

---

Título:

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

# Índice general

|  |           |
|--|-----------|
| <b>1. Tableau: Análisis de los Datos iniciales</b>   | <b>6</b>  |
| <b>2. Clasificación de sentimientos: Análisis, Preproceso y Experimentación</b>                                | <b>10</b> |
| 2.1. Graphical Abstract de la solución . . . . .   | 10        |
| 2.2. Datos . . . . .   | 10        |
| 2.2.1. División entre Train Dev y Test de los datos para entrenar el modelo de predicción del ratings. . . . . | 10        |
| 2.2.2. Distribución de las clases en cada conjunto . . . . .   | 11        |
| 2.2.3. Descripción del preproceso . . . . .  | 12        |
| 2.2.4. Primeros resultados de la tarea de clasificación (Para el 29/04/2025) . . . . .                         | 12        |
| 2.2.5. Últimos resultados de la tarea de clasificación (Para el 20/05/2025) . . . . .                          | 13        |
| 2.2.6. Descripción del Proceso de Submuestreo o Sobremuestreo . . . . .  | 14        |
| 2.2.7. Resultados de la aplicación de los modelos generativos (Para el 20/05/2025) . . . .                     | 14        |
| 2.3. Algoritmos, link a la documentación y nombre de los hiperparámetros empleados . . . .                     | 14        |
| 2.3.1. Experimentación: Algoritmos empleados y Breve Descripción . . . . .                                     | 14        |
| 2.3.2. Resultados sobre el Development . . . . .   | 15        |
| 2.3.2.1. Optimizando los resultados de la clase negativa . . . . .   | 15        |
| 2.3.2.2. Optimizando los resultados de la clase positiva . . . . .   | 15        |
| 2.3.2.3. Sin optimizar ninguna clase en particular . . . . .   | 15        |
| 2.3.3. Discusión sobre el proceso de aprendizaje . . . . .   | 15        |
| 2.3.4. Conclusión sobre la tarea de clasificación . . . . .  | 15        |
| 2.4. Algoritmos, link a la documentación y nombre de los hiperparámetros empleados . . . .                     | 15        |
| 2.4.1. Experimentación: Algoritmos empleados y Breve Descripción . . . . .                                     | 15        |
| 2.4.2. Resultados . . . . .  | 15        |
| 2.4.3. Discusión sobre los descubrimientos realizados . . . . .  | 16        |
| <b>Bibliografía</b>  | <b>16</b> |

# Índice de figuras

|   |    |
|---|----|
| 1.1. Mediana de valoraciones comparando con el competidor . . . . . | 6  |
| 1.2. Tipos de propiedades en Turquía y su frecuencia . . . . .      | 7  |
| 1.3. Tipos de propiedades en Brasil y su frecuencia . . . . .       | 7  |
| 1.4. Mediana de precios comparando con el competidor . . . . .      | 8  |
| 1.5. Distribución de precios en Brasil . . . . .                    | 9  |
| 1.6. Distribución de precios en Turquía . . . . .                   | 9  |
| 2.1. Ejemplo de Graphical Abstract . . . . .                        | 11 |

# Índice de cuadros

|   |    |
|---|----|
| 2.1. División Train, Dev y Test de los datos de AirBnBReviews . . . . .                               | 10 |
| 2.2. División Train, Dev y Test de los datos de TripadvisorHotelReviews . . . . .                     | 11 |
| 2.3. Dev y Test de los datos centrales del estudio que son los contenidos en Airbnn.csv . . . . .     | 11 |
| 2.4. Distribución Train, Dev y Test de AirBnBReviews . . . . .  | 11 |
| 2.5. Distribución Train, Dev y Test de tripAdvisor . . . . .  | 11 |
| 2.6. Resultados sobre el conjunto de desarrollo de AirBnBReviews . . . . .                            | 12 |
| 2.7. Resultados sobre el Dev de los distintos algoritmos y alternativas tripAdvisor . . . . .         | 13 |
| 2.8. Resultados sobre el Dev AirBnBReviews . . . . .  | 13 |
| 2.9. Resultados sobre el Dev de los distintos algoritmos y alternativas tripAdvisor . . . . .         | 14 |
| 2.10. Resultados sobre el dev y el test de airbnb final. Desviación= $(predicho - medio)^2$ . . . . . | 14 |
| 2.11. Resultados sobre el dev y el test de airbnb final. Desviación= $(predicho - medio)^2$ . . . . . | 14 |
| 2.12. Resultados Clase Negativa . . . . .   | 15 |
| 2.13. Resultados Clase Positiva . . . . .   | 15 |
| 2.14. Resultados generales . . . . .  | 15 |

# 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

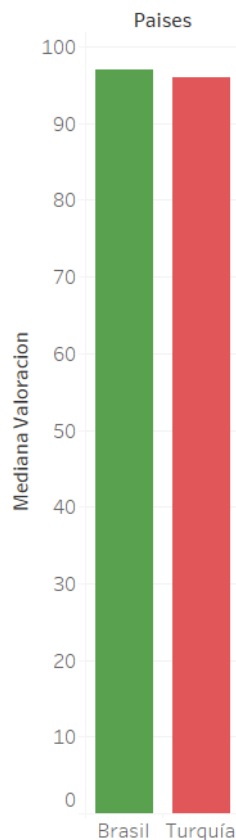
Esta parte es la que también se representará en Tableau. Responsables de esta parte de la documentación; los líderes de Tableau.

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 a representar la información a través de gráficos de quesitos y no suelen ser los mejores. Repasad las transparencias sobre visualización y los links con recomendaciones.

¿Cuál es mi zona y cuál mi competidor?

¿Obtengo mejores o peores valoraciones que mi competidor?

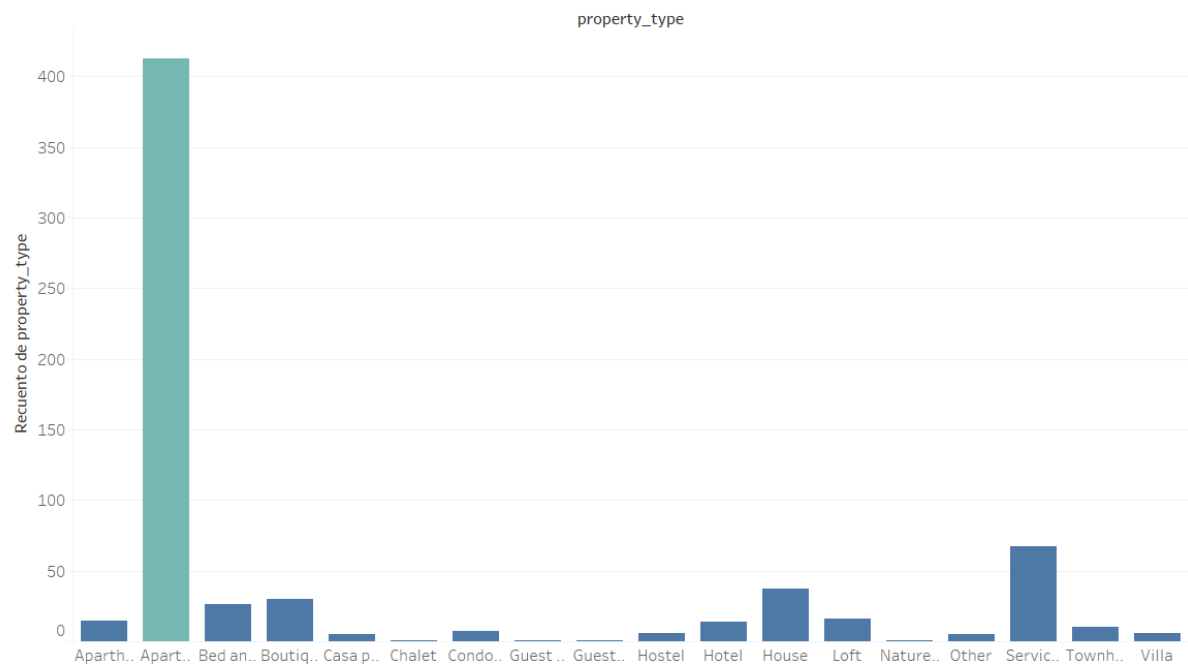
Mediana de las valoraciones de Turquía y Brasil



1.1. Figura: Mediana de valoraciones comparando con el competidor

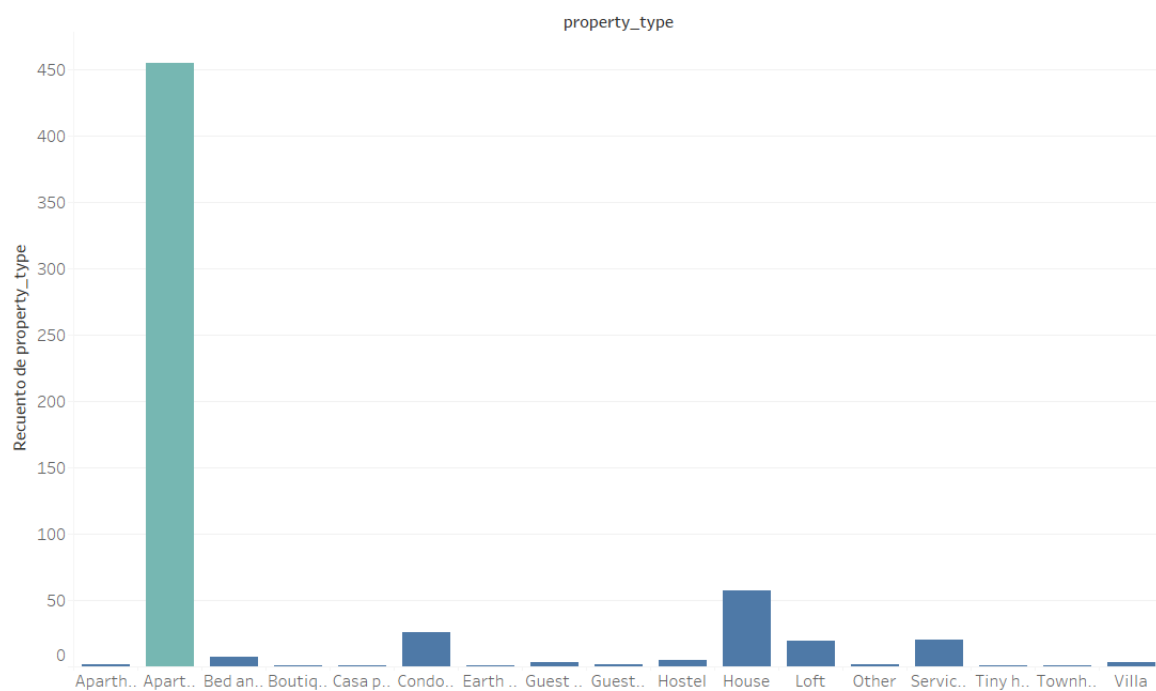
¿Qué tipo de listings son más significativos de mi zona y de mi competidor?

## Tipos de propiedades de Turquía



1.2. Figura: Tipos de propiedades en Turquía y su frecuencia

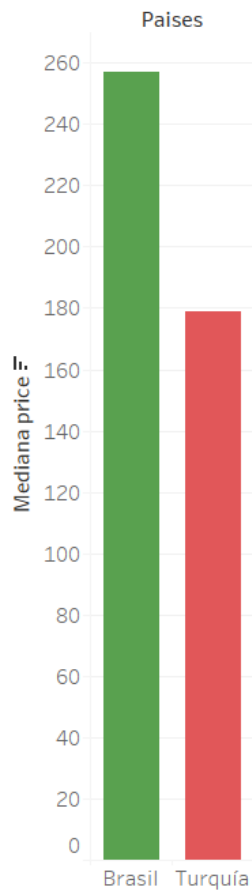
## Tipos de propiedades de Brasil



1.3. Figura: Tipos de propiedades en Brasil y su frecuencia

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

## Mediana de los alquileres de Turquía y Brasil



1.4. Figura: Mediana de precios comparando con el competidor

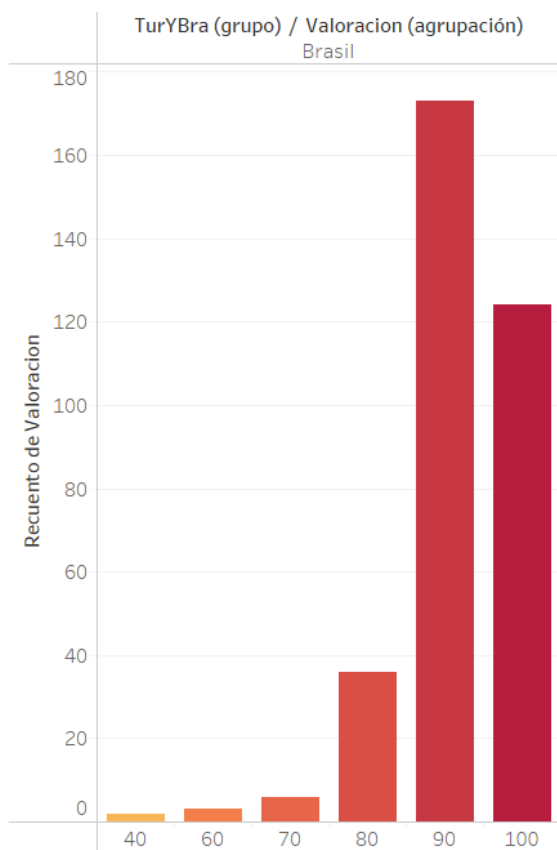
¿Puedo representar en un mapa mis listings y los de mi competidor y están más bien centricos o existe variedad?

¿Se correlacionan los precios con los ratings?

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

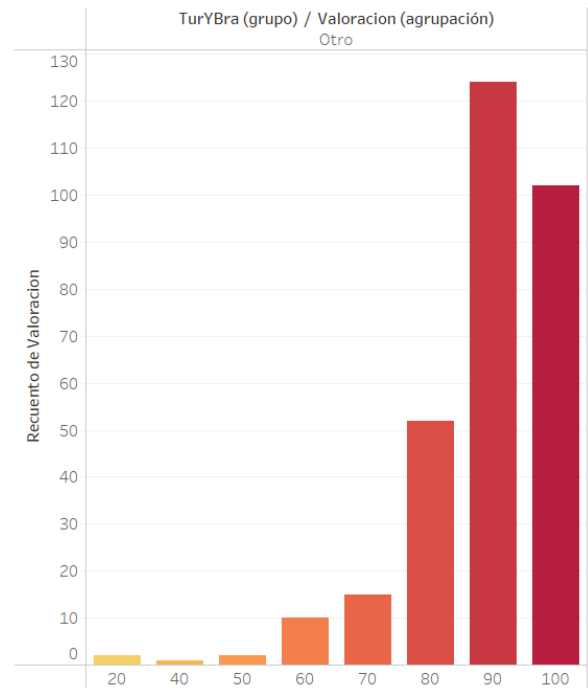


### Grafico de los rating de Brasil



1.5. Figura: Distribución de precios en Brasil

### Grafico de los rating de Turquia



1.6. Figura: Distribución de precios en Turquía

¿Hay diferencias relevantes con respecto a la fecha?

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, Pre-proceso 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.

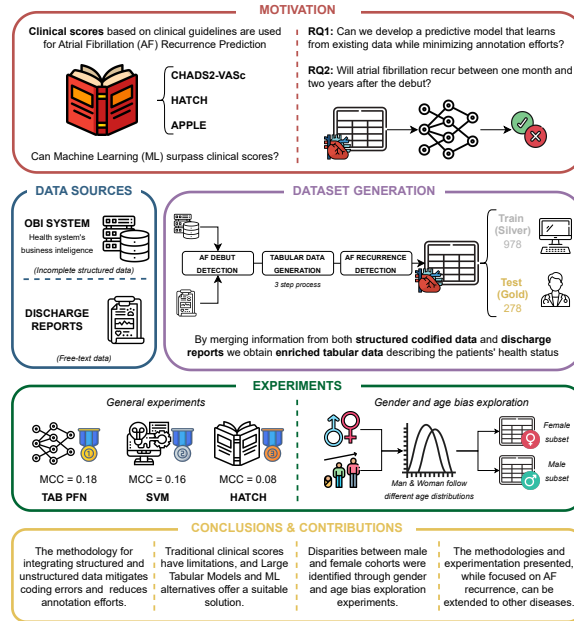
### 2.2. Datos

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

Recordad, esta tarea es 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 todas 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 combinaciones 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.csv

| 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 **XXX**.....

**Añadid aquí la descripción de todos los preprocesos que hayáis realizado.**

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

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.csv`.

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 este primera evaluación deberéis centraros en las dos primeras. `AirBnBReviews.csv`

| 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: False, criterion: entropy, max_depth: 10, max_features: log2, min_samples_leaf: 1, min_samples_split: 5, n_estimators: 50 | F1_Score: 0.944, Accuracy: 0.944, Recall: 0.944, Precision: 0.947 |

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

`tripAdvisor_hotel_reviews.csv`

| 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: 0.73, Recall: 0.70, Precision: 0.77 | F1_Score: 0.69, Recall: 0.67, Precision: 0.70 | F1_Score: 0.75, Recall: 0.73, Precision: 0.77 | F1_Score: 0.83, Recall: 0.83, Precision: 0.82 |
| RandomForest | bootstrap: False, criterion: entropy, 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: 0.54 | 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  | alpha: 0.1, fit_prior: True  | XXX      |
| KNN          | algorithm: auto, leaf_size: 20, n_neighbors: 1, p: 2, weights: uniform   | XXX      |
| RandomForest | bootstrap: False, criterion: entropy, max_depth: 10, max_features: log2, min_samples_leaf: 1, min_samples_split: 5, n_estimators: 50 | 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.csv

| 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 estrategia 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  $\hat{z}=8.5$ ), neutros (por ejemplo para valores  $\hat{z}=8$  y  $\hat{z}=8.5$ ) y negativos (para valores  $\hat{z}=8$ ). 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 (no necesariamente tienen que ser estos sino los que empleeis).

#### ■ Multinomial Naive Bayes:

- Hiperparámetros: alpha 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 hiperparámetros | 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 hiperparámetros | 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 hiperparámetros | 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

#### 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? ¿Cuáles son nuestras fortalezas y nuestros puntos débiles? ¿Y los de nuestros competidores?



## Bibliografía