APRENENTATGE AUTOMÀTIC APLICAT A LA VALORACIÓ D'ALLOTJAMENTS D'AIRBNB

Joan Orteu Saiz

Treball Final de Grau, Grau d'Enginyeria Informàtica

Director: Dr. Santi Seguí

25 de febrer de 2024

Continguts

Continguts

Introducció

Introducció

Definició de Problema

Quines millores es poden realitzar a un allotjament per a augmentar la valoració i quin serà el guany resultant?

Introducció

Definició de Problema

Quines millores es poden realitzar a un allotjament per a augmentar la valoració i quin serà el guany resultant?

Solució Actual

- AirDNA com a *Channel Manager*.
- Llegir els comentaris i estudiar la competencia.

Introducció

Definició de Problema

Quines millores es poden realitzar a un allotjament per a augmentar la valoració i quin serà el guany resultant?

Solució Actual

- AirDNA com a *Channel Manager*.
- Llegir els comentaris i estudiar la competencia.

Solució Plantejada

- Aplicar Aprenentatge Automàtic per a la Predicció de Valoració d'Allotjaments.
- Aplicar tècniques d'XAI com SHAP i LIME per entendre quina és la importància de cada característica.
- Crear una pàgina web per visualitzar els resultats.

Introducció

Motivació

- Aprendre a fer projectes d'aprenentatge automàtic.
- Passió pel sector turístic.

Introducció

Motivació

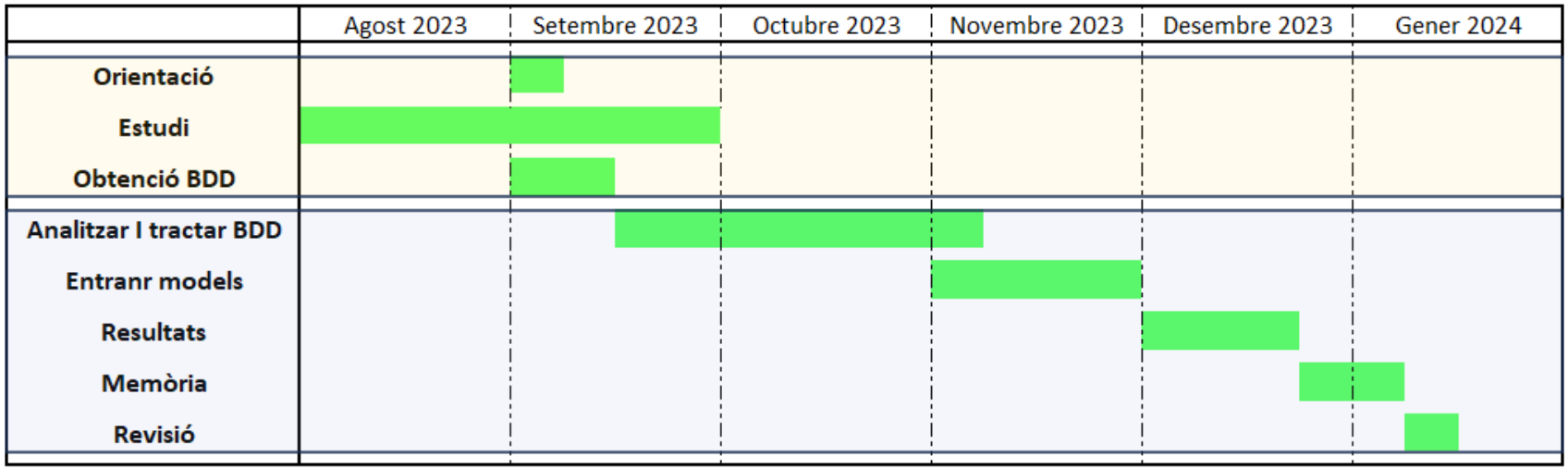
- Aprendre a fer projectes d'aprenentatge automàtic.
- Passió pel sector turístic.

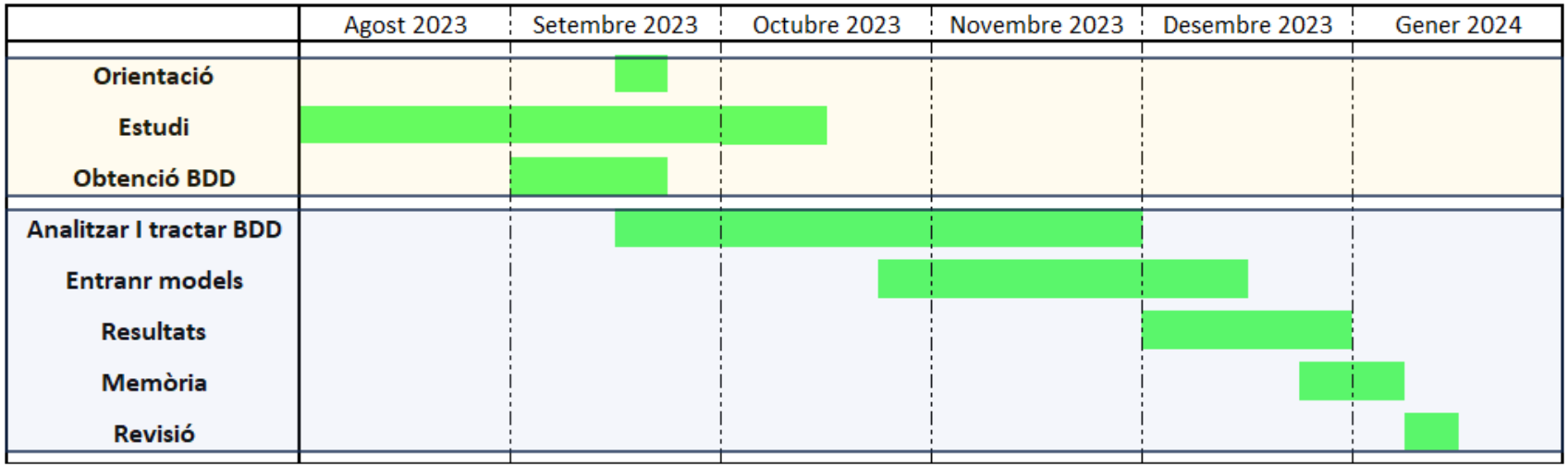
Objectius

- Adquirir i estructurar una base de dades rellevant.
- Realitzar el preprocessament de les dades.
- Desenvolupar un model predictiu de valoracions.

Continguts

Planificació





Continguts

Base de Dades

Obtenció de Dades

Font: **Insideairbnb**.

Contingut: consta de tres taules sobre allotjaments de Barcelona:

- Calendar
- Listings
- Reviews

Atributs Objectiu

- Review_score_rating
- Review_score_accuracy
- Review_score_cleanliness
- Review_score_checkin
- Review_score_communication
- Review_score_location
- Review_score_value

Anàlisi Descriptiu

Tipus de Dades

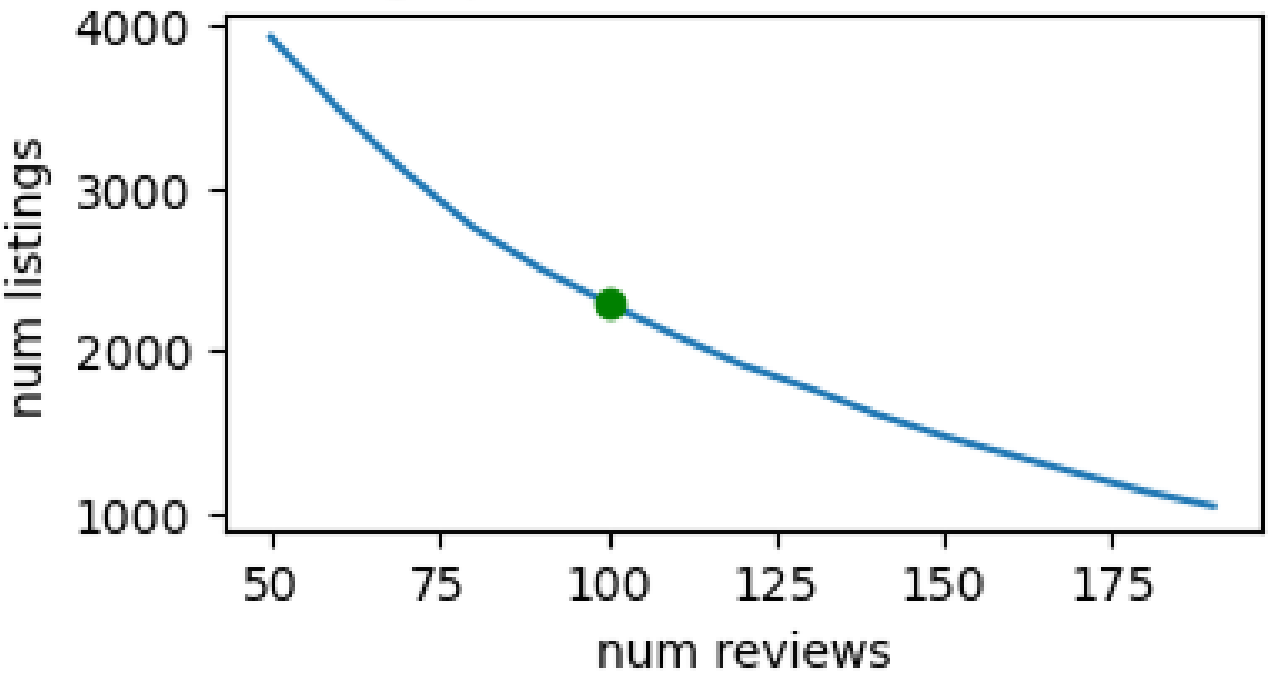
Aquesta és una base de dades tabulades, de tipus:

- Text
- Numèric
- Categòriques
- Dates
- Llistats

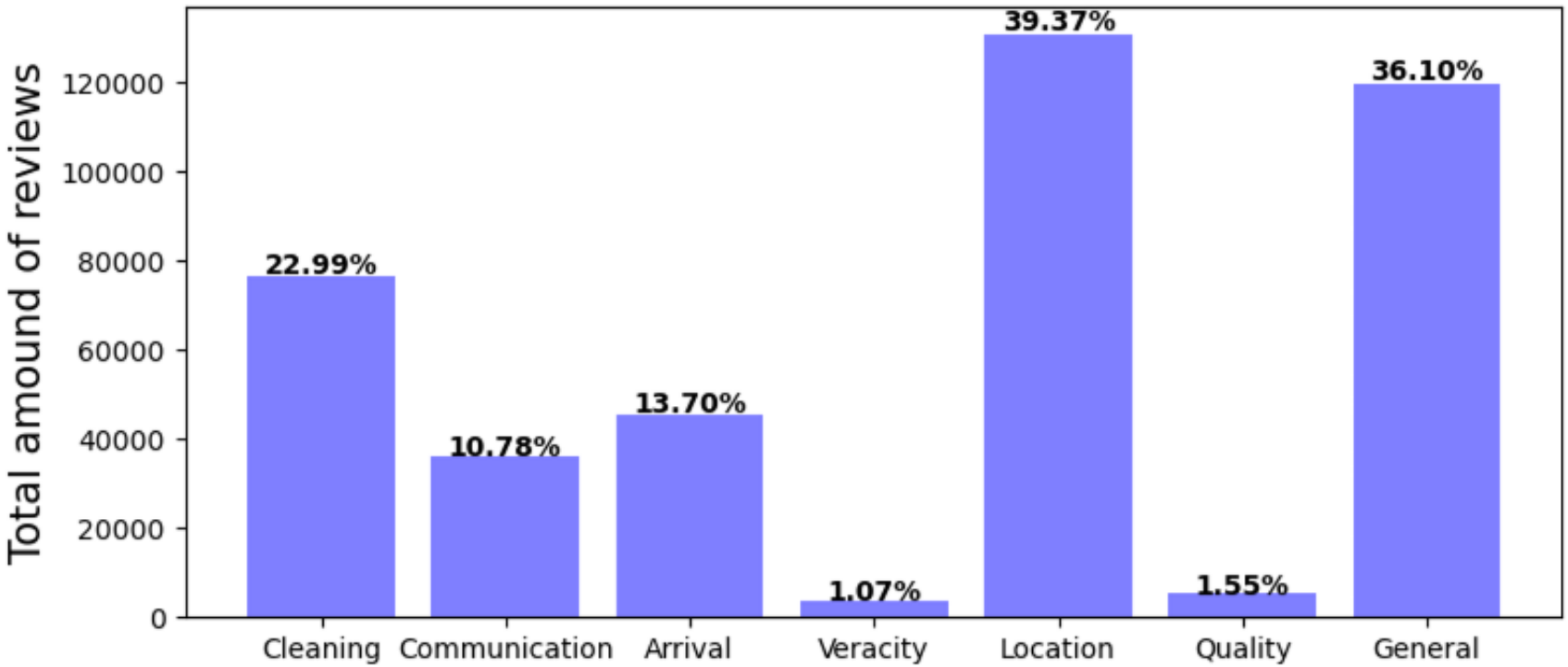
Proves Relitzades

- Número de comentaris per allotjament
- Rellevància dels comentaris
- Correlació entre atributs objectiu
- Valoració de localització

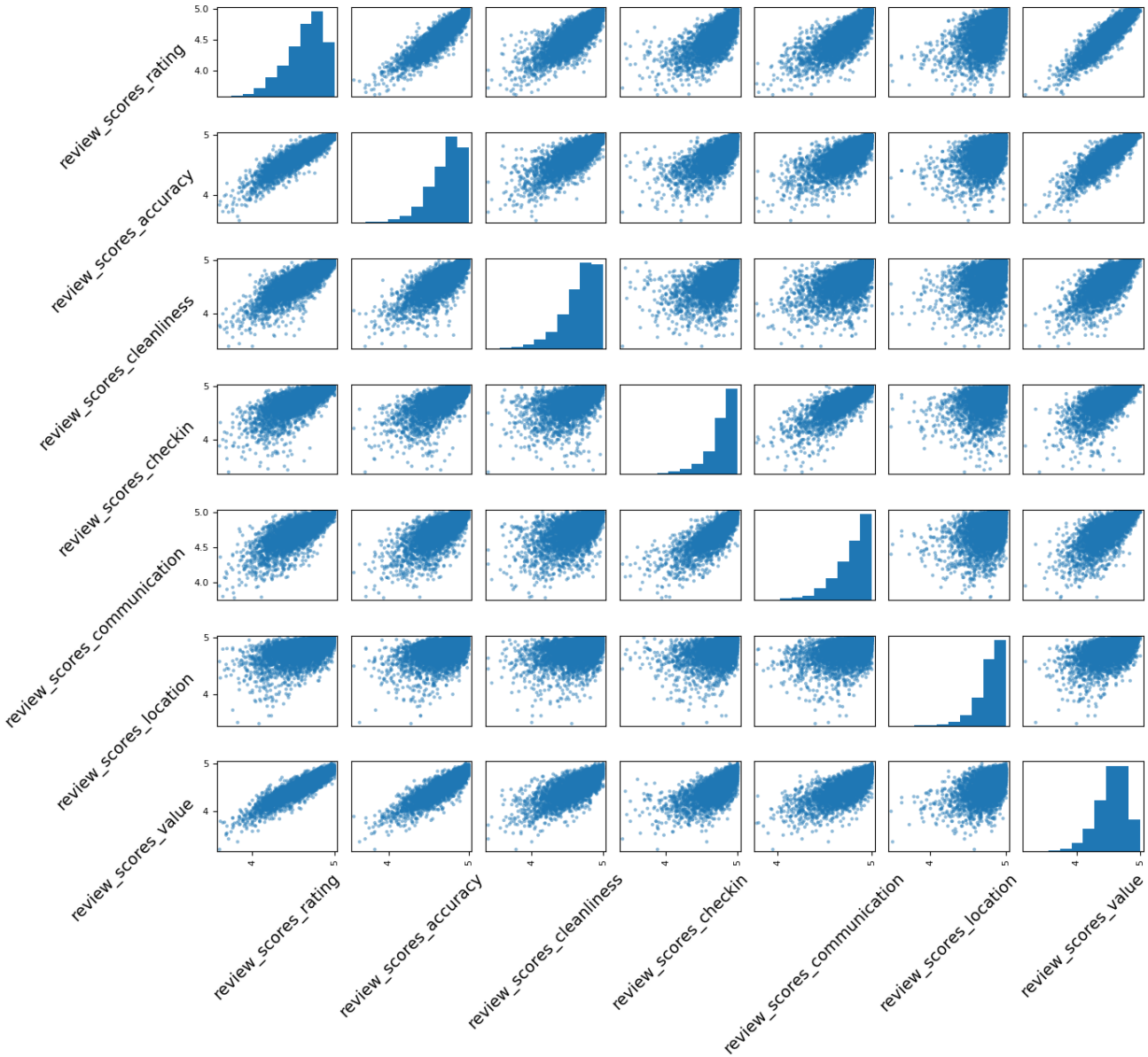
Número de Comentaris per Allotjament



Rellevància dels Comentaris



Correlació Entre Atributs Objectiu

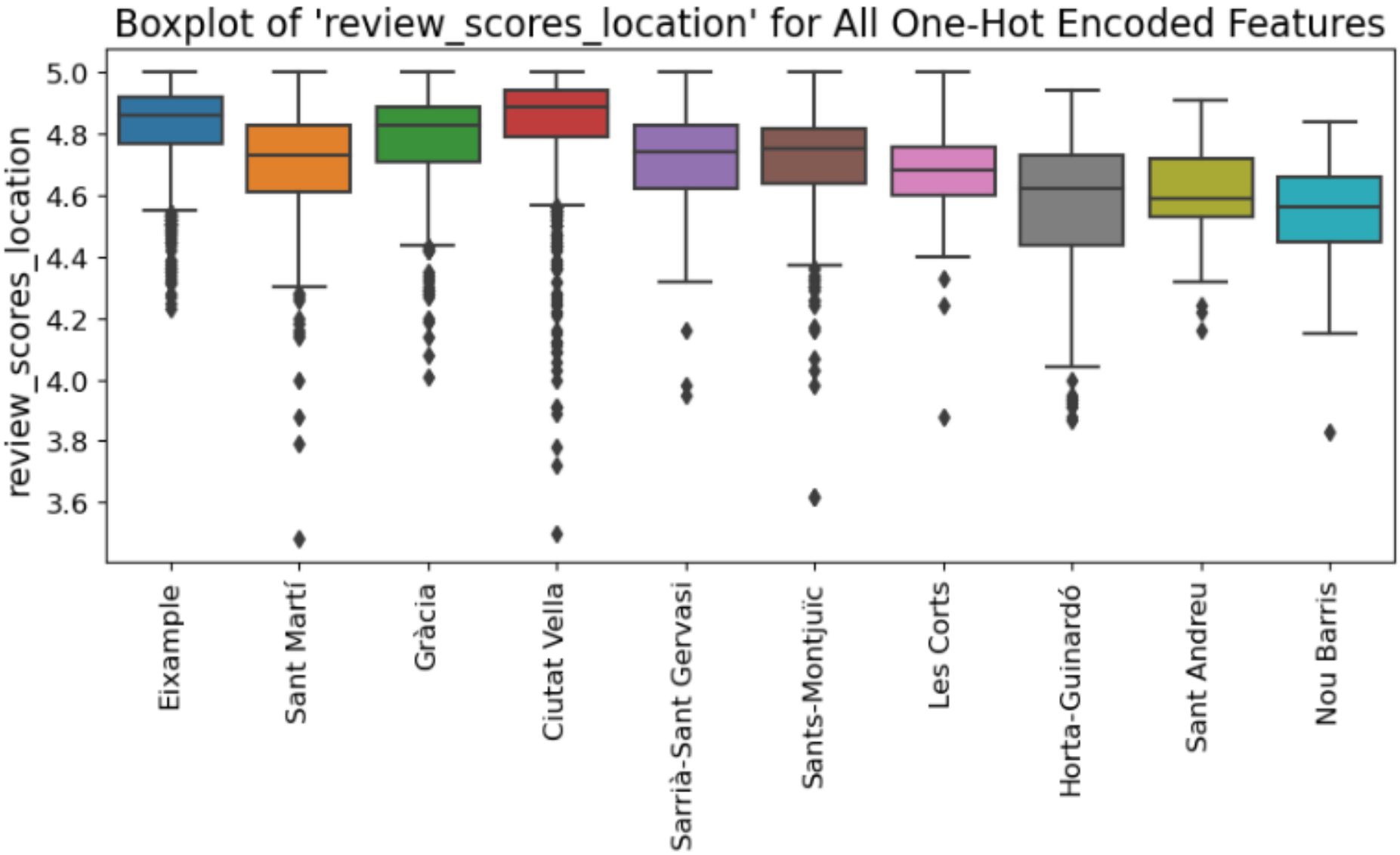


Correlació Entre Atributs Objectiu

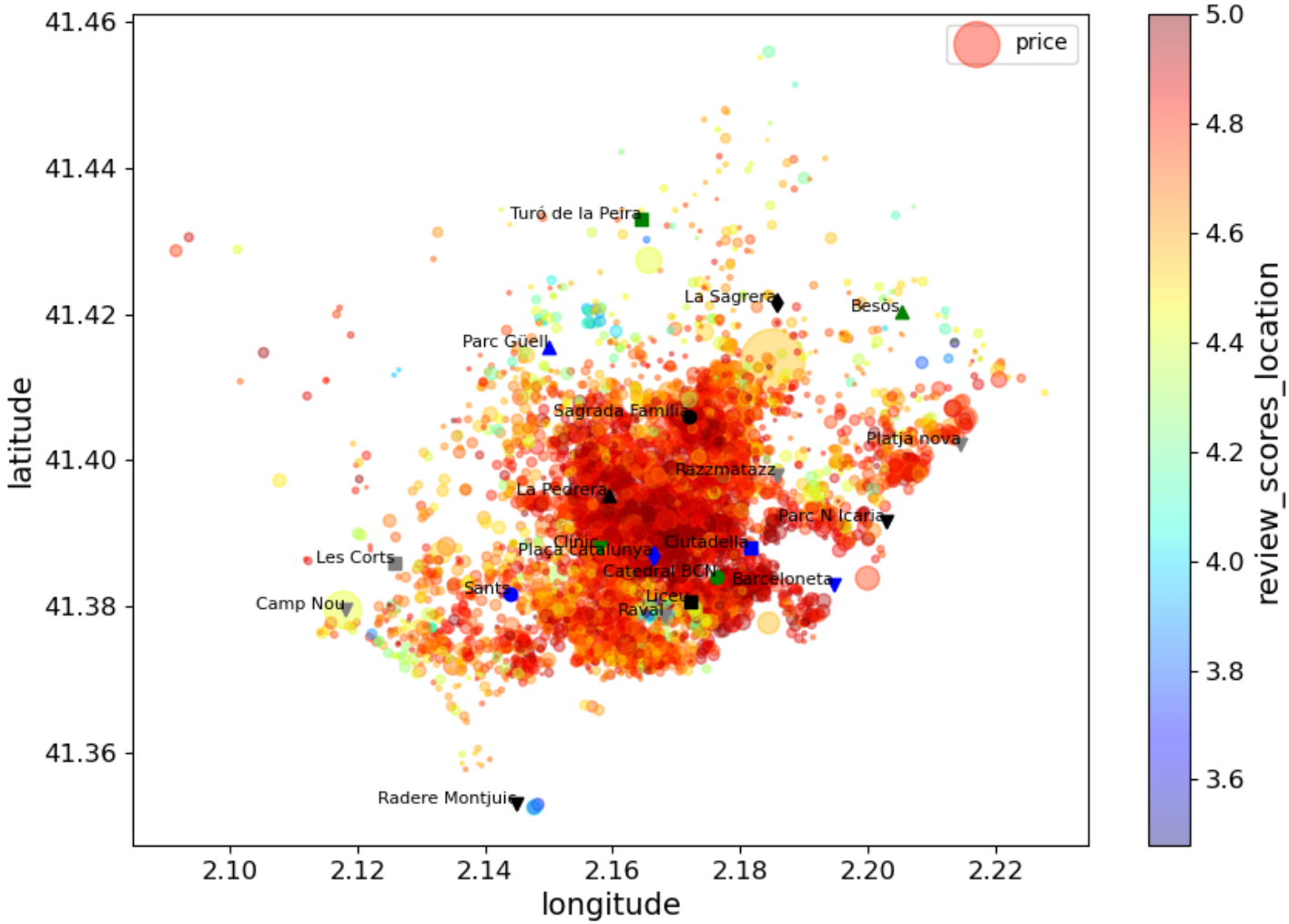
Variable	Corr
host_is_superhost	0.447795
Extra pillows and blankets	0.270498
soap	0.227851
parking	0.217668
First aid kit	0.214844
Room-darkening shades	0.214027
Carbon monoxide alarm	0.208656
number_of_reviews	0.204902
books	0.202167
Host greets you	0.194115

- Indica modificacions que el host pot fer a l'allotjament de forma fàcil

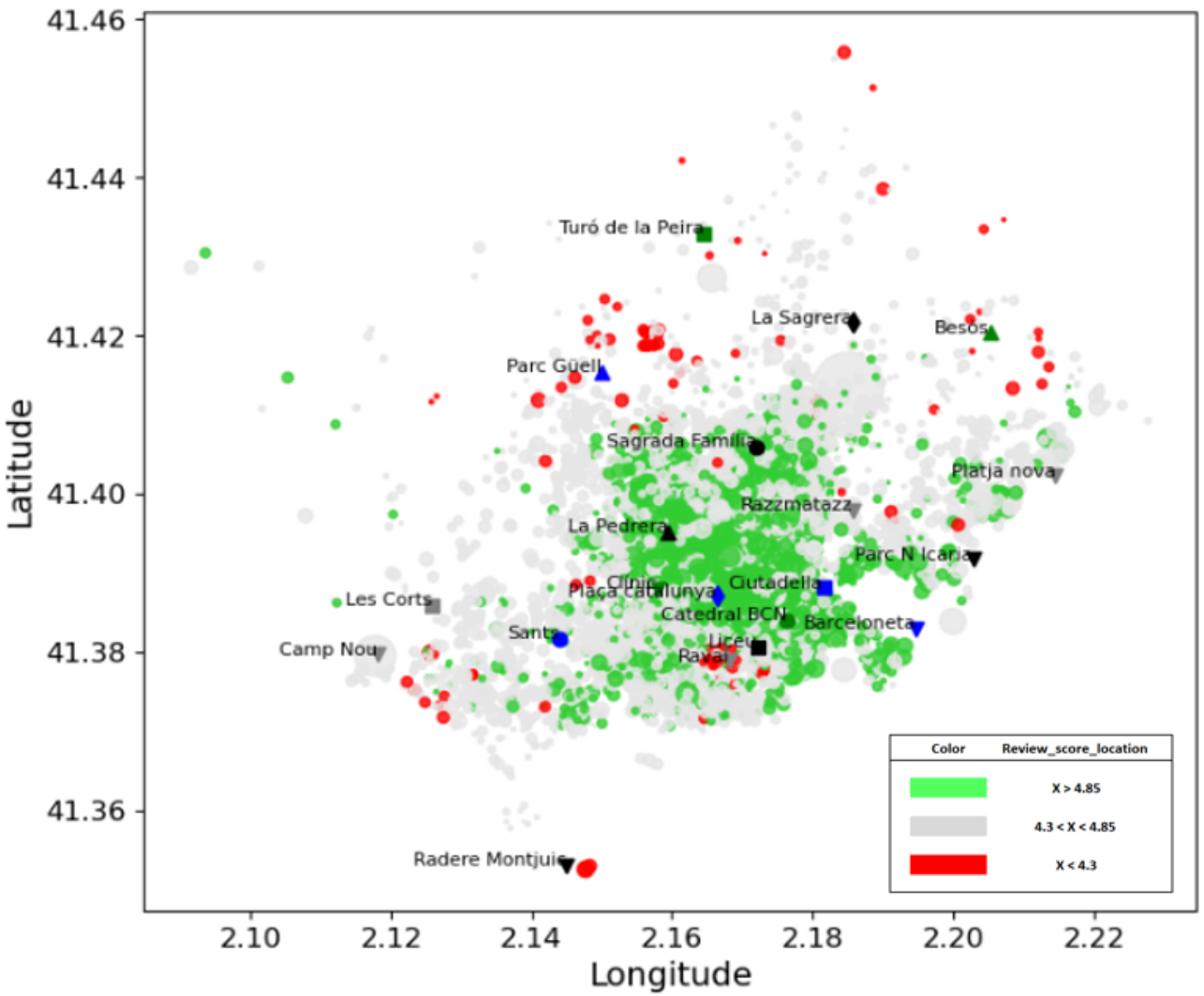
Valoració de Localització



Valoració de Localització



Valoració de Localització



Gestió dels Valors que Falten i Observacions Anòmales

Gestió de Valors que Falten

- Substituir valors que falten per la mediana.
- Substituir l'atribut per un altre booleà que indiqui si apareix.
- Eliminar l'atribut.

Gestió d'Observacions Anòmales

- No s'ha trobat observacions anòmales significatives

Continguts

Tractament de Dades

Base de Dades d'Entrenament

Es divideix la base de dades en dos conjunts: un conjunt d'entrenament i un conjunt de prova.
Això s'ha fet amb la funció de *train_test_split()* de *sklearn* per dividir la base de dades origen en 80% per entrenar i 20% per testejar. En aquest estudi no s'ha utilitzat validació creuada.

El tractament de dades d'aquesta base de dades consta de:

- Tractament de dades categòriques, text i numèriques
- Creació de nous atributs
- Automatitzacions

One Hot Encoding

One Hot Encoding és una tècnica per tractar les variables tipus text categòriques. La majoria d'algoritmes d'aprenentatge automàtic prefereixen treballar amb nombres i aquí és on entra aquesta tècnica, ja que converteix les categories en nombres.

Tractament de Dades Categòriques

One Hot Encoding

One Hot Encoding és una tècnica per tractar les variables tipus text categòriques. La majoria d'algoritmes d'aprenentatge automàtic prefereixen treballar amb nombres i aquí és on entra aquesta tècnica, ja que converteix les categories en nombres.

En aquest treball hem utilitzat la llibreria *sklearn* per processar les dades categòriques utilitzant One Hot Encoding a les següents variables:

- Property_type
- Room_type
- Amenities
- neighbourhood
- neighbourhood_cleansed
- neighbourhood_group_cleansed

Embedding

Text *embedding* fa referència a la representació numèrica d'un text en un espai vectorial continu. És un mètode utilitzat en Processament del Llenguatge Natural (PNL) per transformar paraules o frases en vectors de números.

Tractament de Text

Embedding

Text *embedding* fa referència a la representació numèrica d'un text en un espai vectorial continu. És un mètode utilitzat en Processament del Llenguatge Natural (PNL) per transformar paraules o frases en vectors de números.

En aquest treball hem utilitzat un model preentrenat d'embeddings basat en *transformers* que es diu *bert-base-uncased* i s'ha aplicat a les següents variables:

- Description
- Reviews
- Neighbourhood_overview

Reducció de Components

PCA (Principal Component Analysis)

És un algoritme de reducció dimensional. Primer identifica l'hiperplà que hi ha més a prop de les dades i, després, fa una projecció de les dades a ell. D'aquesta manera es projecten les dades a una dimensió inferior i per tant passem a tenir una base de dades amb menys atributs.

Reducció de Components

PCA (Principal Component Analysis)

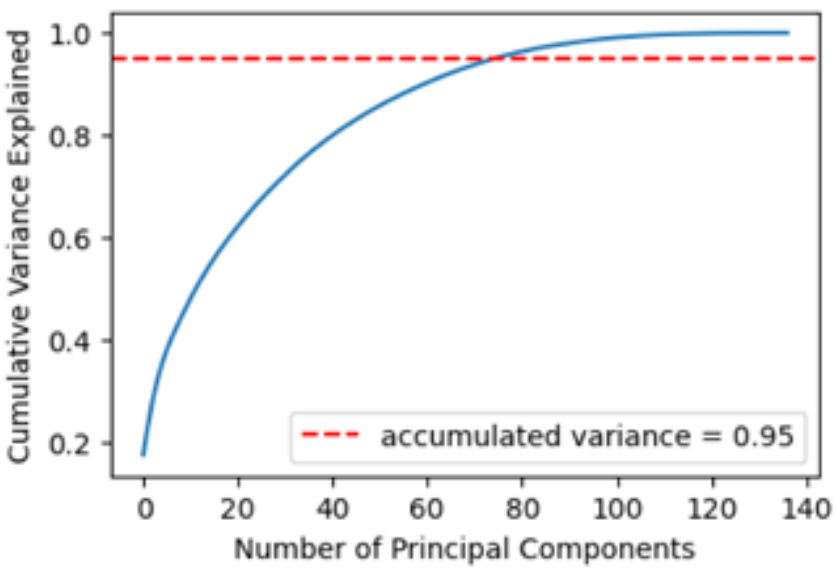
És un algoritme de reducció dimensional. Primer identifica l'hiperplà que hi ha més a prop de les dades i, després, fa una projecció de les dades a ell. D'aquesta manera es projecten les dades a una dimensió inferior i per tant passem a tenir una base de dades amb menys atributs.

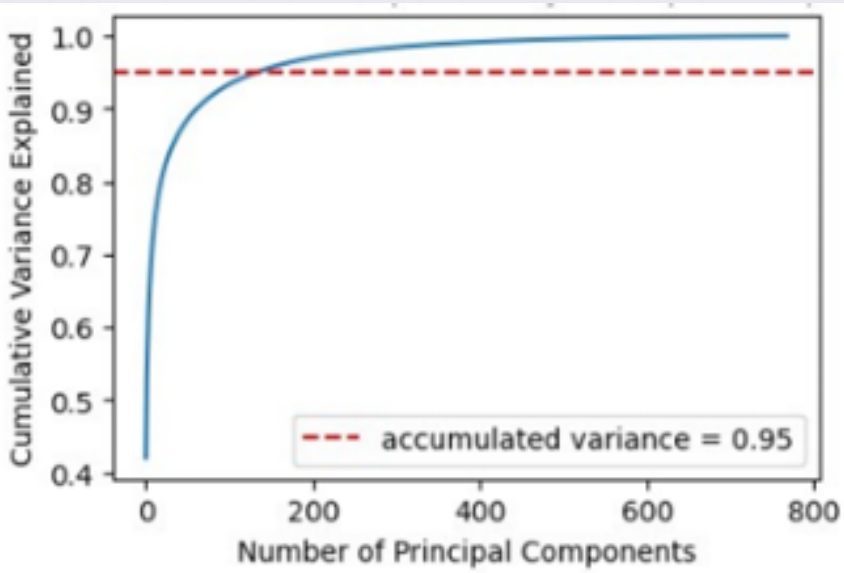
En aquest treball hem utilitzat la llibreria *sklearn* per fer la reducció de dimensió i s'ha aplicat a les següents variables:

- Amenities
- Description

Reducció de Components

Suma de la Variancia Acumulada: *Amenities* i *Description*





Altres Tractaments a Variables de Text

- *Host_is_superhost, host_has_profile_pic i host_identity_verified*: Transformar les variables t i f a una variable booleana ("t" i "f" —> 1 i 0 respectivament)
- *Price*: Treure el símbol de dolar i convertir els valors en float. ("33.3$" —> 33.3)
- *First_review i last_review*: convertir la data en una marca de temps. ("23/12/2022" —> 1671753600000)
- *Host_acceptance_rate i host_response_rate*: Treure el símbol % i convertir els valors en float. ("25%" —> 25)

Tractament de Dades Numèriques

Estandarització

Utilitzar *StandardScaler* per estandarditzar les variables numèriques assegurant que tinguin una mitjana zero i una desviació estàndard d'1.

També es va provar d'utilitzar la Normalització utilitzant *MiniMaxScaler*, però donava resultats iguals o pitjors a Estandarització.

Creació de Nous Atributs

Nous atributs

A part dels atributs generats utilitzant One Hot Encoding, Embeddings i PCA, s'han creat les següents variables:

- Distàncies a llocs remarcats
- Seniment dels comentaris
- Variables descriptives de la descripció (size, capacity, alowed_under_25, family_friendly)

Automatitzacions

Entrenament

Models

- Regressió Lineal
- Aprenentatge conjunt i Random Forest

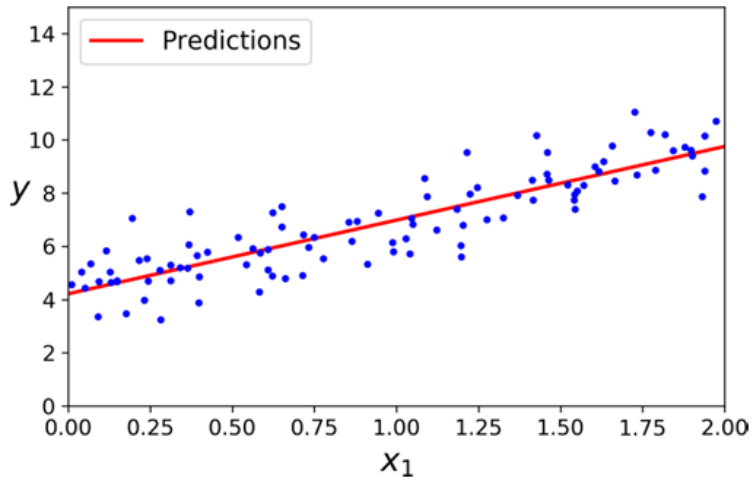
Regressió Lineal

Fòrmula

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n \tag{1}$$

On:

- b_0 és la intersecció amb l'eix y
- $b_1, b_2, \ldots, b_n$ són els coeficients associats a les variables d'entrada $x_1, x_2, \ldots, x_n$.



Aprenentatge Conjunt

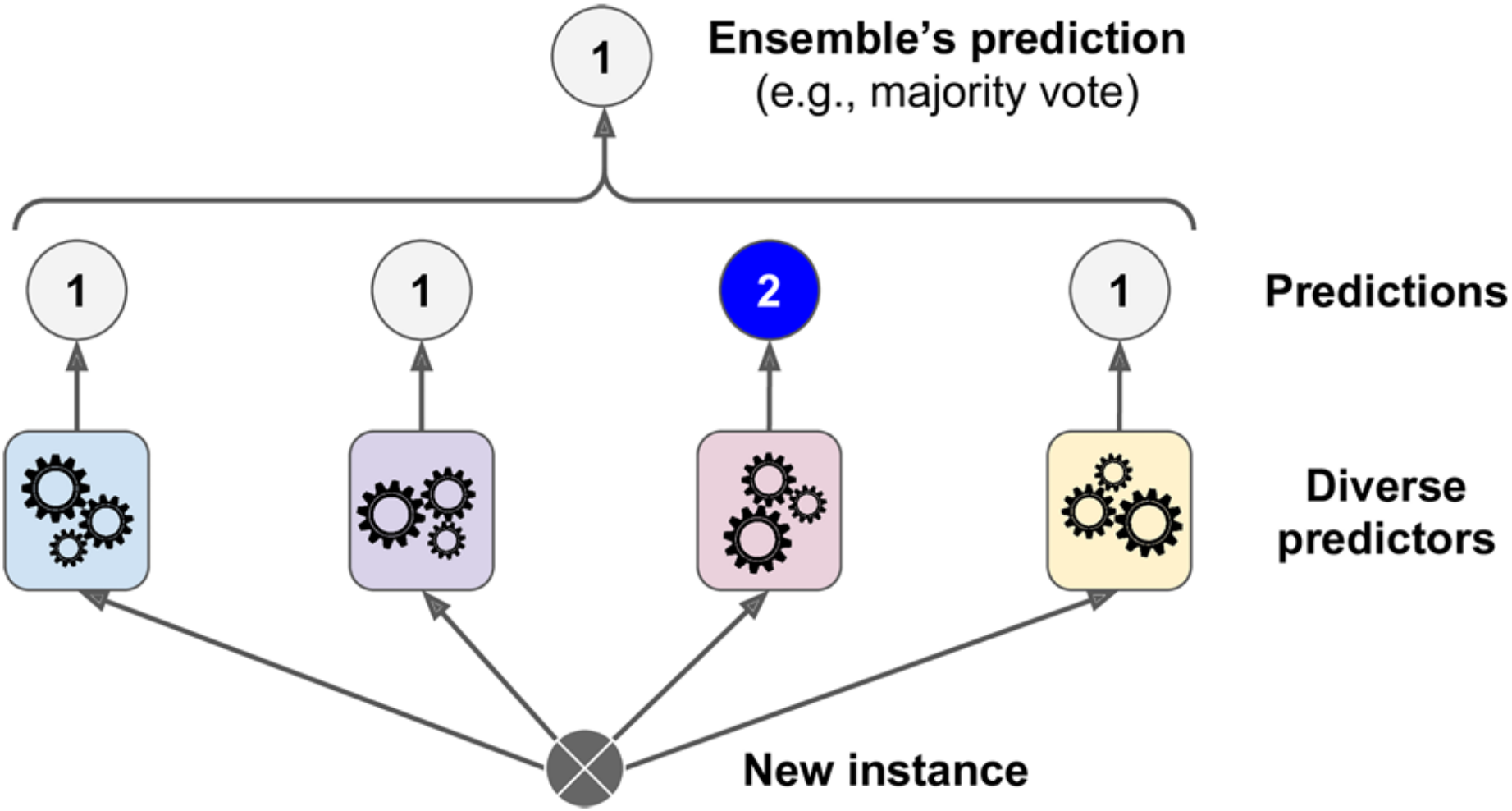
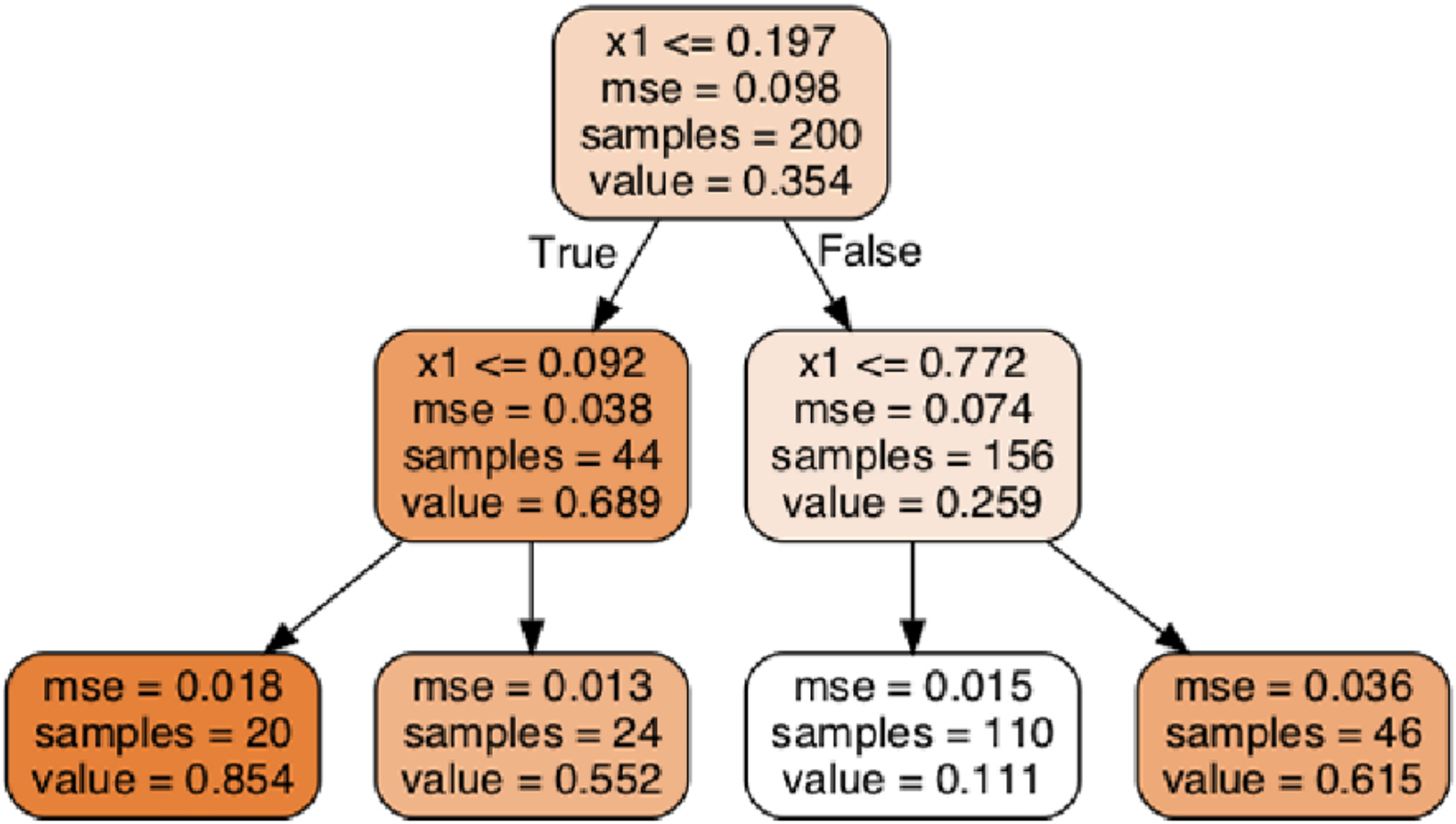


Figura: Algoritmes de votació

Arbres de Decisió



Aprenentatge Conjunt

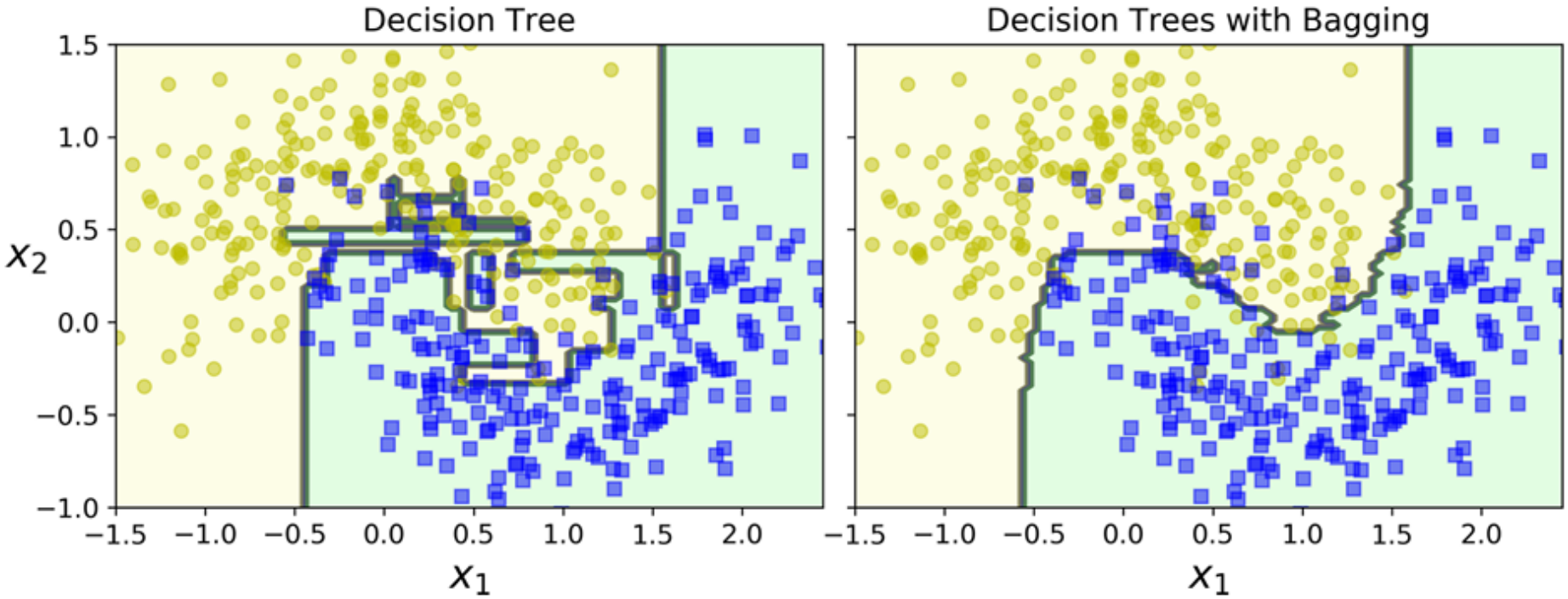


Figura: Bagging

Continguts

Resultats

Mesura de Validació

S'han utilitzat dues mesures de rendiment per avaluar els models:

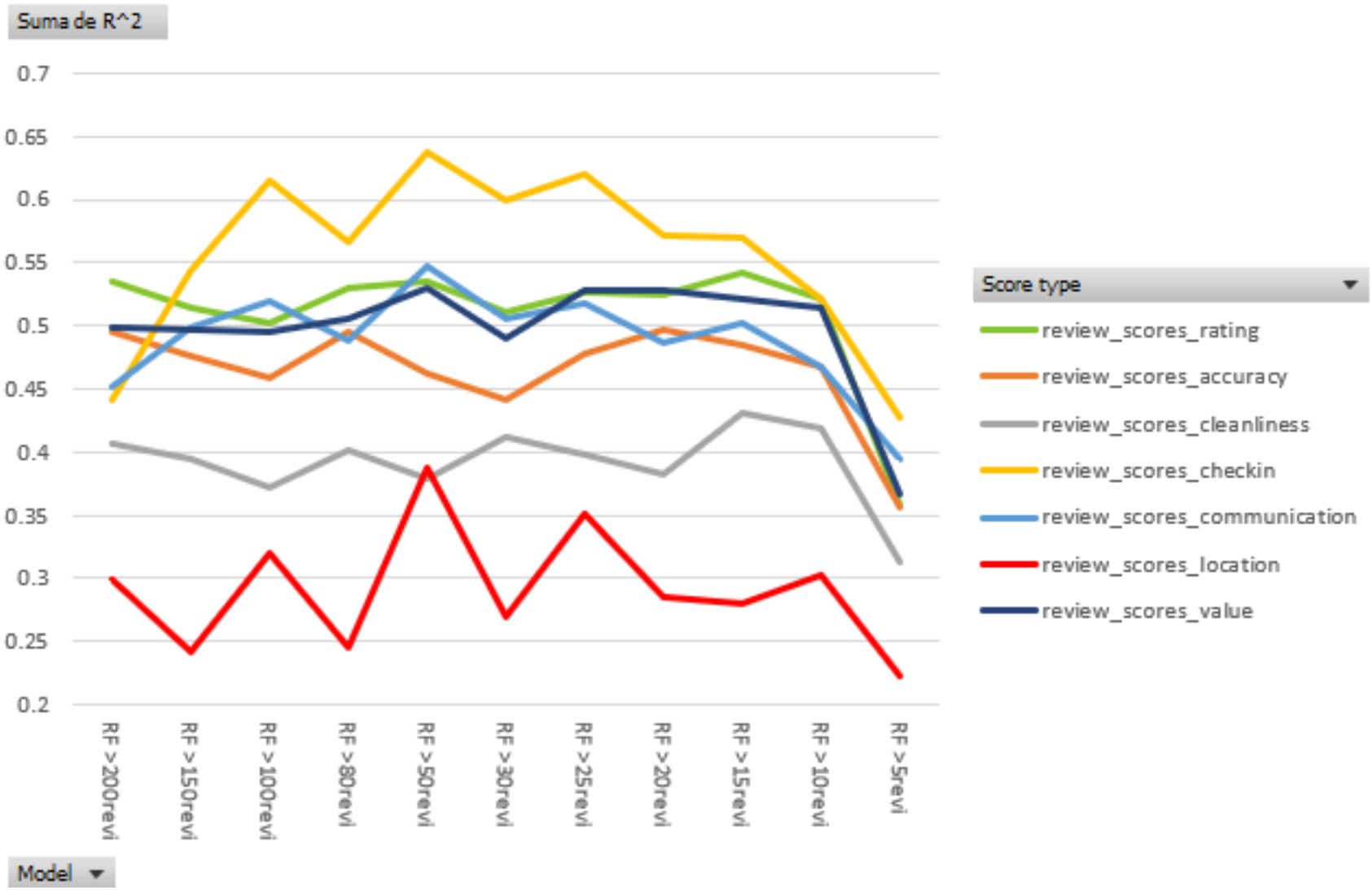
- Coeficient de determinació R^2
- Error Quadràtic Mitjà MSE

Utilitzarem com a mesura del rendiment principal R^2, ja que indica la precisió dels models en predir les dades observades. Tindrà valors entre 0 i 1 on 1 és màxima precisió.

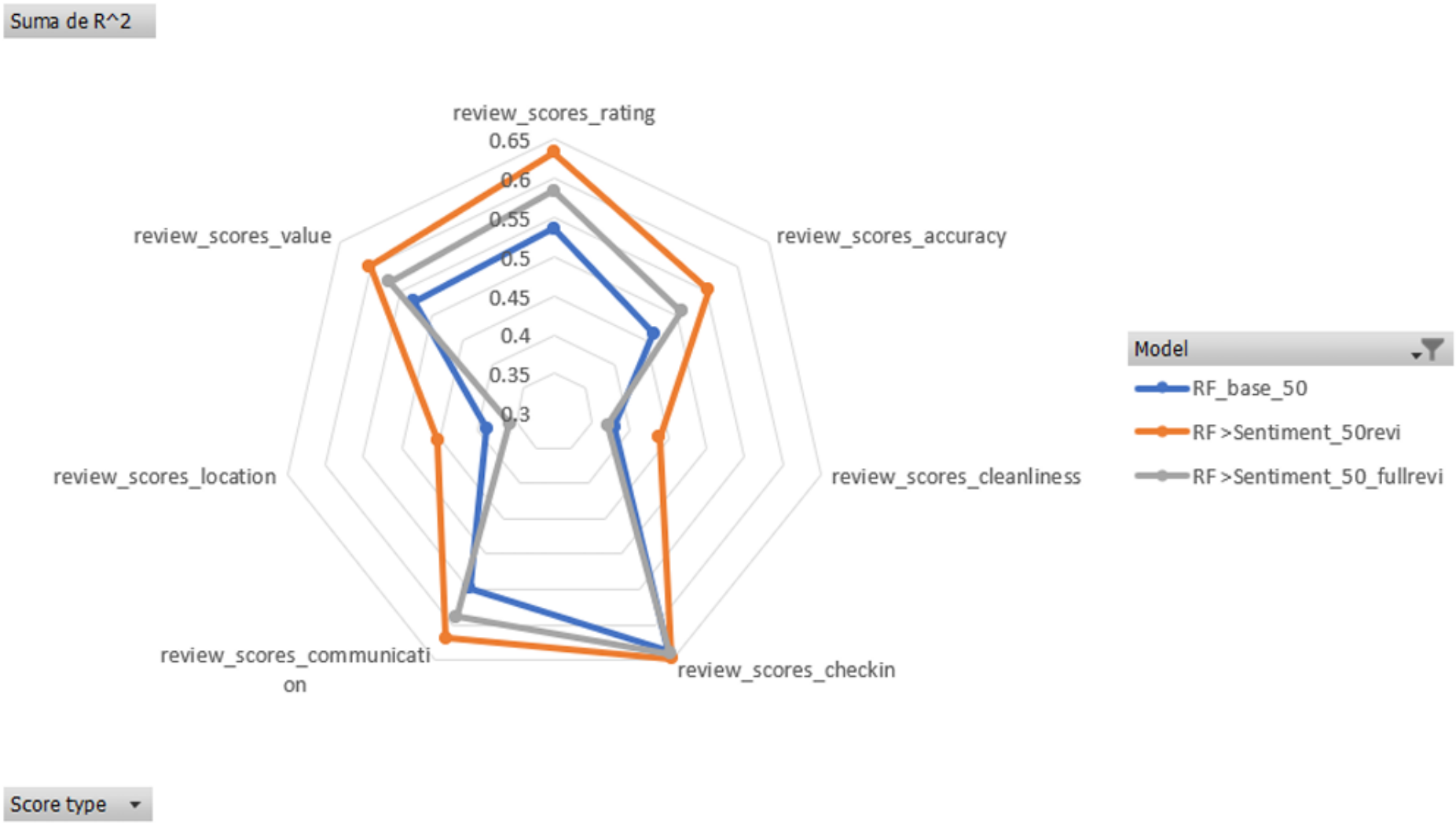
Prova 1: Model de Referència

Score type	mse_RL	mse_RF	r2_RL	r2_RF
rating	0.0227	0.0202	0.3746	0.4442
accuracy	0.0168	0.0148	0.3326	0.4163
cleanliness	0.0287	0.0262	0.2398	0.3073
checkin	0.0135	0.0099	0.4178	0.5706
communication	0.0133	0.0118	0.3751	0.4446
location	0.0123	0.0117	0.2136	0.2513
value	0.0194	0.0179	0.3754	0.4264

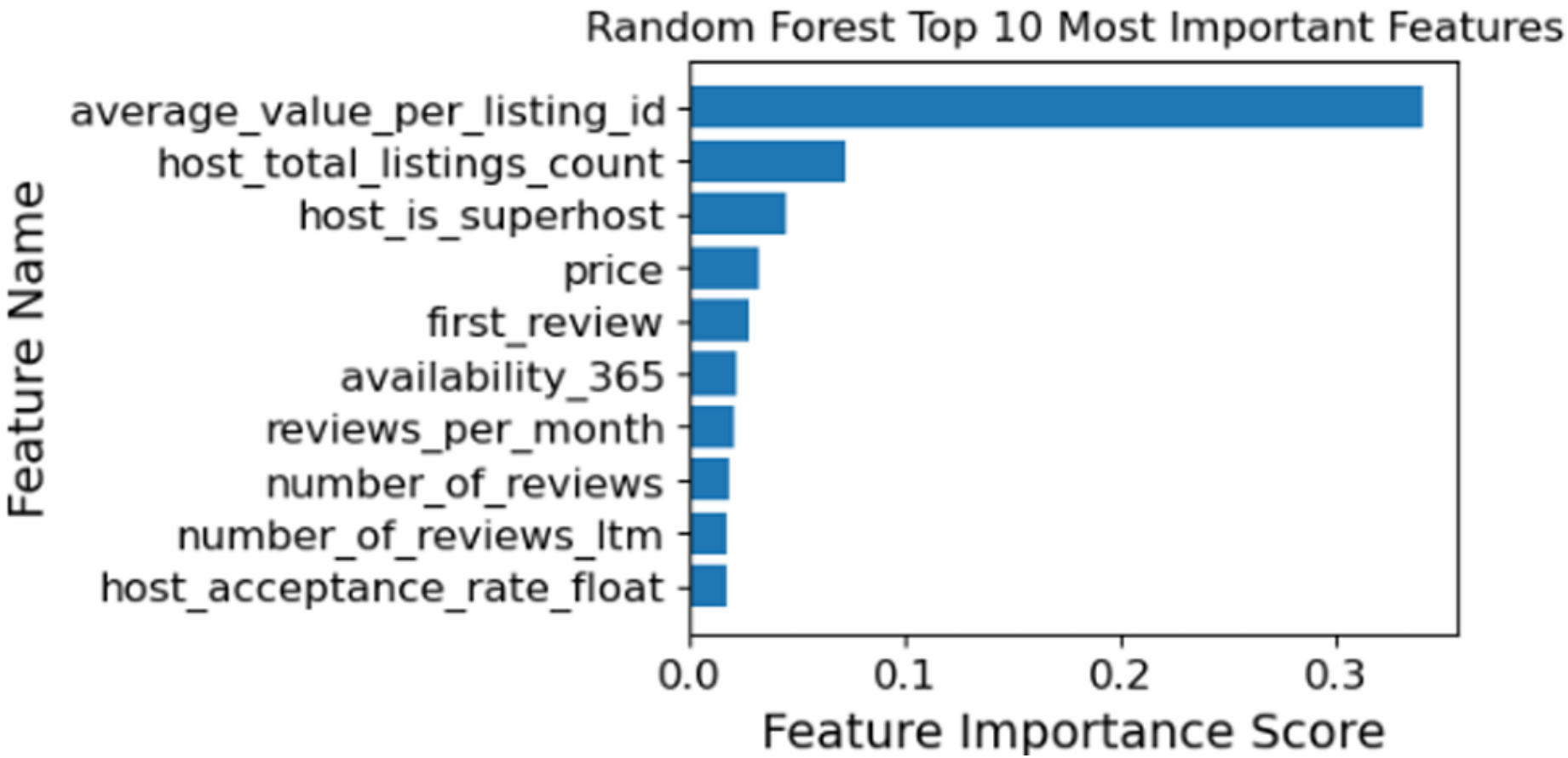
Prova 2: Número Mínim de Valoracions per Allotjament



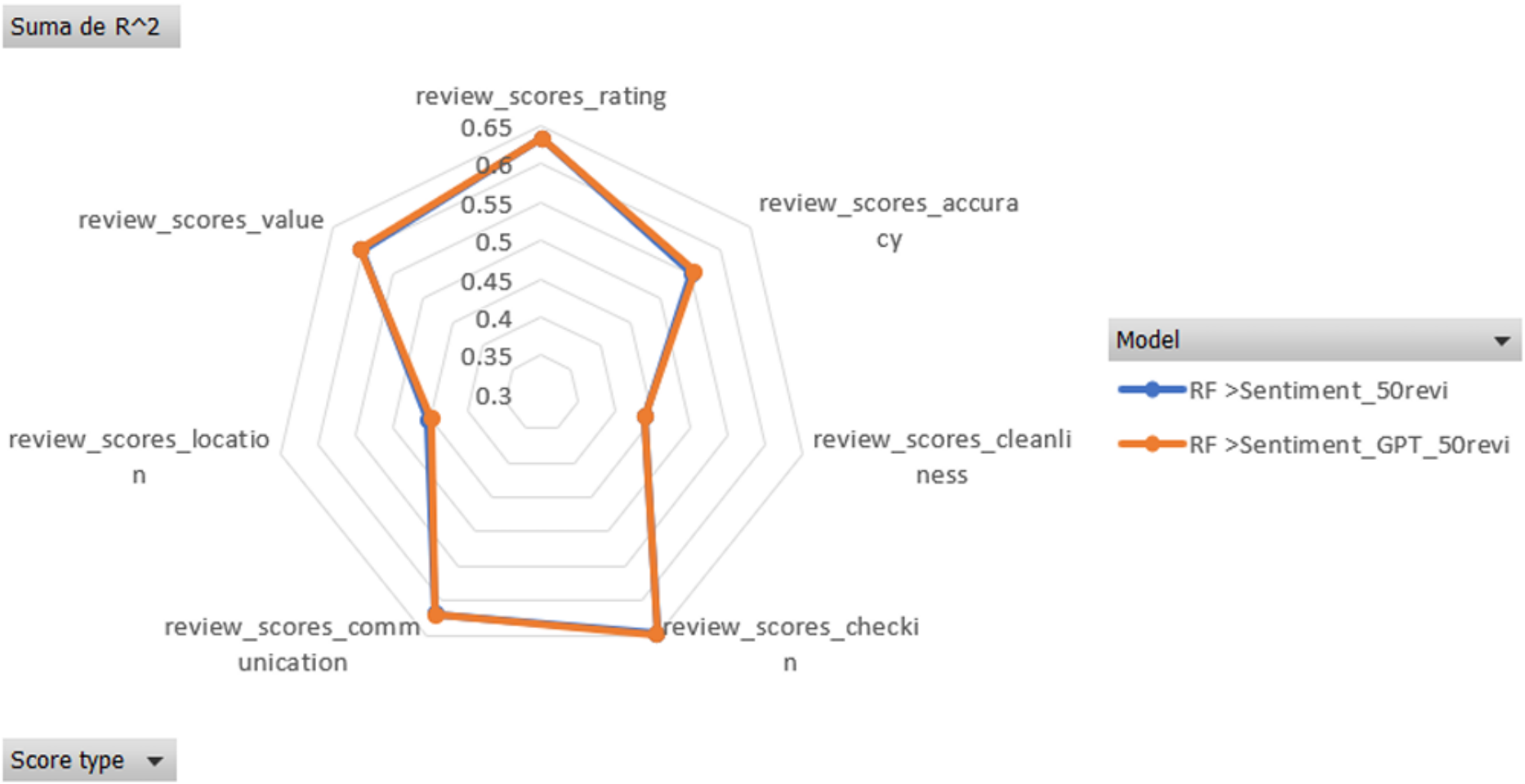
Prova 3: Anàlisi de Sentiment



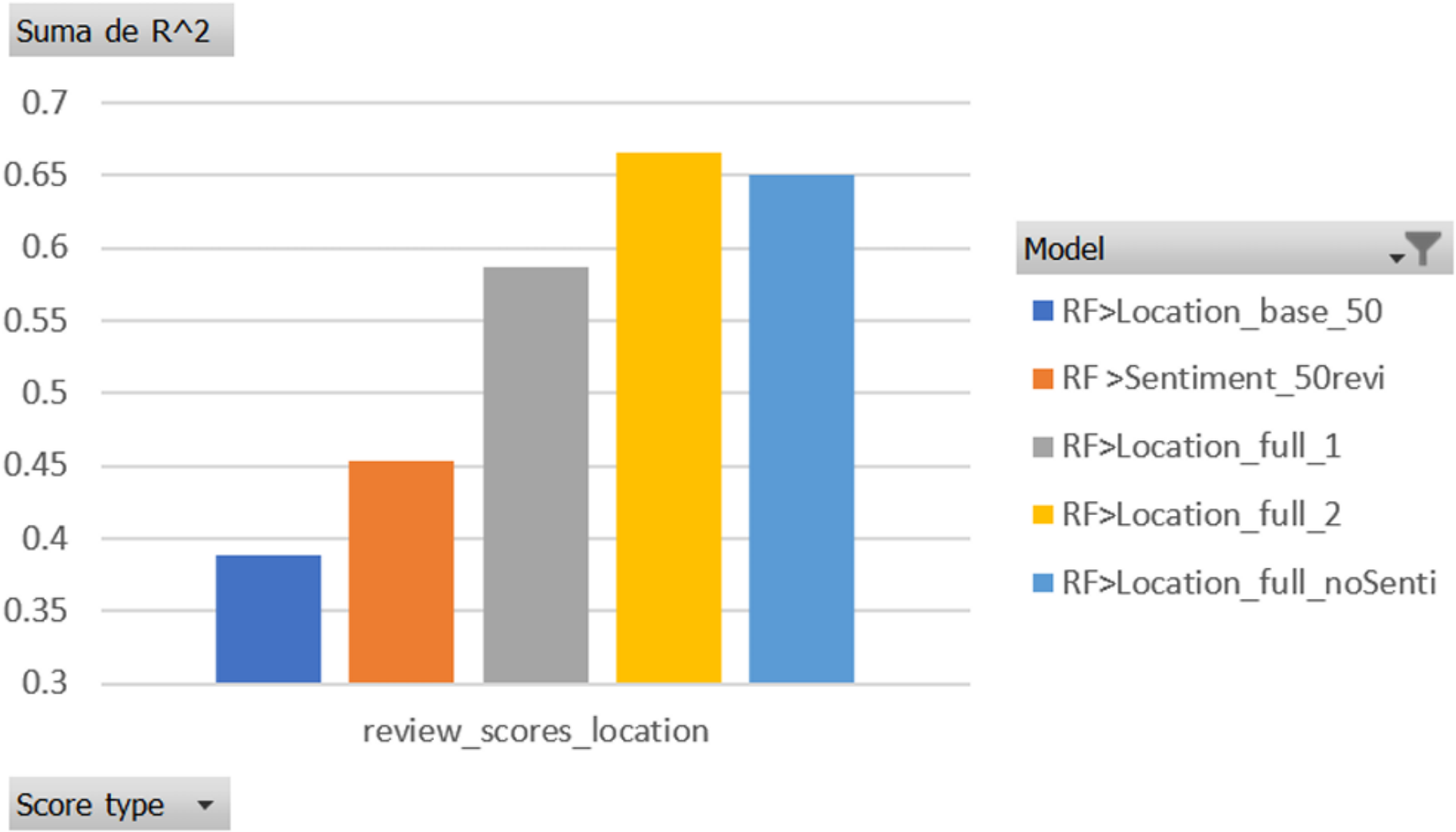
Prova 3: Anàlisi de Sentiment



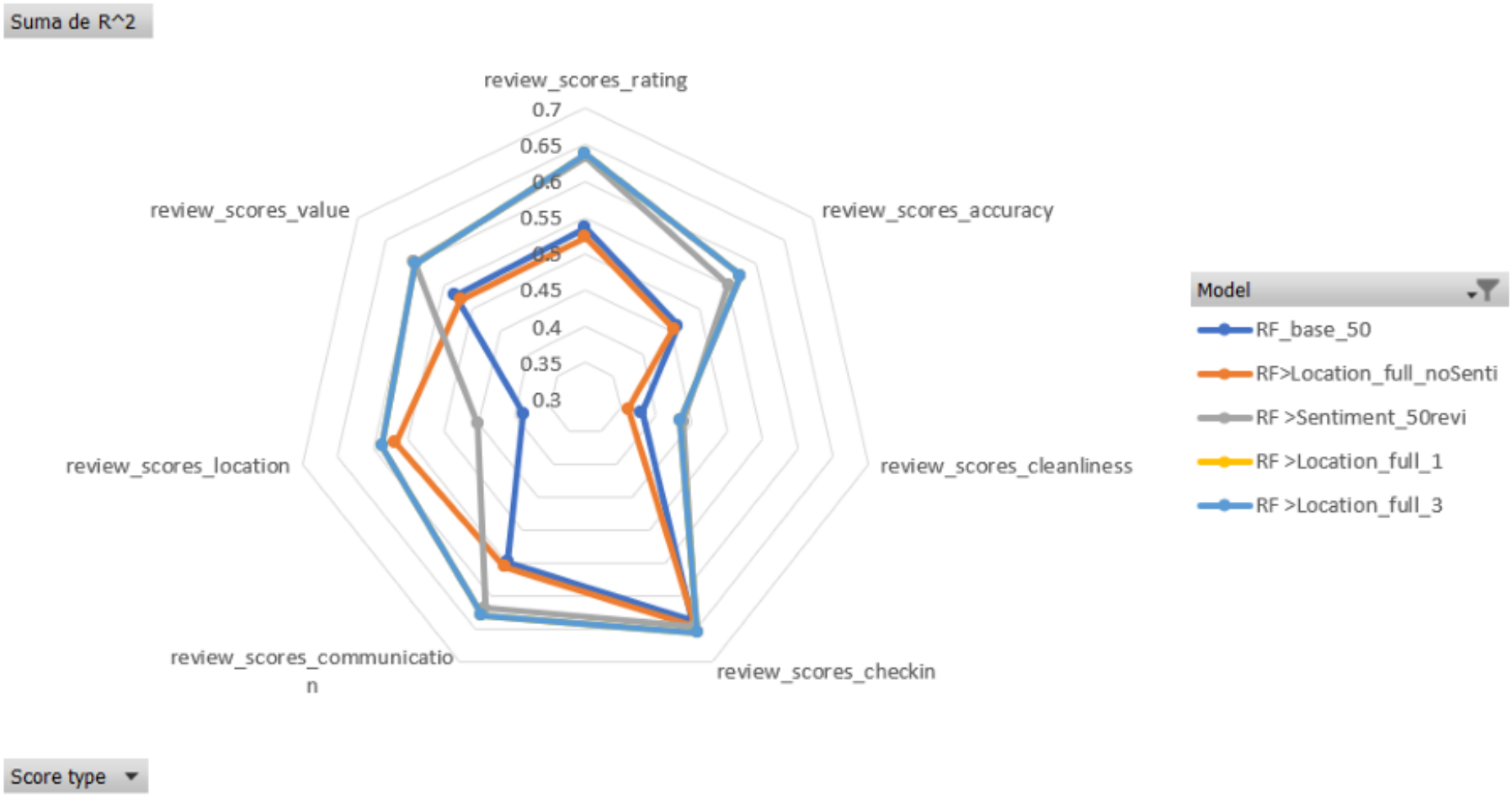
Prova 4: Anàlisi amb ChatGPT



Prova 5: Localització



Prova 5: Localització



Continguts

Conclusions i Futur Treball

Conclusions

- S'ha adquirit i estructurat una base de dades rellevant.
- S'ha realitzat el preprocessament de les dades.
- S'ha desenvolupar un model predictiu de valoracions.

Futur Treball

- Explorar la diferents models i *feature turning* per millorar la precisió.
- Explorar més transformacions d'atributs que ajudin a representar millor la base de dades.
- Interpretacions dels models a través de tècniques de XAI com SHAP i LIME.
- Crear una pàgina web amb la capacitat de mostrar els resultats i seguir aprenent amb les noves dades.