clasificación que utilizamos en el proyecto de los que vimos en clase. Los otros dos algoritmos utilizados son algunos de los más utilizados al clasificar textos, como se hace en este proyecto. No así, elegimos utilizar este algoritmo como un punto de partida y comparación de resultados, y familiarizarnos con el proceso de clasificar textos.

Este algoritmo funciona a partir de la subdivisión del conjunto de datos basado en las características particulares de los conjuntos. Los hiperparámetros de este algoritmo es la profundidad del árbol (que tantos subconjuntos se hacen) y las columnas de decisión. Para encontrar dónde dividir el árbol se hace el cálculo del coeficiente Gini o entropía (dependiendo del que se elija) para cada una de las columnas de los datos, eligiendo la que tenga un valor más significativo. Este proceso se repite tantas veces lo permitan los datos y la profundidad del árbol para formar grupos característicos para cada calificación [0,5].

Para encontrar la profundidad adecuada del algoritmo se hace un proceso de búsqueda de grilla, utilizando una lista de valores plausibles de profundidad, lo cuales se computan para sacar algún coeficiente y compararlos entre sí. Ya con este coeficiente elegido se puede entonces crear el modelo final, dividiendo los datos en grupos 80-20 para entrenar el modelo y hacer pruebas respectivamente.

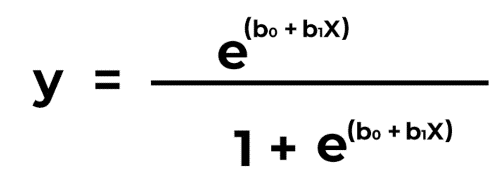
Al finalizar la construcción del modelo dió un resultado de 0.33 en el puntaje f1, métrica de evaluación elegida para este caso, y la siguiente matriz de confusión.   
El modelo entrenado tuvo un puntaje f1 de 0.33, logrando un máximo f1 para predecir los puntajes de 5 en las reseñas, lo que tiene sentido ya que este es el puntaje más común. Este modelo logra ser satisfactorio para predecir los puntajes, sin embargo, sigue faltándole la certeza esperada para poder recomendárselo al negocio.

### Logistic Regression – Daniel Escalante Perez

#### Descripción

La regresión logística es un algoritmo de machine learning de aprendizaje supervisado para la tarea de clasificación, enfocado en la tarea de analítica predictiva. Este algoritmo predice la probabilidad para un ejemplar de que ocurra un evento, en este caso de que se tome cierta clase o no se tome una clase. A pesar de ser un algoritmo para clases binarias, se puede aplicar la técnica OvR (One vs Rest) que implica crear tantos modelos como clases se vayan a analizar. En cada una se toma una clase diferente a analizar y todas las demás clases se ponen en una clase nueva: "todas las que no son la seleccionada", para tener únicamente dos clases. En la predicción con OvR, se busca el clasificador con mayor probabilidad en su clase positiva para el dato estudiado.

La regresión logística se basa en el uso de la función sigmoidea, que mapea cualquier valor real en un rango de 0 a 1.

La fórmula de la regresión logística toma una fórmula similar a la de una regresión lineal con un bias (b0) y un coeficiente (b1) (del entrenamiento). Esto relaciona linealmente la variable independiente con una variable x con la cual se calcula la probabilidad del evento positivo usando la función sigmoidea. De esta manera, cuando un mayor valor de una variable explicativa hace que se tome la clase positiva, el coeficiente de esta variable será positivo para hacer crecer la expresión lineal. Así, incrementa la probabilidad de la clase y es más posible que sea seleccionada.

Con respecto a la configuración de hiperparametros, en scilearn se deben configurar el ‘solver,’ la penalidad y el ‘regularization strength’. El ‘solver’ es el algoritmo que se usa para resolver el problema de optimización. La penalidad es usada para evitar el sobreajuste y generalizar mejor el problema, evitando que el modelo se haga muy complejo. Por último, el ‘regularization strength’ (C), se usa para evitar que el modelo sea muy complejo; entre más pequeño, más regulado está el modelo, entre más alto, más importancia se da a los datos de entrenamiento.

La mejor combinación de hiperparametros se encuentra usando la búsqueda de hiperparámetros con validación cruzada. Esta consiste en dividir los datos de entrenamiento y dejar aparte un pedazo, entrenar sobre el resto y evaluar sobre el pedazo, y repetir cambiando el pedazo. Posteriormente, se saca el promedio de las métricas de las pruebas para toda combinación configurada de hiperparametros.

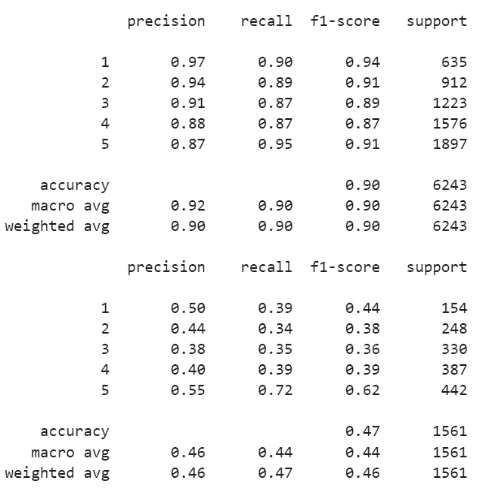
El modelo tiene múltiples supuestos, entre los principales: la no multicolinealidad entre las variables explicativas; que se eviten los valores atípicos extremos; independencia de observaciones; y que exista una relación lineal entre las variables y el logaritmo de éxitos/fallos. Algo recomendable es que el set de datos sea linealmente separable (Que se puede probar entrenando un perceptron, si es linealmente es separable, encontrará un hiperplano en un numero finito de iteraciones).

#### Implementación

Implementando el modelo, se intentó validar varios supuestos. Primero, el supuesto de multicolinealidad, que dice que se deben eliminar las variables correlacionadas con un coeficiente alto. Al hacer esto la calidad del modelo disminuyó considerablemente por lo que se decidió dejar los datos como estaban. Por otro lado, revisando los datos atípicos, se vio que la mayoría de los datos tienen campos atípicos. No obstante, decidimos mantenerlos porque pueden tener valores importantes para el análisis. Como último paso antes de entrenar el modelo, se hizo el entrenamiento del perceptrón. Este convergió, implicando que los datos son linealmente separables, ayudando con el desempeño del modelo.

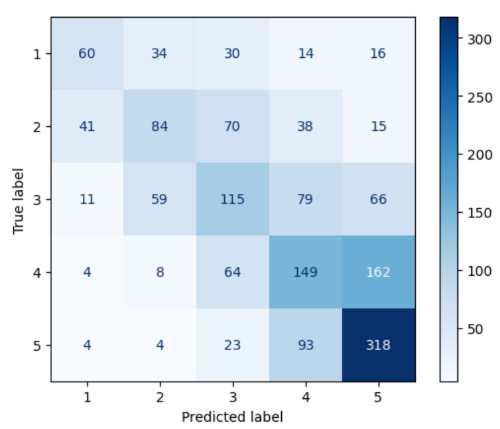
Para crear el modelo, se usó validación cruzada para hallar el valor apropiado del hiperparametros C y se buscó el mejor estimador de 3 agrupaciones de validación cruzada optimizando el f1-score.

#### Análisis Cuantitativo

En la tabla anterior se pueden observar las métricas sobre los clasificadores que componen el clasificador OvR Se puede destacar que se obtuvo un f1-score de 0.47. Esto no es bueno ni malo, es regular. Este puntaje es la media de los f1-scores individuales de cada clasificador de uno contra el resto. El f1-score corresponde a una media armónica entre la precisión y el recall. La precisión es un porcentaje de cuantas se acertaron entre todas las marcadas como dicha clase. El recall es un porcentaje de cuantas se acertaron entre todas las que verdaderamente eran de esa clase.

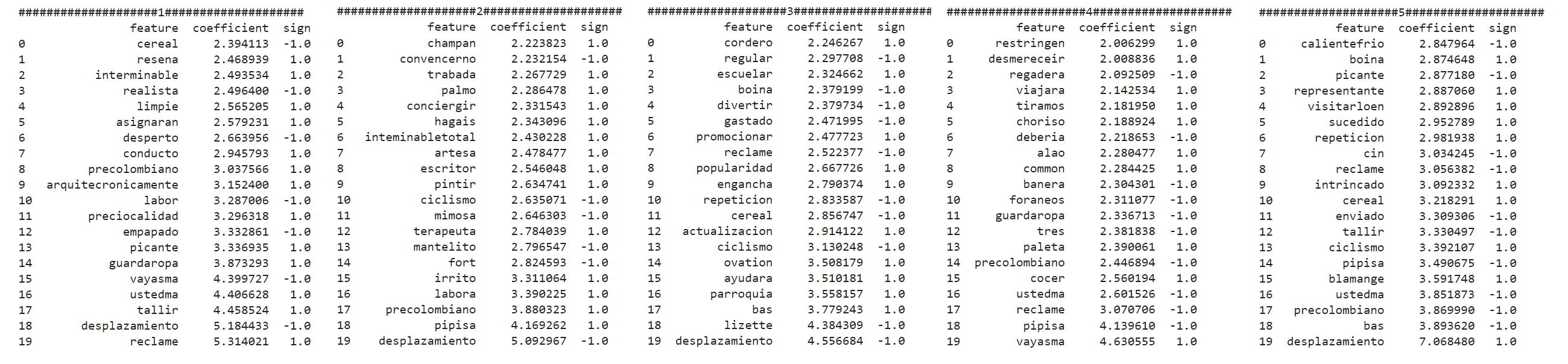
Se observa que el 72 % de las que eran una calificación de 5, se acertaron. Esto no es excelente, pero es bueno porque con el clasificador se puede ver que caracterizan las reseñas de calificación 5 y poder ver los factores en la experiencia de los clientes que influyen en lograr esta calificación. De esta forma se puede lograr el objetivo de ver que hace a los sitios turísticos atractivos. Entre todos los marcados como 5, se acertó el 55 %, lo que es regular porque el modelo podría marcar datos de otra clase como la suya. No obstante, como se verá más adelante, estos falsos positivos podrían corresponder a clases cercanas como 4, que también indican satisfacción. De esta forma se puede mantener que se pueden caracterizar los sitios atractivos.

Los clasificadores de la clase 1 a la clase 4 tienen un recall menor a 0.39, lo cual se podría considerar malo puesto que solo está reconociendo cerca del 40% de los datos que verdaderamente tienen esa calificación. Igualmente ocurre con la precisión, las precisiones de estas clases son menores del 50%, es decir, aciertan la mitad de lo que marcan como de esa clase, lo que se podría comparar con el azar. Como se verá después, y como se mencionó antes, los malos resultados pueden atribuirse a la cercanía entre 2 clases. Como por ejemplo la clase 1 y 2 que ambas indican mala calificación, por esto pueden tener características similares y causar confusión en el modelo y en la clasificación.

En la matriz de confusión no se ven resultados tan negativos como los descritos previamente en las métricas. Esto se debe a que se ve con mayor oscuridad la diagonal de arriba-izquierda a abajo-derecha. En cuanto más oscuros son estos cuadrados, indica que se acertaron más valores de esas clases y mejor será el modelo. Los errores (falsos positivos o falsos negativos), que se representan como todas las cajas diferentes a las de la diagonal mencionada, son en su mayoría más oscuras en los vecinos de los cuadrados de la diagonal. Lo anterior implica que la mayoría de los errores que ocurrieron fueron en clases cercanas a la verdadera. Esto se resalta viendo las filas de la matriz, en estas se pueden ver las fallas en las clases vecinas a la izquierda y derecha de las cajas de la diagonal.

Con todo esto en cuenta, se puede decir que el modelo puede distinguir que es una buena experiencia y que es una mala puesto que se acerca a clases similares así sea errónea la predicción. Se observa como en la clase 1 (terrible) los errores son principalmente con la clase 2. En la clase 4, se ve como los errores son principalmente con la clase 5. Entonces, se podría decir que para las clases cercanas se le dificulta distinguir que clase es la correcta ya que pueden representar cosas similares (ejemplo, 1 = terrible y 2 = malo). Otra razón de los resultados regulares del modelo puede ser la falta de datos y el desbalance de clases ya que habían muchos datos de la calificación 5 y pocos de calificaciones bajas.

Con la caracterización y con el mecanismo que califique sitios turísticos se podrá cumplir con el objetivo de negocio de ver oportunidades de mejora. Puede que no sea muy preciso el mecanismo automático, pero se acercara a las calificaciones reales por una unidad de puntaje en la mayoría de los casos como se observa en la matriz.



Anteriormente, se pueden ver 5 tablas de coeficientes del modelo con las palabras que más influenciaban cada clasificador. En estas tablas se puede ver la fuerza que ejercían sobre las clasificaciones. Adicionalmente, se puede ver la dirección del efecto, los negativos llevaban a la clase negativa (ósea que no es de esa clase que estaba identificando el clasificador), los positivos llevaban a la clase positiva (la clase del clasificador). De esta forma, viendo los positivos y de mayor coeficiente, podemos ver las que más influencian la clase a estudiar. Este análisis se hará en la parte cualitativa donde se observarán los conceptos principales que caracterizan cada clase.

Aun así, se puede ver cómo hay palabras que en las clases buenas (4 y 5) tienen un efecto positivo mientras que en las clases bajas (1, 2 y 3) tienen efecto negativo, respaldando la separación entre bueno y malo mencionada previamente.

## Elección del mejor modelo

Al momento de seleccionar el mejor modelo de los 3 que se realizaron, utilizando árboles de decisión, Naive Bayes y regresión logística, una buena forma de comparar la calidad de los modelos es con el F1 score. Esta métrica representa la precisión y exhaustividad del modelo en una sola medida, lo que lo hace útil para evaluar el equilibrio entre falsos positivos y falsos negativos. Un F1 score más alto indica un mejor equilibrio entre precisión y exhaustividad, lo que significa que el modelo tiene un mejor rendimiento general en la clasificación.

En cuanto a los F1 scores obtenidos para cada modelo, fueron 0.33 para árboles de decisión, 0.29 para Naive Bayes y 0.47 para regresión logística, tal y como se puede ver en las secciones anteriores. A partir de estos resultados, se determina que el modelo de regresión logística es el más efectivo, ya que alcanzó el F1 score más alto entre los tres algoritmos evaluados.

#### Análisis Cualitativo

Para hacer el análisis cualitativo del mejor modelo y ver la información que se puede extraer de los modelos, con el fin de cumplir con los objetivos del negocio, se harán análisis sobre las palabras que fueron más relevantes sobre la elección de cada clase dentro del clasificador desarrollado.

**Calificación 1:** Para esta primera clase que representa una terrible experiencia en el sitio turístico. Entre las palabras influyentes para elegir esta clase, la de mayor incidencia e influencia en que una reseña se marcara como 1 fue "reclame", que probablemente se refiere a la necesidad de realizar reclamos dentro del sitio turístico. Otra palabra que aportaba a la selección de esta clase es "guardarropa", es posible que los visitantes tengan algún problema con este servicio y que debería revisarse. Otra palabra para destacar entre las más influyentes es "limpie", lo cual puede indicar alguna situación que se esté presentando con la limpieza de los sitios. Por último, se puede resaltar "interminable", lo cual puede referirse a esperas o aburrimiento durante algún tipo de servicio ofrecido por el sitio turístico.

**Calificación 2:** Para esta calificación que denota también un mal servicio se pueden resaltar múltiples aspectos. Primero que todo, se nota como una de las palabras con mayor incidencia es "champan" lo que puede indicar algún problema con los licores dentro del sitio turístico. Otra palabra es "conciergir" que puede referirse a los concierge del sitio turístico los cuales pueden haber prestado mal servicio en las instalaciones, en la orientación o en la reserva y planeación de servicios. Otra palabra observada es "irrito" lo cual indica irritación por parte del cliente por molestias dadas o problemas de actitud en el lado de los empleados. Otra palabra por mencionar es "trabada" lo que puede señalar algún mal funcionamiento de algún equipo en las atracciones o demoras en algún servicio.

**Calificación 3:** En la calificación 3, se suele indicar un servicio mediocre o regular, pero ni bueno o ni malo. Para esta calificación es difícil determinar alguna otra observación ya que, al ser un nivel medio, no queda claro si una palabra indica algo positivo o negativo en el sitio turístico. Ejemplos de esto es "popularidad", "parroquia" o "promocionar".

**Calificación 4:** Esta calificación indica algo que fue bueno, no excelente pero bueno. Otra palabra que se observa es "chorizo" lo cual puede indicar algún tipo de satisfacción con esta comida dentro del sitio turístico. Se notan también algunas palabras que alejan la reseña de la calificación de 4 que se presentaban en otras calificaciones inferiores, una de estas es "guardarropa", lo que valida el valor negativo de este aspecto en la calificación de 1. Una última palabra para mencionar que afecto positivamente es "paleta" lo cual puede referirse al alimento y la satisfacción de los clientes respecto este producto.

**Calificación 5:** Esta calificación representa un servicio excelente. En esta calificación se pueden observar algunas palabras que alejaban de puntuaciones más bajas pero que acá se observa que acercan a este nivel de satisfacción. Un ejemplo de esto es "desplazamiento", es posible que esto se refiere a que los servicios son cercanos entre si dentro de un sitio turístico o que el sitio ofrece desplazamiento entre localizaciones. Otra palabra es "ciclismo", lo que indica que los sitios turísticos que ofrecen alquiler o espacios para esta actividad suelen satisfacer a los clientes. Una palabra que aporta a la calificación de 5 es "cereal" lo que puede indicar la disponibilidad de este producto o alimento y la satisfacción gracias a este. Una última palabra para mencionar es "representante" que puede referirse a la disponibilidad o satisfacción con los representantes dentro del sitio turístico que pueden guiar, ayudar o hasta proveer los servicios.

En base a la información obtenida se pueden tomar múltiples estrategias por parte de la organización para mejorar los hoteles, aumentar su popularidad y fomentar el turismo.

1. Hacer revisión de los servicios de limpieza en los hoteles. Esto se debe garantizar en base a la influencia de este componente en las calificaciones negativas.

2. Mejorar la atención ante reclamos y mejorar las ayudas que provee el hotel ante inquietudes o problemas que tienen los visitantes. Esto se debe realizar ya que las ayudas pueden mejorar la experiencia de los visitantes ya que sus inconvenientes son solucionados y se evitan reclamos repetitivos o incomodidades futuras.

3. Realizar evaluación de la eficiencia de los procesos y la calidad del entretenimiento ya que la palabra "interminable" destacó. Para esto se pueden revisar opiniones sobre el entretenimiento y los tiempos promedios en los servicios y solucionar si no son satisfactorios los valores tomados.

4. Establecer o revisar la presencia y calidad de los representantes o conserjes. Esto se hace como consecuencia a que se vieron problemas con el término " conserje" y positividad con "representante". Es fundamental mejorar estos personajes que ofrecen servicios y facilitan la consolidación de las expectativas de los visitantes, además de ayudar y proveer orientación.

5. Implementar o evaluar servicios clave observados. Entre estos servicios se observó el ciclismo, por lo que se puede ofrecer clases, espacios o alquiler de equipos de esta diciplina. También, se puede ofrecer desplazamiento entre lugares turísticos o terminales de transporte y el hotel.

6. Ampliación y revisión de los servicios alimenticios. Se puede proponer esta actividad como resultado de la positividad que aportaron a la calificación los alimentos como el chorizo, las paletas y los cereales. Incluir estos componentes, y proveerlos con calidad pueden mejorar la satisfacción en los clientes. Si estos ya están presentes, entonces se puede revisar la calidad como puede ser con los licores, ya que el champagne fue marcado como algo negativo, por lo que hay que proveer calidad en este producto.

7. Evaluación de los guardarropas. Se debe revisar este servicio, tanto en los closets de las habitaciones como el servicio de almacenamiento de prendas en eventos. Este es esencial por la confianza que ponen los clientes al entregar artículos personales.

8. Evaluación de empleados. Se vio que la irritación fue algo negativo que fue resaltado por el modelo. Por esto se debe revisar la actitud de los empleados por si tienen poca paciencia con los clientes o están causándoles disgusto. Por otra parte, también se debe evaluar el desempeño de estos y la calidad de sus trabajos para evitar la irritación de los clientes.

Con esta información y estrategias, se puede concluir que la organización podrá tanto mejorar sus hoteles, ser más populares y mejorar la percepción de los clientes. Todo esto resultaría en mayor flujo de clientes, y claramente, beneficios económicos que pueden ayudar a crecer la organización, lo que beneficia a los stakeholders. Los stakeholders se verán beneficiados de otras formas que se verán en la siguiente sección.

## Mapa de actores relacionado con el producto de datos creado

Una organización que puede beneficiarse del resultado del modelo analítico planteado puede ser una cadena de hoteles, por ejemplo: Estelar, Hilton o Holiday Inn. Estas son organizaciones que operan múltiples establecimientos de alojamiento bajo un mismo nombre. Normalmente, estos buscan una marca representativa y consistente a través de sus establecimientos. Por esto, se pueden encontrar similitudes en diseño, servicio y calidad a través de sus establecimientos. Comúnmente, estas cadenas tienen presencia en varias partes dentro de un país o en varios países.

|  |  |  |  |
| --- | --- | --- | --- |
| Rol | Tipo | Beneficio | Riesgo |
| Dirección general | Financiador | Mejora en la calidad de la operación y acercamiento a los objetivos de caracterización de sitios turísticos. Información basada en analítica para acciones estratégicas. | Si los resultados no son los esperados se pueden perder los recursos usados en el proyecto. Si los resultados fallan, se podrían tomar acciones erróneas y causar daños a la cadena de hoteles. |
| Empleados | Empleados | Retroalimentación de su trabajo para poder mejorar y abarcar aspectos deficientes. Retroalimentación constructiva de las labores hechas. | Si el modelo falla en reconocer las características, los empleados podrían ser criticados sin fundamento o erróneamente, adoptando o cambiando la forma de trabajo innecesariamente. |
| Control de Calidad | Cliente | Poder identificar problemas de calidad dentro de la organización o comunes en otras para poder evitarlos y/o abarcarlos. | Que las características encontradas por el modelo se localicen en una región con más reseñas o en un hotel con muchas reseñas negativas y no poder aprender aspectos propios a mejorar por falta de generalización. |
| Marketing | Cliente | Ver aspectos atractivos para promocionarlos y destacarlos para atraer clientes. | Que el modelo resalte aspectos particulares difíciles de usar para promociones. |
| Huéspedes | Beneficiado | Mejora en sus experiencias en estos hoteles. Mejores productos ofrecidos por parte de la cadena. | Que el modelo no tome en cuenta o no encuentre importancia en ciertas características que para un segmento de clientes si son factores decisivos. |

## Trabajo en equipo

Cómo equipo para el desarrollo de este proyecto nos dividimos de la siguiente manera:

* Daniel Felipe Vargas Ulloa - [d.vargasu@uniandes.edu.co](mailto:d.vargasu@uniandes.edu.co)
  + Rol: Lider de negocio
* Daniel Escalante Perez - [d.escalante@uniandes.edu.co](mailto:d.escalante@uniandes.edu.co)
  + Roles: Líder de datos y Líder de analítica
* Santiago Chamie Rey - [s.chamie@uniandes.edu.co](mailto:s.chamie@uniandes.edu.co)
  + Rol: Lider de proyecto

Basado en estos roles, las tareas las dividimos así:

* Daniel Felipe Vargas Ulloa - [d.vargasu@uniandes.edu.co](mailto:d.vargasu@uniandes.edu.co)
  + Tareas
    - Modelado y evaluación: Naive Bayes classifier [3 horas]
    - Entendimiento del negocio y enfoque analítico [2 horas]
    - Planeación del video [1.5 horas]
    - Segmento del video [3 horas]
* Daniel Escalante Perez - [d.escalante@uniandes.edu.co](mailto:d.escalante@uniandes.edu.co)
  + Tareas
    - Modelado y evaluación: Logistic Regression [4 horas]
    - Mapa de actores relacionado con el producto de datos creado [1 hora y media]
    - Resultados Texto [3 horas]
    - Segmento del video [2 horas]
* Santiago Chamie Rey - [s.chamie@uniandes.edu.co](mailto:s.chamie@uniandes.edu.co)
  + Tareas:
    - Entendimiento y preparación de los datos. [3 horas]
    - Modelado y evaluación: Árbol de decisión [3 horas]
    - Segmento del video [2 horas]

Nos reunimos en múltiples ocasiones para reflexionar sobre nuestro progreso en el proyecto, en los siguiente horarios y motivos:

* Reunión de lanzamiento y planeación [4:00 pm | Marzo 20, 2024] [2h]
* Reunión de ideación [3:00 pm | Marzo 26, 2024] [30 min]
* Reunión de seguimiento semana 1 [6:00 pm | Marzo 30, 2024] [1h]
* Reunión de seguimiento semana 2 [6:00 pm | Abril 4, 2024] [30 min]
* Reunión de finalización [6:00 pm | Abril 6, 2024] [30 min]

El reto principal planteado por este proyecto fue aprender a aplicar los conceptos vistos sobre algoritmos de clasificación, a un tipo de clasificación particular, que es la clasificación de textos. Para esto nos tocó investigar maneras de preparar los datos para funcionar dentro de un esquema de clasificación además de buscar algoritmos nuevos que se adaptaran mejor a este tipo de ejercicio. Otro reto principal fue encontrar un modelo que pudieramos recomendarle a la empresa utilizar. Usando los algoritmos vistos en clase como KNN y árboles de decisión los resultados de las métricas no fueron los mejores, por lo que implementamos otro tipo de algoritmos, como el de regresión logística, el cual se usa comunmente para este tipo de clasificación. Otro reto que apareció fue el manejo de la RAM disponible por Google Colab, ya que, al estar trabajando en esta plataforma, tenemos una cantidad limitada de RAM para utiliza, y gracias al tamaño de los datos en este proyecto esta RAM se acababa. Para solucionar esto, al finalizar cada etapa del desarrollo borrábamos las variables que no se iban a volver a utilizar, así abriendo espacio para el código posterior.

Dentro de los miembros del grupo repartiríamos los 100 puntos equitativamente entre todos ya que sentimos que todos contribuimos de manera significativa en el desarrollo del proyecto y sin la ayuda de todos lo hubiéramos logrado los resultados que obtuvimos. Por eso a cada uno le asignaríamos 33.33 o 100/3 puntos. No nos parecería justo darle a alguno más que al resto debido a que todos contribuimos por igual.

Nuestra dinámica de grupo ha sido bastante buena durante el desarrollo del proyecto al igual que los laboratorios por lo que no tenemos quejas al respecto. Sin embargo, los puntos a mejorar y trabajar de ahora en adelante como grupo son la sincronización y la tranquilidad. Con respecto a la sincronización, cómo Google Colab no permite trabajar a dos personas a la vez sin crear conflictos de versión, tuvimos momentos donde alguien estaba trabajando y alguien más llegaba y creaba conflictos. Esto se solucionaría trabajando en intervalos distintos o creando horarios de trabajo exclusivos. Con respecto a la tranquilidad, durante el desarrollo hubo varios momentos críticos donde el estrés a veces gana y lleva a tomar decisiones apuradas que pueden no ser las mejores.

## Referencias

Coursesteach. (2023, September 25). Deep learning (part 7)-logistic regression cost function. Medium. https://medium.com/@Coursesteach/deep-learning-part-7-6e78057a9ca6

Editorial Team. (2023, August 19). The 12 Main Hotel Departments and their key responsibilities. Indeed. https://sg.indeed.com/career-advice/career-development/hotel-departments

Euroinnova. (n.d.). Organigrama de un hotel: Departamentos y cargos principales. Euroinnova Business School. https://www.euroinnova.edu.es/articulos/escuelaiberoamericana/organigrama-de-un-hotel#direccion-general

Gunjal, S. (2020, June 1). Multiclass logistic regression using Sklearn. Kaggle. https://www.kaggle.com/code/satishgunjal/multiclass-logistic-regression-using-sklearn

Gusarov, M. (2022, April 9). Do I need to tune logistic regression hyperparameters?. Medium. https://medium.com/codex/do-i-need-to-tune-logistic-regression-hyperparameters-1cb2b81fca69

Jurafsky, D., & Martin, J. (2024, February 3). Logistic regression. Stanford. https://web.stanford.edu/~jurafsky/slp3/5.pdf

Kanade, V. (2022, April 18). Logistic regression: Equation, assumptions, types, and best practices. Spiceworks. https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/

Lorena. (2024, March 29). Chain Hotels O Cadena Hotelera: Características Principales. IHCS. https://www.ihcshotelconsulting.com/es/blog/chain-hotels-o-cadena-hotelera/

Martins, J. (2024, February 18). ¿Quiénes son los stakeholders de un proyecto?. Asana. https://asana.com/es/resources/project-stakeholder

Pramoditha, R. (2023, November 29). Logistic regression for multiclass classification - 3 strategies you need to know. Medium. https://rukshanpramoditha.medium.com/logistic-regression-for-multiclass-classification-3-strategies-you-need-to-know-0a3e74574b96

Scikit. (n.d.). Sklearn.linear\_model.Perceptron. Scikit Learn. <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Perceptron.html>

Weinberger, K. (n.d.). Cornell. Lecture 3: The Perceptron. <https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote03.html#:~:text=The%20Perceptron%20was%20arguably%20the,%2C%20it%20will%20loop%20forever>.)

IBM. (2023). What are naïve Bayes classifiers? <https://www.ibm.com/topics/naive-bayes>

Saritaş, M. M., & Yaşar, A. (2019). Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification. International Journal Of Intelligent Systems And Applications In Engineering, 7(2), 88-91. <https://doi.org/10.18201/ijisae.2019252786>