Análisis Comparativo de Alojamientos Airbnb: Turquía vs. Brasil :

Participates: Bruno Izaguirre, Oscar Lijeron, Raul Alvarez, Iker Marín

Índice general

Índice de figuras

Índice de cuadros

Acrónimos

- **LR**: Logistic Regression
- **XGB**: XGBoost
- **MNB**: Multinomial Naive Bayes
- **BoW**: Bag of Words
- **Tf-Idf**: Term frequency – Inverse document frequency

1. Tableau: Análisis de los Datos iniciales

Esta parte es la que también se representará en Tableau. Responsables de esta parte de la documentación; los lideres de Tableau.

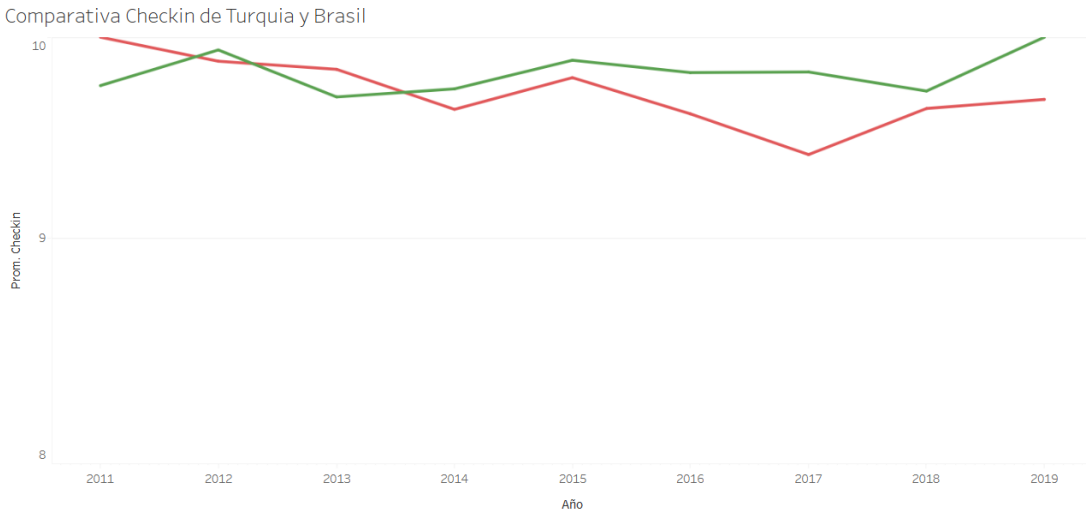
Presentarán una descripción de los datos a través de gráficos de Tableau. Por ejemplo sería interesante representar la siguiente información tanto para mi empresa como para mi competidor. Recordad que es importante elegir el gráfico adecuado en cada ocasión. Tenéis tendencia e representar la información a través de gráficos de quesitos y no suelen ser los mejores. Repasad las transparencias sobre visialización y los links con recomendaciones.

¿Cúal es mi zona y cúal mi competidor?

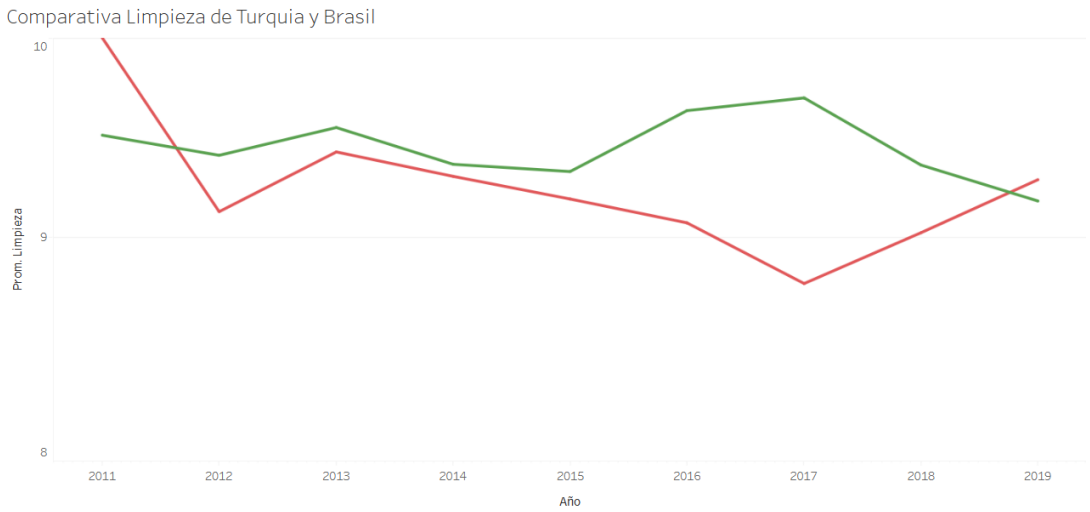
Nuestra zona es todo el territorio de Turquía y la zona de nuestros competidores sería Brasil.

¿Obtengo mejores o peores valoraciones que mi competidor?

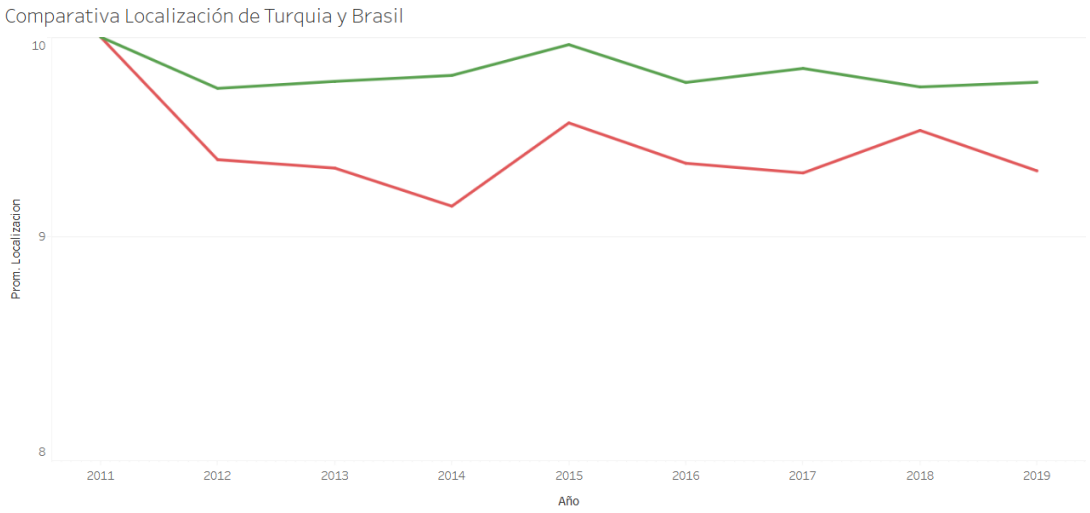
La media de las valoraciones según la muestra que hemos trabajado Turquía tendría una valoración de 91.6 sobre 100. En cambio, Brasil tiene una valoración de un 94.8 sobre 100. Es decir, Brasil gana respecto a Turquía por un 3.2 por ciento.



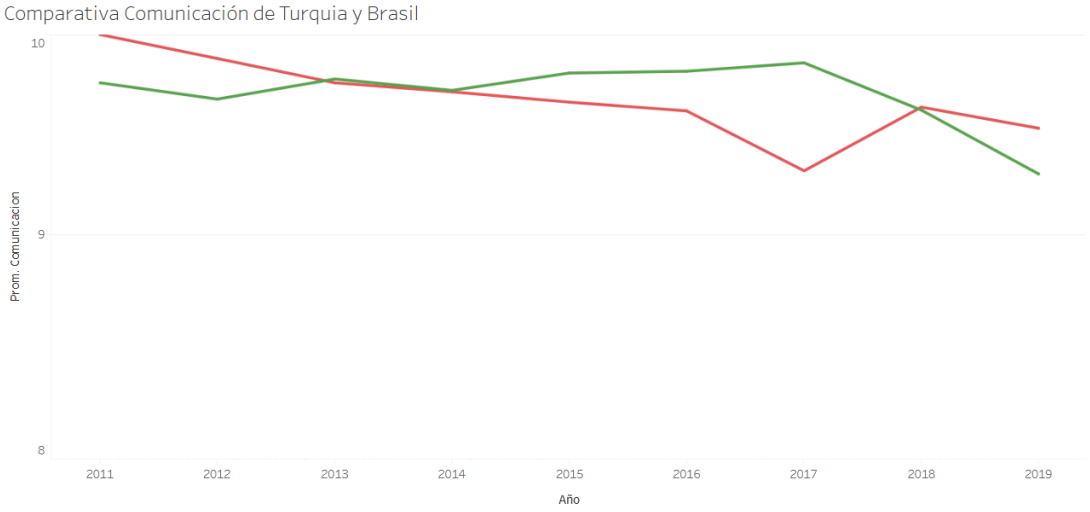
1.1. Figura: Comparativa de Checkin de Turquía y Brasil



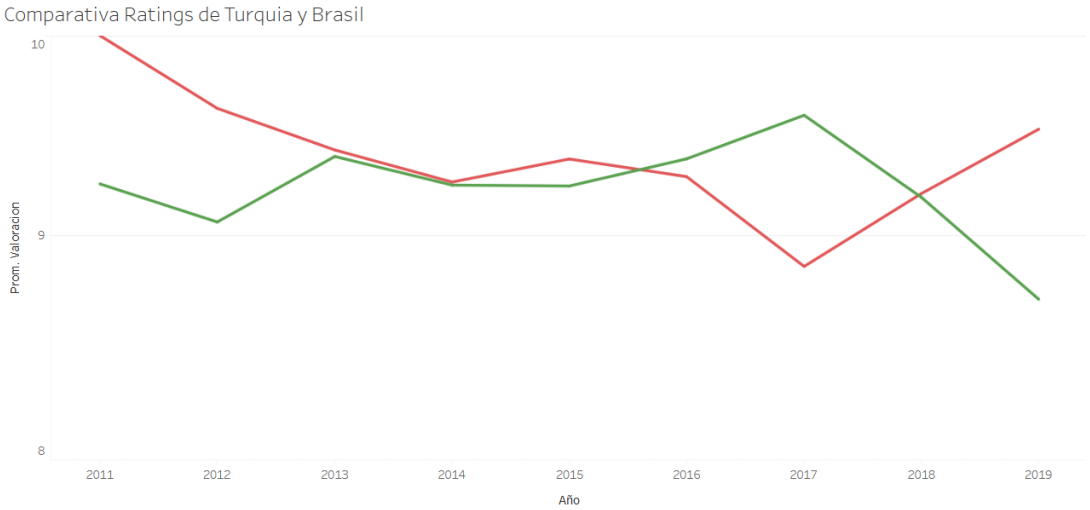
1.2. Figura: Comparativa de Limpieza de Turquía y Brasil



1.3. Figura: Comparativa de Localización de Turquía y Brasil



1.4. Figura: Comparativa de Comunicación de Turquía y Brasil



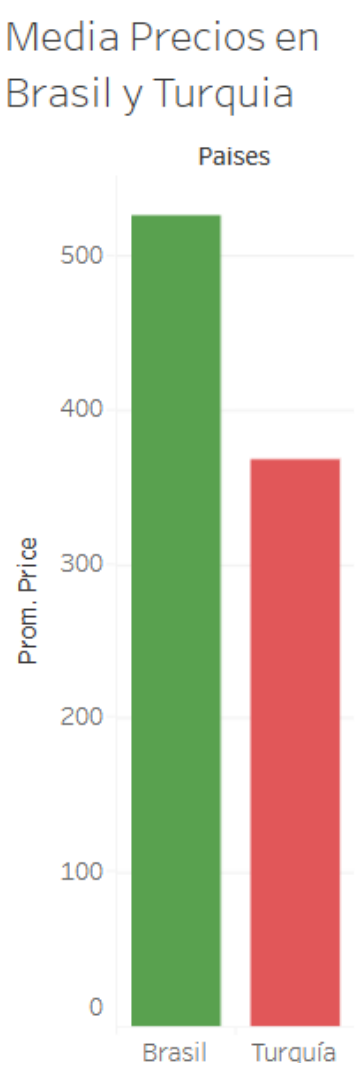
1.5. Figura: Comparativa de Ratings de Turquía y Brasil

¿Qué tipo de listings son más significativos de mi zona y de mi competidor?

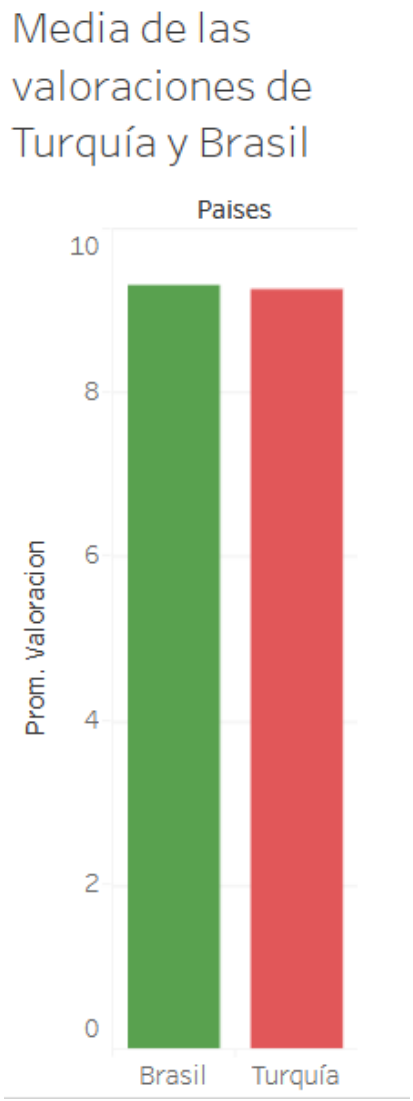
Los listings mas significativos serían los precios de los alquileres, sus valoraciones medias, asi como sus valoraciones respecto a la limpieza, la comunicación, la localización y el checkin. Otro interesante podría ser el tipo de propiedades de cada país.

¿Qué precios medios, que tipos de listings describen mejor a mi zona y a la de mi competidor?

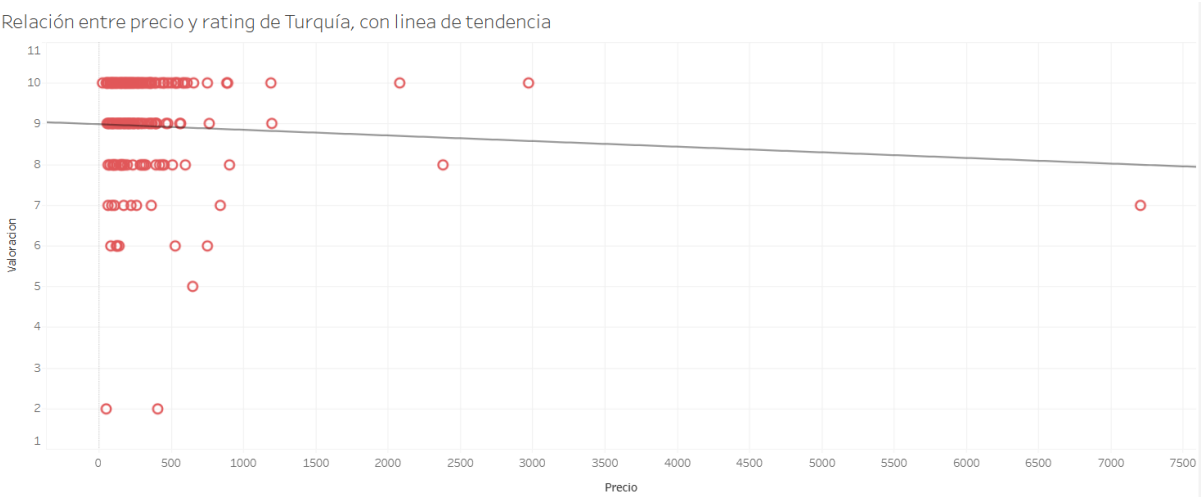
La media de los precios de Turquia es de 368 euros. En cambio, Brasil tiene una medio de precios de 525 euros. Comparando la media de las valoraciones y de los precios podemos observar Brasil (nuestro competidor) sale ganando respecto a Turquía (nuestra zona).



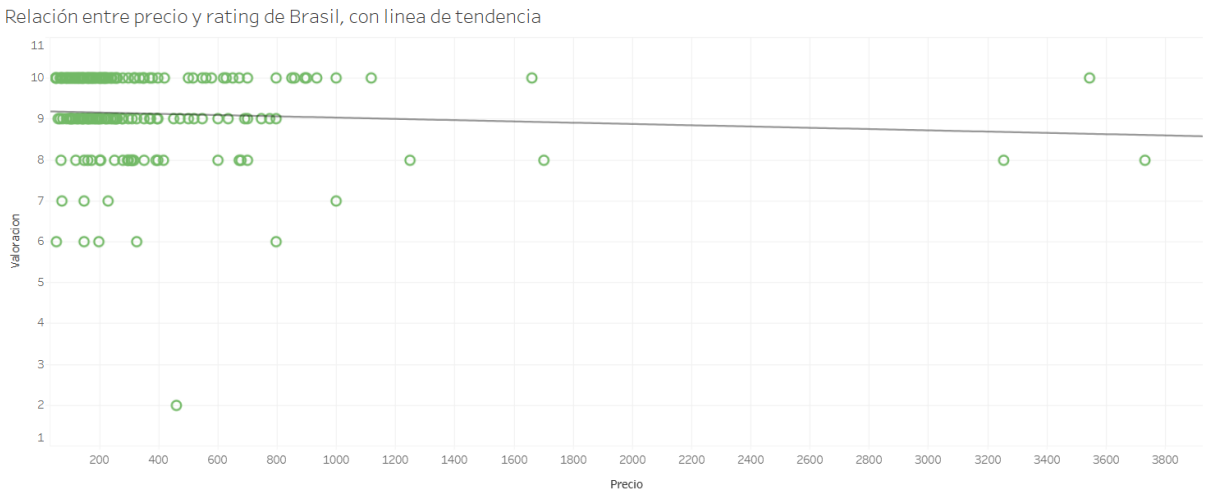
1.6. Figura: Media de precios



1.7. Figura: Media de Ratings



1.8. Figura: Relación entre los precios y ratings de Turquia



1.9. Figura: Relación entre los precios y ratings de Brasil

¿Puedo representar en un mapa mis listings y los de mi competidor y están más bien centricos o existe variedad?

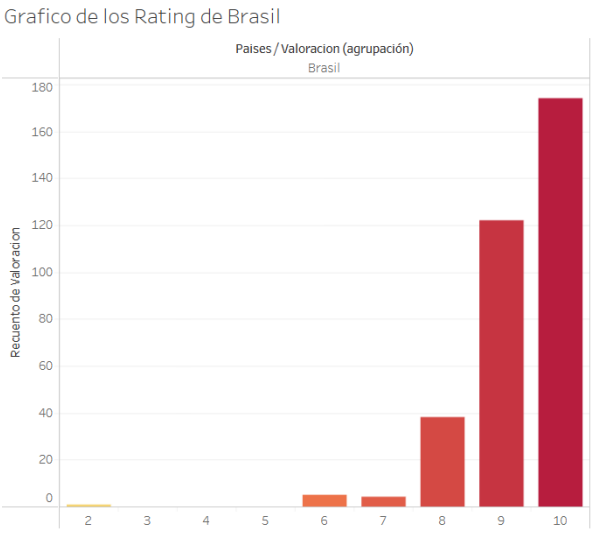
En general, no existe una gran variedad respecto a los listing excepto en los precios de los alquileres de Turquía, cuya media es inferior al de Brasil por amplia diferencia.

¿Se correlacionan los precios con los ratings?

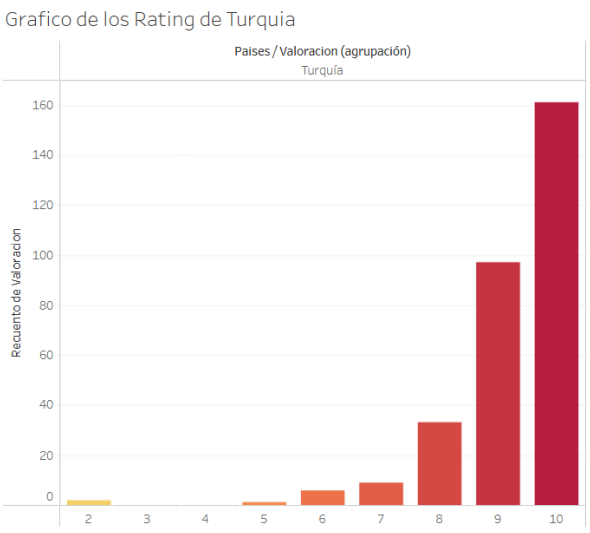
Podemos ver que no hay grandes diferencias en cuanto a valoraciones teniendo en cuenta el precio, aunque, de hecho, viendo las lineas de tendencias podemos entender que tanto en nuestra zona (Turquia) como en la de nuestro competidor (Brasil) cuanto más caros son los alquileres menos valoración tienen.

¿Se pueden representar en un mapa los listings y sus ratings por colores?

Si, y vemos que en ambos casos que el mayor numero de ratings estan entre 90-100.



1.10. Figura: Distribución de precios en Brasil



1.11. Figura: Distribución de precios en Turquía

¿Hay diferencias relevantes con respecto a la fecha?

Por lo general, la fecha no determina la calidad de los alquileres. Las valoraciones respceto a cada año no varia significativamente y oscila entre 9 y 10.

Estas preguntas son solo ejemplos de información que podría ser relevante. Añadid otras que tras analizar los datos os parezcan relevantes para contar vuestra historia

2. Clasificación de sentimientos: Análisis, Preproceso y Experimentación

Esta sección la desarrollarán entre todos los componentes del grupo.

Se puede encontrar el original de los datos en este link con los datos

2.1. Graphical Abstract de la solución

El Graphical abstract busca representar los módulos y el proceso seguido en vuestra experimentación así como las fuentes de datos empleadas (adicionalmente podéis incluir en el documento más gráficos explicativos y figuras adicionales para representar cada proceso de aprendizaje individualmente a lo largo del documento).

La figura 2.1 contiene un ejemplo de graphical abstract de un artículo de investigación para que podáis tomar de referencia de los que es un graphical abstract.

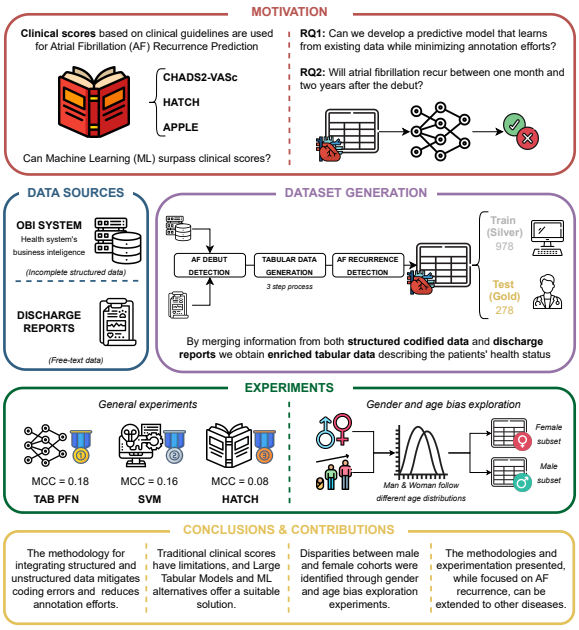
2.2. Datos

2.2.1. División entre Train Dev y Test de los datos para entrenar el modelo de predicción del ratings.

Recordad, esta tarea en una tarea de clasificación, es decir de aprendizaje supervisado. Para llevar a cabo esa tarea de aprendizaje supervisado entrenaréis clasificadores basados en los algoritmos que habéis aprendido en el curso a partir de los ficheros de AirBnBReviews y tripAdvisorHotelReviews donde cada review muestra la clase a la que pertenece. Ambos juegos de datos son distintos, uno permite realizar un aprendizaje binario y el otro un aprendizaje multiclase. Así pues realizaréis al menos dos conjuntos de experimentación. Generaréis un clasificador binario a partir de los datos de AirBnBReviews y generaréis un clasificador multiclase a partir de los datos de tripAdvisor. Vuestros clasificadores seguirán dos evaluaciones. Por un lado la evaluación local sobre el subconjunto dev asociado a AirBnBReviews (para el clasificador binario) y al dev tripAdvisorReviews (para el clasificador multiclase). De todos las combinaciones de hiperparámetros y algoritmos que hayáis realizado, pasaréis a evaluar sobre el dataset OBJETIVO o CENTRAL de vuestro estudio (esto es AirBnB.csv) y las combinaciónes que mejores resultados hayan arrojado sobre el dev y ¡CUIDADO! repito la evaluaréis sobre el dev de vuestros DATOS OBJETIVO que son los que se encuentran en el fichero AirBnB.csv. La media que obtenga vuestro algoritmo para los reviews del dev que hayáis generado deberá de coincidir lo mejor posible con la media que aparece en el fichero AirBnB.csv. Es importante tener en cuenta que las escalas no son equivalentes, así que uno de los trabajos a realizar es hacer compatibles ambas escalas.

Conjunto De Datos	% de instancias	Num. de instancias
Train	75	343
Dev	12,5	57
Test Final	12,5	58

2.1. Cuadro: División Train, Dev y Test de los datos de AirBnBReviews



2.1. Figura: Ejemplo de Graphical Abstract

Conjunto De Datos	% de instancias	Num. de instancias
Train	75	21093
Dev	12,5	3000
Test Final	12,5	3000

2.2. Cuadro: División Train, Dev y Test de los datos de TripadvisorHotelReviews

Conjunto De Datos	% de instancias	Num. de instancias
Dev	XXX	XXX
Test Final	XXX	XXX

2.3. Cuadro: Dev y Test de los datos centrales del estudio que son los contenidos en Airbnn.csv

2.2.2. Distribución de las clases en cada conjunto

AirBnBReviews.csv

Conjunto De Datos	Clase Neg	Clase Pos.
Train	176	167
Dev	23	34
Test Final	30	28

2.4. Cuadro: Distribución Train, Dev y Test de AirBnBReviews

tripAdvisor_hotel_reviews.cvs

Conjunto De Datos	Rating 1	Rating 2	Rating 3	Rating 4	Rating 5
Train	1057	1362	1643	4473	6833
Dev	182	218	255	756	1150
Test Final	182	213	286	810	1071

2.5. Cuadro: Distribución Train, Dev y Test de tripAdvisor

2.2.3. Descripción del preproceso

Dado que algunos de los reviews del fichero central airbnb.csv estaban en otras lenguas que no son el inglés lo primero que hemos hecho es traducir todos los reviews a inglés. Para ello hemos empleado Ollama y el modelo de lenguaje generativo Llama2 (o Llama3 si vuestro equipo os lo permite). El prompt que hemos empleado para hacer la traducción ha sido: XXXXX. Hemos seleccionado X reviews y hemos visto que la traducción coincide considerablemente con la traducción realizada por google translator.

Por otro lado, hemos convertido los textos en vectores empleando tf-idf, aunque para entrenar el modelo con el csv de tripAdvisor tuvimos que limitar el número de palabras máximas del tf-idf por problemas de rendimiento. Pasamos las palabras a minúsculas, quitamos stopwords y caracteres especiales y lematizamos.

2.2.4. Primeros resultados de la tarea de clasificación (Para el 29/04/2025)

Algoritmo	Hiperparámetros	Clase Weighted
Naive Bayes	alpha: 0.1, fit_prior: True	F1_Score: 0.979, Accuracy: 0.979, Recall: 0.979, Precision: 0.980
KNN	algorithm: auto, leaf_size: 20, n_neighbors: 1, p: 2, weights: uniform	F1_Score: 0.973, Accuracy: 0.973, Recall: 0.973, Precision: 0.975
RandomForest	bootstrap: True, criterion: entropy, max_depth: 10, max_features: log2, min_samples_leaf: 1, min_samples_split: 10, n_estimators: 50	F1_Score: 0.944, Accuracy: 0.962, Recall: 0.962, Precision: 0.962

2.6. Cuadro: Resultados sobre el conjunto de desarrollo de AirBnBReviews

tripAdvisor_hotel_reviews.cvs

Algoritmo	Hiperparámetros	Rating 1	Rating 2	Rating 3	Rating 4	Rating 5
Naive Bayes	alpha: 0.1, fit_prior: false	F1_Score: 0.88, Recall: 0.88, Precision: 0.88	F1_Score: 0.26, Recall: 0.30, Precision: 0.23	F1_Score: 0.30, Recall: 0.34, Precision: 0.32	F1_Score: 0.51, Recall: 0.51, Precision: 0.51	F1_Score: 0.77, Recall: 0.70, Precision: 0.74
KNN	algorithm: auto, leaf_size: 20, n_neighbors: 1, p: 2, weights: uniform	F1_Score: 0.88, Recall: 1, Precision: 0.79	F1_Score: 0.27, Recall: 0.14, Precision: 0.19	F1_Score: 0.20, Recall: 0.18, Precision: 0.22	F1_Score: 0.36, Recall: 0.36, Precision: 0.36	F1_Score: 0.61, Recall: 0.61, Precision: 0.60
RandomForest	bootstrap: True, criterion: Gini, max_depth: 10, max_features: sqrt, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 50	F1_Score: 0.85, Recall: 0.99, Precision: 0.74	F1_Score: 0, Recall: 0, Precision: 0	F1_Score: 0, Recall: 0, Precision: 0	F1_Score: 0.02, Recall: 0.01, Precision: 0.53	F1_Score: 0.68, Recall: 0.94, Precision: 0.54

2.7. Cuadro: Resultados sobre el Dev de los distintos algoritmos y alternativas tripAdvisor

2.2.5. Últimos resultados de la tarea de clasificación (Para el 20/05/2025)

Tendréis que aplicar vuestro ingenio para enfrentar el reto de aprender de unos datos y evaluar en otros. Por eso en el siguiente apartado (2.1.5) se os plantea presentar al menos 3 tablas:

Una de los resultados entrenando sobre el subset train de AirBnBReviews.csv (clasificación binaria) y evaluando sobre el dev y test del mismo tipo de dato (AirBnBReviews.csv).

Otra entrenando sobre el set de train de tripadvisor_hotel_reviews.csv (clasificación multiclase) y evaluando sobre el dev y test esos mismos tipos de datos (tripadvisor_hotel_reviews.csv).

Por último ver como funcionan los resultados a la hora de predecir los scores medios que aparecen en el fichero airbnb.cvs.

Daros cuenta de:

- Los datos que tenéis para aprender son o binarios (0,1 AirBnBReviews) o multiclase (1-5 tripadvisor_hotel_reviews.csv). Este tema va a ser clave en la contrucción de los clasificadores.
- Los datos finales sobre los que tenéis que evaluar tienen valores que van del 0-10.

Para esta última evaluación deberéis centraros en obtener la última tabla y mejorar las primeras y decidir la estrategia más apropiada para obtener la última tabla. AirBnBReviews.csv

Algoritmo	Hiperparámetros	Clase ??
Naive Bayes	XXX	XXX
KNN	XXX	XXX
RandomForest	XXX	XXX

2.8. Cuadro: Resultados sobre el Dev AirBnBReviews

tripAdvisor_hotel_reviews.cvs

Algoritmo	Hiperparámetros	Rating 1	Rating 2	Rating 3	Rating 4	Rating 5
Naive Bayes	XXX	XXX	XXX	XXX	XXX	XXX
KNN	XXX	XXX	XXX	XXX	XXX	XXX
RandomForest	XXX	XXX	XXX	XXX	XXX	XXX

2.9. Cuadro: Resultados sobre el Dev de los distintos algoritmos y alternativas tripAdvisor

tripAdvisor_hotel_reviews.cvs

Algoritmo	Hiperparámetros	Rating medio predicho	Rating medio real	Desviacion
????	XXX	XXX	XXX	XXX

2.10. Cuadro: Resultados sobre el dev y el test de airbnb final. Desviación=$(predicho - medio)^2$

2.2.6. Descripción del Proceso de Submuestreo o Sobremuestreo

Dado el caracter desbalanceado de las opiniones (positivas, negativas, neutras) se ha probado realizar Over y Undersampling para ello se ha empleado la técnica Oversampling implementada en la librería imblearn.over_sampling, en concreto RandomOverSampler

2.2.7. Resultados de la aplicación de los modelos generativos (Para el 20/05/2025)

(Añadir la descripción de la estratégia empleada: zero-shot, one-shot, few-shot y el prompt así como el modelo que se ha empleado)

Aquí no habrá entrenamiento de un clasificador empleando AirBnBReviews.csv o tripadvisor_hotel_reviews.csv. La predicción se hará sobre los datos finales, probando distintos prompts, estrategias y modelos sobre la muestra dev para finalmente aplicar la mejor combinación sobre el test. Tiene sentido aquí plantear la Desviación o convertiréis las puntuaciones en positivo (por ejemplo para valores ¿= 8.5), neutros (por ejemplo para valores ¿=8 y ¡8.5) y negativos (para valores ¡8). Esto dependerá mucho de vuestros datos y no se puede saber con antelación. Requiere que lo penseis, lo analicéis detenidamente y me expliquéis vuestra decisión.

Algoritmo	Hiperparám.	¿Rating medio predicho?	¿Rating medio real?	¿Desviacion?
????	XXX	XXX	XXX	XXX

2.11. Cuadro: Resultados sobre el dev y el test de airbnb final. Desviación=$(predicho - medio)^2$

2.3. Algoritmos, link a la documentación y nombre de los hiperparámetros empleados

2.3.1. Experimentación: Algoritmos empleados y Breve Descripción

Hemos empleado los siguientes algoritmos con los siguientes hiper-parámetros.

- **Multinomial Naive Bayes**:
  - Hiperparámetros: aplha y fit_prior
  - Link: Sklearn MultinomialNB
- **Random Forest**:
  - Hiperparámetros: n_estimators, criterion, max_depth, min_samples_split, min_samples_leaf, max_features y bootstrap.
  - Link:
- **kNN**:
  - Hiperparámetros: n_neighbors, weights, algorithm, leaf_size y p.
  - Link:
- **Modelos Generativos (p.e. Llama)**:
  - Hiperparámetros:
  - Link:

2.3.2. Resultados sobre el Development

En esta sección se presentan los resultados obtenidos para el development.

2.3.2.1. Optimizando los resultados de la clase negativa

Algoritmo	Combinación hyperparámetros	Prec	Rec	F-sco
XXX	XXX	XXX	XXX	XXX

2.12. Cuadro: Resultados Clase Negativa

2.3.2.2. Optimizando los resultados de la clase positiva

Algoritmo	Combinación hyperparámetros	Prec	Rec	F-sco
XXX	XXX	XXX	XXX	XXX

2.13. Cuadro: Resultados Clase Positiva

2.3.2.3. Sin optimizar ninguna clase en particular

Algoritmo	Combinación hyperparámetros	Prec	Rec	F-sco
XXX	XXX	XXX	XXX	XXX

2.14. Cuadro: Resultados generales

2.3.3. Discusión sobre el proceso de aprendizaje

La combinación XXXX obtiene mejores resultados para XXX pero peores para XXXX. La razón puede ser XXXX.

Bla bla bla

Se han presentado los siguientes problemas con el algoritmo XXXX, que se intentó probar pero sucedió XXXXX y no pudo acabarse la prueba.

2.3.4. Conclusión sobre la tarea de clasificación

Para la optimización de los resultados sobre la clase negativa se ha seleccionado finalmente se ha seleccionado XXXXX por XXXXX.

Para la optimización de los resultados generales se ha seleccionado finalmente se ha seleccionado XXXXX por XXXXX.

2.4. Algoritmos, link a la documentación y nombre de los hiperparámetros empleados

2.4.1. Experimentación: Algoritmos empleados y Breve Descripción

Creemos que deberiamos usar el algortimo naive bayes para predecir por tener mejores resultados que el kNN y el random forest tanto en el barrido de hiperpárametros como en el test. Puede que naive bayes sea mejor para el sentimental análisis porque es eficiente y rápido con datasets de texto grandes y permite mas cantidad de palabras del tf-idf a tomar en cuenta,y porque funciona bien con datos desbalanceados gracias al suavizado.

2.4.2. Resultados

Mostrar gráficamente la distribución de las palabras más significativas de cada rating en la vuestra zona y en la competidora ¿Son similares o referencian los atributos de los datos iniciales como por ejemplo *localización, limpieza*, etc? ¿Habéis descubierto razones más allá de los atributos mencionados?

Mostrar gráficamente la distribución de los conceptos significativos.

2.4.3. Discusión sobre los descubrimientos realizados

¿Qué habéis descubierto con respecto a vosotros y vuestros competidores?¿Cúales son nuestras fortalezas y nuestros puntos débiles?¿Y los de nuestros competidores?

Bibliografía