Informe de transparencia y equidad del modelo

16.09.2020

ALTRAN
Part of Capgemini

Informe de transparencia y equidad del modelo

16.09.2020

Content

1 Introducción

El objeto del presente documento es **analizar la equidad** de los tres modelos desarrollados para el categorizador de tickets, mediante el cálculo de las métricas *Precision*, *Recall* y *F1 Score* para los grupos de tickets seleccionados, y **analizar la transparencia** de los tres modelos desarrollados para el categorizador de tickets, a través de los resultados obtenidos al aplicar técnicas de visualización y técnicas interactivas a esos modelos con el objetivo de entender y explicar las decisiones que toman dichos modelos.

2 Recopilación de modelos

En este apartado se describen las tareas realizadas para recopilar los modelos seleccionados o desarrollados durante la ejecución del proyecto piloto desarrollado por la Generalitat de Catalunya para implantar una serie de dispositivos inteligentes en el proceso de entrada de mails al servicio de atención al usuario SAU del CTTI, que actuarán en el momento de recibir los correos electrónicos dirigidos al buzón genérico del SAU, y que realizarán tres tipos de acciones:

1. En primer lugar, usando herramientas de procesamiento del lenguaje natural, clasificarán la "intención" del correo electrónico según una clasificación establecida (si es una incidencia, un soporte, ...).
2. En segundo lugar, usando herramientas de procesamiento del lenguaje natural, clasificarán todos aquellos correos que hayan sido clasificados como incidencias según el bloque técnico al que correspondan (puesto de trabajo, CPDs o aplicaciones).
3. En tercer lugar, usando herramientas de integración y automatización crearán un ticket en la herramienta de Remedy mediante los servicios de integración de los que dispone esta herramienta.

En la siguiente figura puede verse un diagrama general de la solución propuesta para el categorizador de tickets.

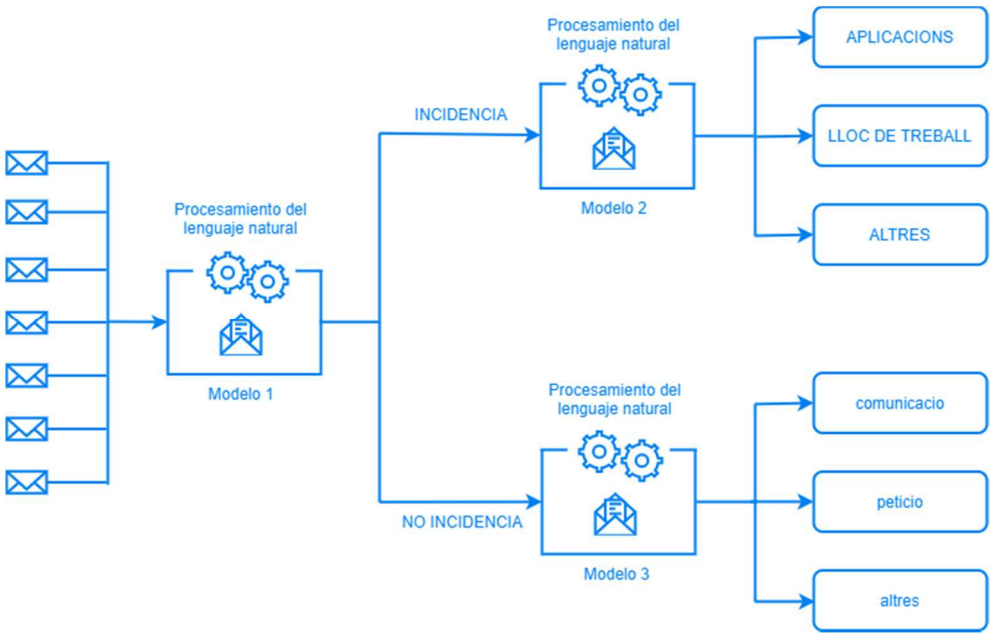


Figura 1 Diagrama general del categorizador de tickets.

Los modelos seleccionados, por tanto, permitirán categorizar los correos electrónicos dirigidos al buzón genérico del SAU, una vez que hayan sido tratados mediante técnicas de procesamiento del lenguaje natural, para que el sistema finalmente cree los tickets correspondientes a dichos correos en la herramienta Remedy.

Los modelos proporcionados por el proveedor encargado de la ejecución del proyecto piloto están en formato PKL, y en la siguiente tabla puede verse el nombre de los distintos ficheros proporcionados, así como su tamaño y breve descripción de su contenido.

Nombre fichero	Tamaño fichero	Descripción
isInc.pkl	129 MB	Primera versión del modelo que clasifica los tickets en INCIDENCIA y NO INCIDENCIA.
isInc_v2.pkl	3,93 MB	Segunda versión del modelo que clasifica los tickets en INCIDENCIA y NO INCIDENCIA.
wheninc.pkl	4,54 MB	Primera versión del modelo que clasifica los tickets que son incidencia en ALTRES, APLICACIONS o LLOC DE TREBALL.
wheninc_v2.pkl	4,54 MB	Segunda versión del modelo que clasifica los tickets que son incidencia en ALTRES, APLICACIONS o LLOC DE TREBALL.
whenNonInc.pkl	681 KB	Primera versión del modelo que clasifica los tickets que son incidencia en ALTRES, COMUNICACÓO o PETICIÓ.
whenNonInc_v2.pkl	681 KB	Segunda versión del modelo que clasifica los tickets que son incidencia en ALTRES, COMUNICACÓO o PETICIÓ.

Tabla 1 Modelos seleccionados para ser utilizados en el categorizador de tickets.

En la siguiente figura puede verse un diagrama de flujo que aclara cómo interactuarán los correos electrónicos dirigidos al buzón genérico del SAU, y los distintos modelos seleccionados para clasificar dichos correos.

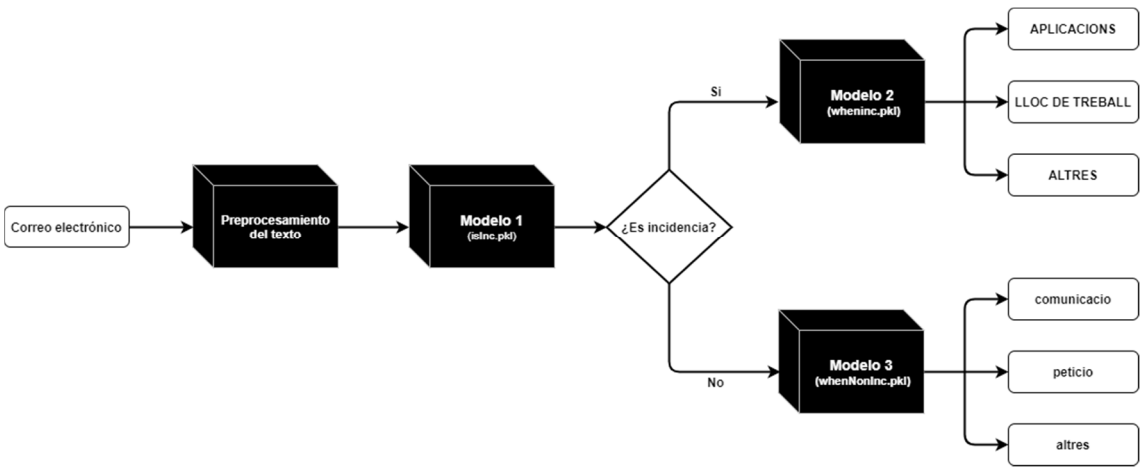


Figura 2 Diagrama de flujo del categorizador de tickets.

Junto con los ficheros que contenían los modelos seleccionados, también se recopilaron otros ficheros que se habían utilizado para procesar el texto existente en las descripciones de los tickets, mediante técnicas de procesamiento natural. En la siguiente tabla puede verse el nombre de los distintos ficheros recopilados, así como su tamaño y breve descripción de su contenido.

Nombre fichero	Tamaño fichero	Descripción
stop_words.txt	5,12 KB	Lista de stop words que se eliminarán del texto existente en los correos electrónicos.
saludos.txt	409 bytes	Lista de saludos que se eliminarán del texto existente en los correos electrónicos.
output.txt	321 KB	Stems (no utilizado en ninguna de las versiones de los modelos recopilados).
voc.txt	457 KB	Vocabulario (no utilizado en ninguna de las versiones de los modelos recopilados).
parse.py	4,00 KB	Fichero de código que implementa la funcionalidad que permite eliminar sender, receiver, issue, footer del texto existente en los correos electrónicos.

Tabla 2 Ficheros utilizados para procesar el texto existente en las descripciones de los tickets.

Debido a que los modelos cuya equidad y transparencia se va a evaluar tienen su origen en un proyecto piloto desarrollado por la Generalitat de Catalunya que no se ha finalizado todavía, los análisis se han realizado sobre modelos que pueden variar en la versión definitiva del categorizador de tickets, y por tanto es necesario tener en cuenta que los resultados y conclusiones obtenidas podrían no ser aplicables a dicha versión final.

3 Construcción de datos

Para la visualización de los tres modelos comprendidos en categorizador de tickets se crearon además tres conjuntos nuevos de datos donde simplemente seleccionaban aquellos datos a los que se aplicaba cada modelo:

Nombre fichero	Descripción
preprocessed_filtered_isInc.csv	Conjunto de datos equivalente a "preprocessed.csv" en el que se ha añadido la columna isInc existente en el fichero "preprocessed.csv". Esta columna recoge si el ticket es una incidencia o no.
preprocessed_filtered_whenInc.csv	Conjunto de datos equivalente a "preprocessed_filtered.csv" en el que se han seleccionado o filtrado solo las incidencias, y en el que además se ha añadido la columna whenInc que indica el tipo de incidencia y que posteriormente se utiliza para llevar a cabo la categorización.
preprocessed_filtered_whenNonInc.csv	Conjunto de datos equivalente a "preprocessed_filtered.csv" pero en el que se han seleccionado o filtrado solo las no incidencias, y en el que además se ha añadido la columna whenNonInc que indica el tipo de no incidencia y que posteriormente se utiliza para llevar a cabo la categorización.
preprocessed_tickets.csv	Conjunto de datos equivalente a "preprocessed.csv" en el que se han añadido las columna isInc, whenInc y whenNonInc. Se han eliminado los tickets que tuviesen alguna de las columnas con nan, y se utilizará para visualizar y seleccionar el texto de los tickets, en la aplicación web desarrollada para experimentar con LIME.

Tabla 3 Conjuntos de datos utilizados para realizar la visualización del modelo categorizador de tickets.

La columna isInc mencionada en la Tabla 3 se calcula a partir del valor de la columna "RESOLUTION_CATEGORY". Como explico el proveedor del categorizador de tickets, la categoría NO INCIDÈNCIA se sustituye por no y el resto se engloban en la categoría yes.

La columna whenInc mencionada en la Tabla 3 se calcula a partir del valor de la columna "CLOSURE_PRODUCT_CATEGORY_TIER1". Como explico el proveedor del categorizador de tickets, las categorías APLICACIONS y LLOC DE TREBALL no se modifican y el resto se engloban en la categoría ALTRES.

La columna whenNonInc mencionada en la Tabla 3 se construye a partir del valor de la columna "RESOLUTION_CATEGORY_TIER2". Como explico el proveedor del categorizador de tickets, las categorías ÉS UN CANVI y ÉS CONSULTA se sustituyen por COMUNICACIÓ, la categoría ÉS UNA PETICIÓ se sustituye por PETICIÓ y la categoría ÉS DESENVOLUPAMENT se sustituye por ALTRES.

4 Métricas

A continuación se definirán las distintas métricas utilizadas en el presente documento, usando la nomenclatura inglesa: True Negative [TN], True Positive [TP], False Positive [FP], False Negative [FN]:

- **Accuracy** (exactitud). Mide el porcentaje de casos en que el modelo ha acertado. Esta métrica puede llevar a engaño, ya que puede hacer que un modelo parezca que es mucho mejor de lo que es realmente.

  Se calcula con la siguiente fórmula:

$$accuray = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision** (*precisión*). Mide la calidad del modelo en tareas de clasificación. Respondería, por ejemplo, a una pregunta del tipo, ¿qué porcentaje de los clientes que contactemos estarán interesados?

  Se calcula con la siguiente fórmula:

$$precision = \frac{TP}{TP + FP}$$

- **Recall** (exhaustividad). Mide la cantidad casos verdaderos que el modelo es capaz de identificar. Respondería, por ejemplo, a una pregunta del tipo. ¿qué porcentaje de los clientes que están interesados somos capaces de identificar?

  Se calcula con la siguiente fórmula:

$$recall = \frac{TP}{TP + FN}$$

- **F1-score** (valor F1). Combina las medidas de precision y recall en un sólo valor. Esto permite comparar de forma más sencilla el rendimiento combinado de la precisión y la exhaustividad en varios modelos.

  Se calcula haciendo la media armónica entre la precisión y la exhaustividad:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

5 Transparencia

Este apartado tratará de aplicar las diferentes técnicas de visualización utilizadas para poder entender y explicar el modelo categorizador de tickets.

En concreto este modelo está formado por tres submodelos:

- ***IsInc_v2.pkl***: Este submodelo contiene dos etapas, en la primera etapa del modelo se cuantifica el texto de los correos electrónicos y se transforma las palabras en vectores numéricos utilizando el vectorizador *TfidfVectorizer*. Estos vectores numéricos expresan cuan relevante es una palabra en cada correo electrónico. Una vez vectorizadas las palabras en la segunda etapa se aplica un modelo de regresión logística binomial para categorizar si correo el electrónico es una incidencia o no.
- ***Wheninc_v2.pkl***: Este segundo submodelo solamente se aplica en aquellos correos electrónicos que han sido categorizados como incidencias y al igual que en el caso anterior, contiene dos etapas. De la misma manera en la primera etapa del modelo se cuantifica el texto de los correos electrónicos mediante el vectorizador *TfidfVectorizer*, y una vez vectorizadas las palabras, en la segunda etapa se aplica una regresión logística multiclase que categoriza el correo electrónico en *ALTRES*, *APLICACIONS* o *LLOC DE TREBALL*.
- ***WhenNoninc.pkl***: Este tercer submodelo solamente se aplica en aquellos correos electrónicos que han sido categorizados como no-incidencias con el submodelo y al igual que los casos anteriores contiene dos etapas. En la primera etapa del modelo cuantifica el texto de los correos electrónicos mediante el vectorizador *CountVectorizer*. En este caso, los vectores numéricos expresan la frecuencia de la palabra en cada correo electrónico. Una vez vectorizadas las palabras en la segunda etapa se aplica una regresión logística multiclase que categoriza el correo electrónico en *ALTRES*, *COMUNICACÓO* o *PETICIÓ*.

Las diferentes herramientas de visualización han sido aplicadas en cada uno de los tres submodelos con el fin de explicar y visualizar en detalle los pasos que se siguen en el modelo completo.

5.1 Visualización Ad-Hoc

Como se ha detallado en la introducción, cada uno de los tres submodelos de los que se compone el modelo completo del categorizador de tickets contiene una primera etapa en la que vectorizando las palabras se expresa numéricamente la relevancia de cada una de ellas en el correo electrónico, y una segunda etapa en la que a partir de estos vectores numéricos y aplicando una regresión logística, ya sea binomial o multiclase, se clasifica el correo electrónico en una u otra categoría. Ambas etapas son algoritmos o modelos simples, interpretables y explicables de por si, por lo tanto se puede desarrollar una herramienta visualización ad-hoc que especifique de manera clara, inequívoca y sin llevar a cabo ningún tipo de aproximación como se ha realizado la clasificación.

En concreto la **regresión logística binomial** asigna un peso a cada palabra del correo electrónico e indica cómo esta palabra contribuye a que cierto correo sea clasificado como una categoría o la opuesta. Es decir, por ejemplo en el modelo **IsInc_v2.pkl** si la regresión logística asigna un peso positivo a cierta palabra *X*, *esta* contribuirá a que el correo sea clasificado como *Incidencia,* y si por el contrario le asigna un peso negativo, la palabra *X* contribuirá a que sea clasificado como *No-incidencia*. Cuanto más grande sea el peso la palabra *X* más contribuirá a la clasificación, ya se a favor de *Incidencia* o *No-incidencia*. La suma total de los pesos de todas las palabras determina si el correo electrónico es clasificado como *Incidencia* o *No-incidencia*.

En la Figura 3 se muestran las 20 palabras con mayor peso para el modelo *IsInc_v2.pkl* (izquierda), es decir las que más contribuyen a la clasificación cuando aparecen en un correo electrónico dado, y la distribución de pesos para todas las palabras que contribuyen en este mismo modelo. Como se observa cuando la palabra *pica* o *tais* aparecen en el

correo electrónico, estas palabras contribuyen a que el correo sea clasificado como no *No-incidencia*, puesto que tienen un peso negativo, mientras que *contrasenya* y *borsa* contribuyen a que sea clasificado como *Incidencia*, ya que tienen un peso positivo.

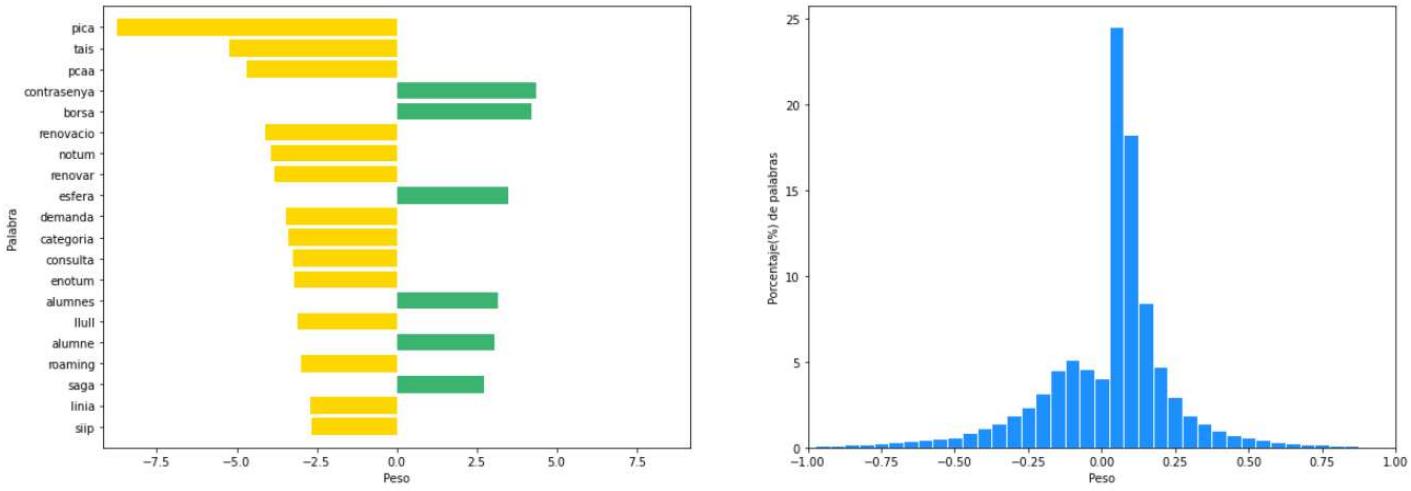


Figura 3: Representación de los pesos para el modelo IsInc_v2.pkl. Izquierda: Las 20 palabras con mayor peso para el modelo. Derecha: Distribución de pesos para todas las palabras que contribuyen en el modelo, debido que el porcentaje de palabras con pesos mayores (menores)a 2 (-2) es muy bajo la gráfica se ha centrado en el intervalo [-2,2] para una mejor visualización.

Es importante reseñar, que además de los pesos dependientes de las palabras del correo existe un valor constante denominado **intercepto** que se suma también a los pesos de las palabras y que indica la predisposición por defecto o inherente del modelo (independientemente de las palabras que aparecen en él) a que un correo sea clasificado como *Incidencia* o *No-incidencia*. Es decir, un correo será clasificado como *Incidencia* cuando la suma de los pesos de las palabras y el intercepto sea positiva y como *No-incidencia* cuando la suma sea negativa.

En el supuesto caso de que el correo estuviese vacío y no contuviese ninguna palabra, la clasificación vendrá determinada por el intercepto: *Incidencia* si es positivo y *No-incidencia* si es negativo. Para el caso del modelo *IsInc_v2.pkl* el intercepto tiene un valor de 2,10, es decir, cuando la suma total de palabras (Figura 4 barras negras) sea menor a -2,10 (Figura 4 línea vertical roja) será clasificado como *No-incidencia,* puesto que la probabilidad dada por el modelo de ser incidencia será menor a 0.5, y cuando sea mayor será clasificado como *Incidencia* (Figura 4 curva azul). La probabilidad del modelo está definida de 0 a 1, indicando que cuando la $0 < Probabillidad_{Incidencia} < 0.5$ la clasificación es *No-incidencia* y $0.5 \leq Probabillidad_{Incidencia} \leq 1$ la clasificación es *Incidencia* (Figura 4 eje vertical).

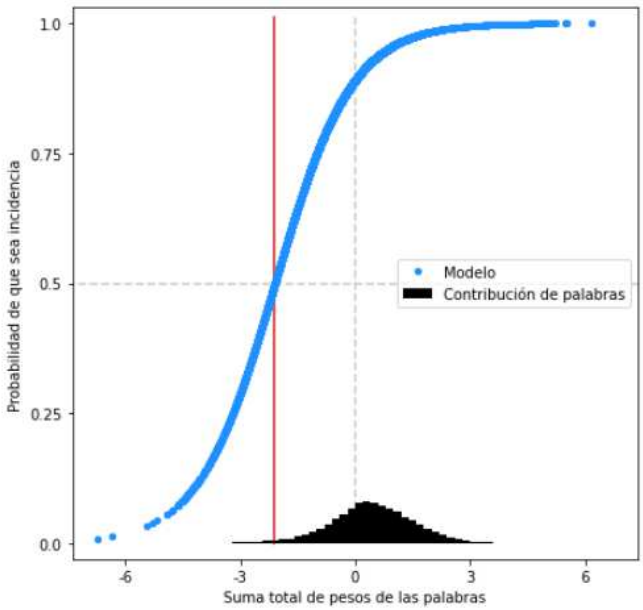


Figura 4: Probabilidad de que un correo electrónico sea clasificado como Incidencia con respecto a la contribución total de sus palabras. En la curva azul está indicada la probabilidad total que da el modelo IsInc_v2.pkl, en las barras negras la distribución de la suma total de pesos para todos los correos electrónicos que se encuentran en los datos preprocessed_filtered_isInc.csv y en la línea roja la suma total de pesos para la cual la clasificación pasa de No-Incidencia a Incidencia, es decir el valor –Intercepto (Intercepto multiplicado por -1).

Los **pesos de cada palabra y el intercepto ya han sido ajustados durante la fase de entrenamiento y ajuste del modelo,** desarrollada por el proveedor encargado de la ejecución del proyecto piloto del categorizador de tickets, y están presentes en el fichero PKL de cada modelo recopilado, por lo que la herramienta de visualización ad-hoc simplemente los extrae, los lee y los visualiza de manera simple e intuitiva, para ayudar al entendimiento del proceso de clasificación.

Además, con el fin de estudiar en detalle el comportamiento del modelo se ha comparado la probabilidad real y la probabilidad que da el modelo para que sea clasificado como incidencia.

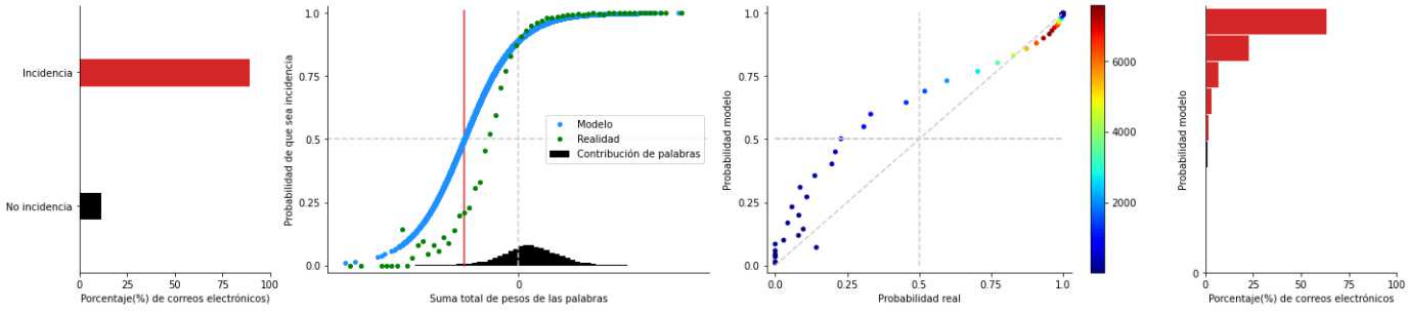


Figura 5: Probabilidad de clasificación del modelo vs real. Sub-figura 1: Distribución de correos electrónicos Incidencia (rojo) y No-Incidencia para los datos preprocessed_filtered_isInc.csv. Sub-figura 2: Probabilidad de clasificación incidencia del modelo (verde) y real (azul) para todos los correos electrónicos. Sub-figura 3: Probabilidad de clasificación incidencia del modelo (eje vertical) versus real (eje horizontal). El gradiente de color indica el número de correos electrónicos para cada para de probabilidad real-modelo. Sub-figura 4: Distribución de probabilidad que da el modelo para los datos correos electrónicos de preprocessed_filtered_isInc.csv.

En la Figura 5 se comprueba que a pesar de haber una diferencia sustancial entre la probabilidad estimada por el modelo (Figura 5 sub-figura 2 puntos azules verdes y Figura 5 sub-figura 3 eje vertical) y la real (Figura 5 sub-figura 2 puntos verdes y Figura 5 sub-figura 3 eje horizontal) el impacto en la clasificación de los correos es mínima (Figura 5 sub-figura 1 versus Figura 5 sub-figura 4 barras rojas y negras). Esto es debido a que, como se ha comentado en numerosas ocasiones, el número de *No-incidencias* es mucho menor y estás son las para las que le modelo hace una estimación peor que difiere de la real (Figura 5 sub-figura 3 los puntos cuando la $Probabilidad_{real} < 0.5$ y $Probabilidad_{modelo} > 0.5$ tienen muy pocos correos electrónicos).

En la Figura 6 se muestra la visualización generada por el modelo ad-hoc para un correo electrónico ejemplo:

Ticket completo:

benvolguts/des,

als efectes d'informar en un procés de recuperació de períodes d'inscripció, necessito que informeu si un usuari va accedir al web per a renovar la demanda

- aaron henriques iser. 43671394y. data aproximada 26/04/2018

cordialment

isabel domingo egea
directora oficina de treball de dante

c/argimon 10-12 | 08032 barcelona | tel. 93 4072651

isabel.domingo@gencat.cat

Es **No incidencia** y el modelo dice que es **No Incidencia**

Unigramas:
efectes informar proces recuperacio periodes inscripcio necessito informeu usuari accedir web renovar demanda aaron henriques iser data aproximada isabel domingo egea directora oficina treball dante argimon tel isabel domingo

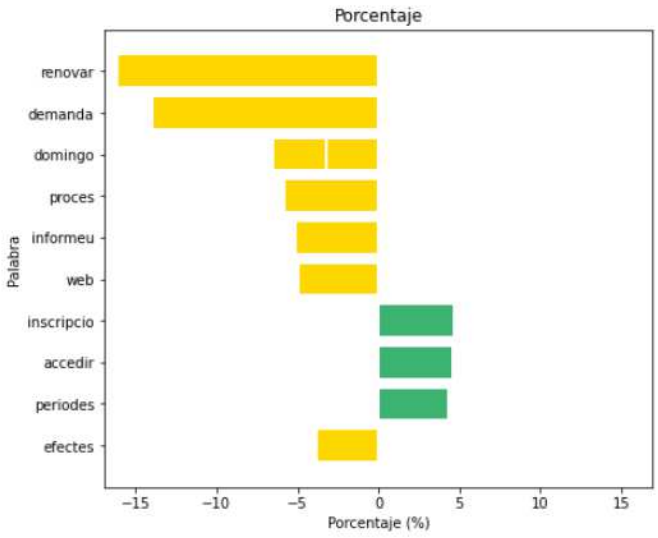


Figura 6: Visualización generada mediante el modelo ad-hoc para un correo electrónico dado que es No-incidencia y ha sido clasificado correctamente como No-incidencia.

En la parte superior se muestra un correo electrónico concreto y en la parte central, se muestran las palabras que han tenido impacto en la clasificación señaladas en verde o amarillo, en este caso particular la clasificación del correo ha sido no-incidencia. Las palabras señaladas en amarillo (verde) indica que tienen un peso negativo (positivo) y por lo tanto que contribuyen a favor de que el correo sea clasificado como *no-incidencia* (*incidencia*).

En la parte inferior aparece la contribución exacta de cada palabra, es decir el porcentaje parcial de la clasificación (la proporción particular de la suma de pesos total) que explica cada una de las palabras. En este ejemplo concreto se ha optado por visualizar solamente las 10 palabras que más contribuyen. Si una palabra aparece *N* veces en un texto se sumará su peso *N* veces, véase la palabra *domingo* en la Figura 6.

Mediante la herramienta de visualización ad-hoc se observa que para el ejemplo concreto de la Figura 6 la mayoría de las palabras tienen una contribución negativa, siendo la palabra que más contribuye *renovar*, que de hecho explica un *17%* de la clasificación.

Por otro lado, para las **regresiónes logísticas multiclase** que aparecen en los modelos ***Wheninc_v2.pkl*** y ***WhenNoninc_v2.pkl***, se aplica el mismo razonamiento, pero al ser en este caso la clasificación multiclase con 3 posibles categorías, cada palabra tendrá tres pesos asignados, tal y como puede verse en la Figura 7. Es decir, para el modelo ***Wheninc_v2.pkl***, por ejemplo, cada palabra tendrá un peso que contribuirá a favor de la clasificación *ALTRES o NO-ALTRES* (Figura 7 primer texto central señalado en verde), un segundo peso que contribuirá a favor de la clasificación *APLICACIONS o NO-APLICACIONS* (Figura 7 segundo texto central señalado en naranja*)* y un último peso a favor de la clasificación *LLOC DE TREBALL* o *NO-LLOC DE TREBALL* (Figura 7 tercer texto central señalado en azul). Cuando la palabra contribuye de manera positiva a la clasificación, la visualización ad-hoc marca la palabra en color intenso y cuando contribuye de manera negativa la marca en color claro.

Ticket completo:

de: pedro ferre galvan [pferre2@xtec.cat]
enviat el: diumenge, 6 / maig / 2018 10:06
per a: sau generalitat de catalunya
tema: re: contrasenya
hola perdoneu, faig referència a:

saga/esfer@/atri

moltes gràcies

el dia 5 de maig de 2018 a les 20:04, sau generalitat de catalunya <sau.tic@gencat.cat> ha escrit:
benvolgut,

ens podria indicar a quina contrasenya fa referència:

-correu xtec.
-saga/esfer@/atri

moltes gràcies

servei d'atenció a l'usuari, d'espai de treball i col•laboració
centre de telecomunicacions i tecnologies de la informació
tel : 900 82 82 82
sau.tic@gencat.cat
_____
de: pferregalvan@gmail.com [pferregalvan@gmail.com] en nom de pedro ferre [pferre2@xtec.cat]
enviat el: dissabte, 5 / maig / 2018 19:58
per a: sau generalitat de catalunya
tema: contrasenya
hola soc el pedro ferré galván (39678342f) i m'ha fallat la contrasenya i no hi ha manera d'entrar.
podeu recuperar-la ho dona una d'auxiliar.
gràcies
_____

Es APLICACIONS y el modelo dice que es APLICACIONS

Unigramas:
referencia escrit benvolgut podria indicar contrasenya referencia correu xtec servei atencio usuari espai treball col laboracio centre telecomunicacions tecnologies informacio tel

referencia escrit benvolgut podria indicar contrasenya referencia correu xtec servei atencio usuari espai treball col laboracio centre telecomunicacions tecnologies informacio tel

referencia escrit benvolgut podria indicar contrasenya referencia correu xtec servei atencio usuari espai treball col laboracio centre telecomunicacions tecnologies informacio tel

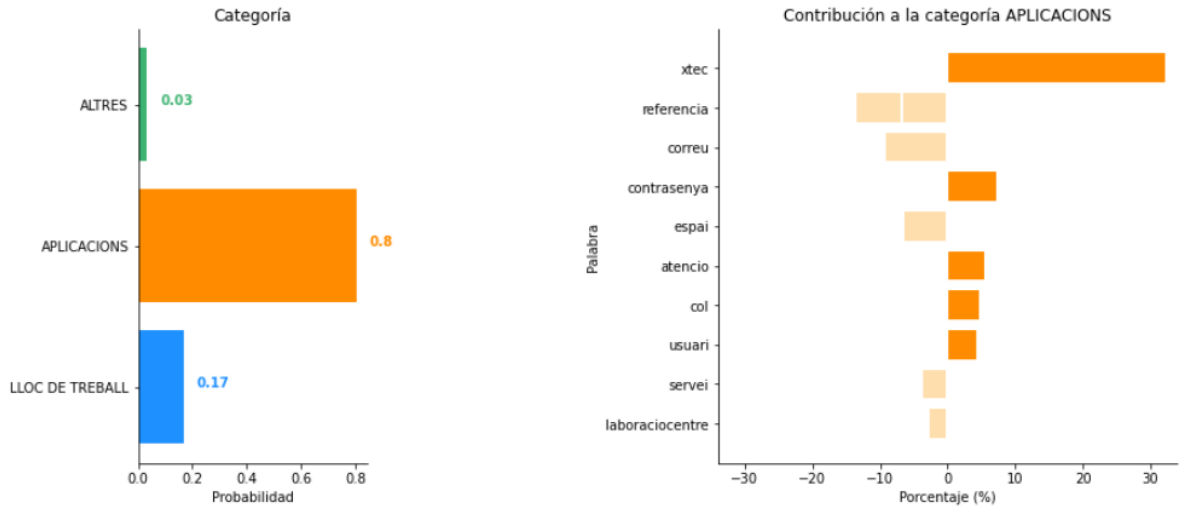


Figura 7: Visualización generada mediante el modelo ad-hoc para un correo electrónico incidencia dado que pertenece a la categoría ALTRES y ha sido clasificado correctamente como ALTRES.

La visualización ad-hoc muestra además que probabilidad muestra el correo electrónico concreto de ser clasificado en cada una de los posibles categorías (Figura 7 inferior izquierda), puesto que el modelo calcula tres gracias a los 3 diferentes pesos por palabra, en este caso concreto la probabilidad más alta es para la categoría *APLICACIONS* con un *0.8* sobre *1,* como consecuencia el correo es clasificado como *APLICACIONS*.

Al igual que se hacía para el modelo binomial, también se muestra la contribución exacta de cada palabra para la categorización más probable (Figura 7 inferior derecha). En este ejemplo concreto los pesos a favor de la clasificación *APLICACIONS* en naranja intenso y a favor de la clasificación *NO-APLICACIONS* (o lo que es lo mismo en contra de la clasificación *APLICACIONS*) en naranja claro.

Para el modelo ***WhenNoninc_v2.pkl*** la visualización para cada correo electrónico será la misma, pero en este caso las tres categorías de clasificación serán *altres*, *comunicacio* y *peticio*. A modo resumen se muestran las palabras que más contribuyen para el modelo *Wheninc_v2.pkl* (Figura 8) y para *WhenNoninc_v2.pkl* (Figura 9) para cada una de sus posibles categorías de clasificación.

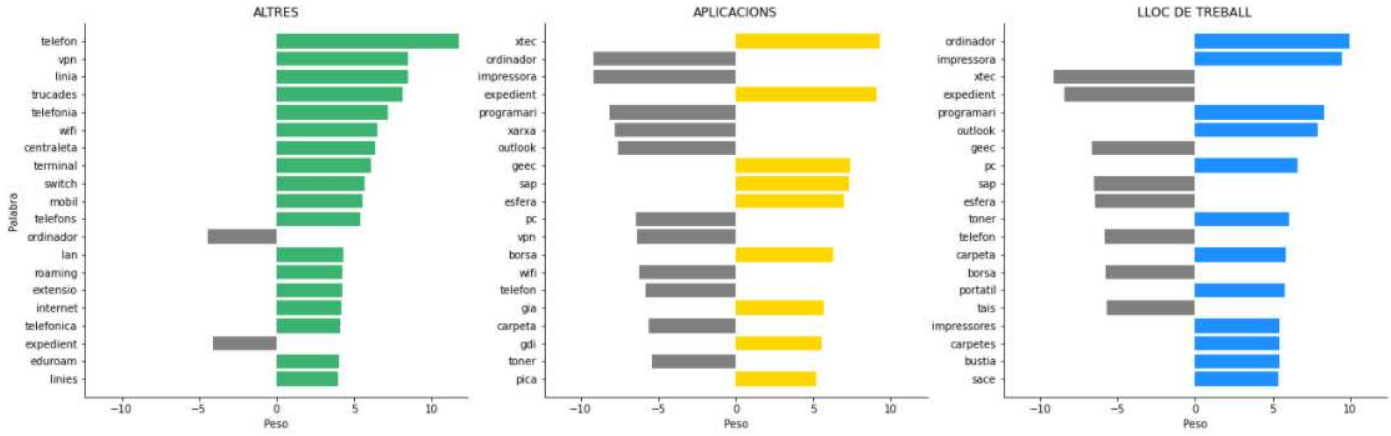


Figura 8: Representación de las 20 palabras con mayor peso para cada categoría (ALTRES, APLICACIONS y LLOC DE TREBALL) del modelo Wheninc.pkl.

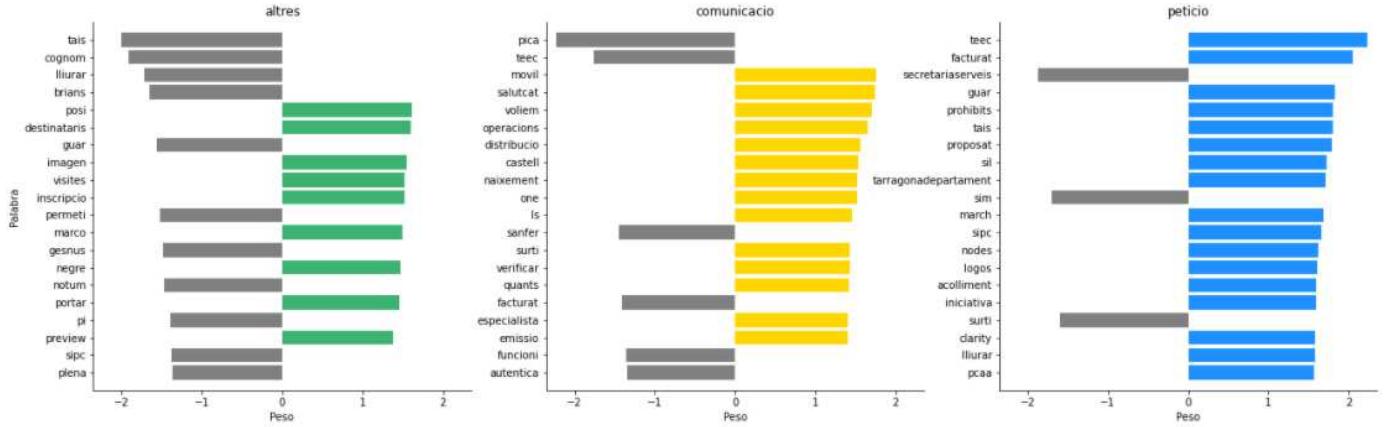


Figura 9: Representación de las 20 palabras con mayor peso para cada categoría (altres, comunicacio y peticio) del modelo WhenNoninc.pkl.

En definitiva, como ocurría para el caso de la clasificación binomial (Figura 6) **la herramienta ad-hoc ilustra de manera clara y sencilla como se ha valorado cada correo electrónico y que se ha tenido en cuenta para determinar su clasificación** multiclase.

Ya sea para las regresiones logísticas binomiales o multiclase, a la hora ajustar la regresión y obtener los pesos y el intercepto, el proveedor encargado de la ejecución del proyecto piloto del categorizador de tickets solo tuvo en cuenta los unigramas (la contribución de cada palabra por separado). Sin embargo, la herramienta de visualización ad-hoc también ha sido diseñada para representar contribución de bigramas (secuencia de dos palabras) y trigramas (secuencia de tres palabras) en caso de requerirlo.

5.2 LIME

LIME (Local Interpretable Model-Agnostic Explanations) es independiente del modelo, lo que significa que se puede aplicar a cualquier modelo de Machine Learning, e intenta comprender el modelo perturbando las muestras de datos que se introducen al modelo, para comprender cómo cambian las predicciones. LIME permite interpretar de forma local el modelo, es decir, modifica una sola muestra de datos ajustando los valores de una característica determinada, y observa el impacto resultante en la salida, permitiendo así conocer la importancia de cada variable en la predicción.

Como el modelo consta de dos etapas, donde la segunda es dependiente de la predicción efectuada en la primera etapa, no será posible aplicar LIME directamente con los modelos proporcionados en los ficheros PKL, si no que será necesario realizar algunos ajustes. De esta forma, LIME explicará dos situaciones:

- Por un lado, si el correo se clasifica como *INCIDENCIA*, explicará qué palabras del texto contribuyen a la probabilidad de que el correo pertenezca a la clase *ALTRES, APLICACIONS*, o *LLOC DE TREBALL*.
- Por otro lado, si el correo se clasifica como *NO-INCIDENCIA*, entonces la explicación tratará de resaltar aquellas palabras que más contribuyen a la clasificación del correo entre las clases *ALTRES*, *COMUNICACIÓ* o *PETICIÓ*.

Un ejemplo del primer caso se muestra en la Figura 10, que es un caso particular de un correo que se clasifica como *INCIDENCIA* y, dentro de ella, en la categoría *LLOC DE TREBALL*.

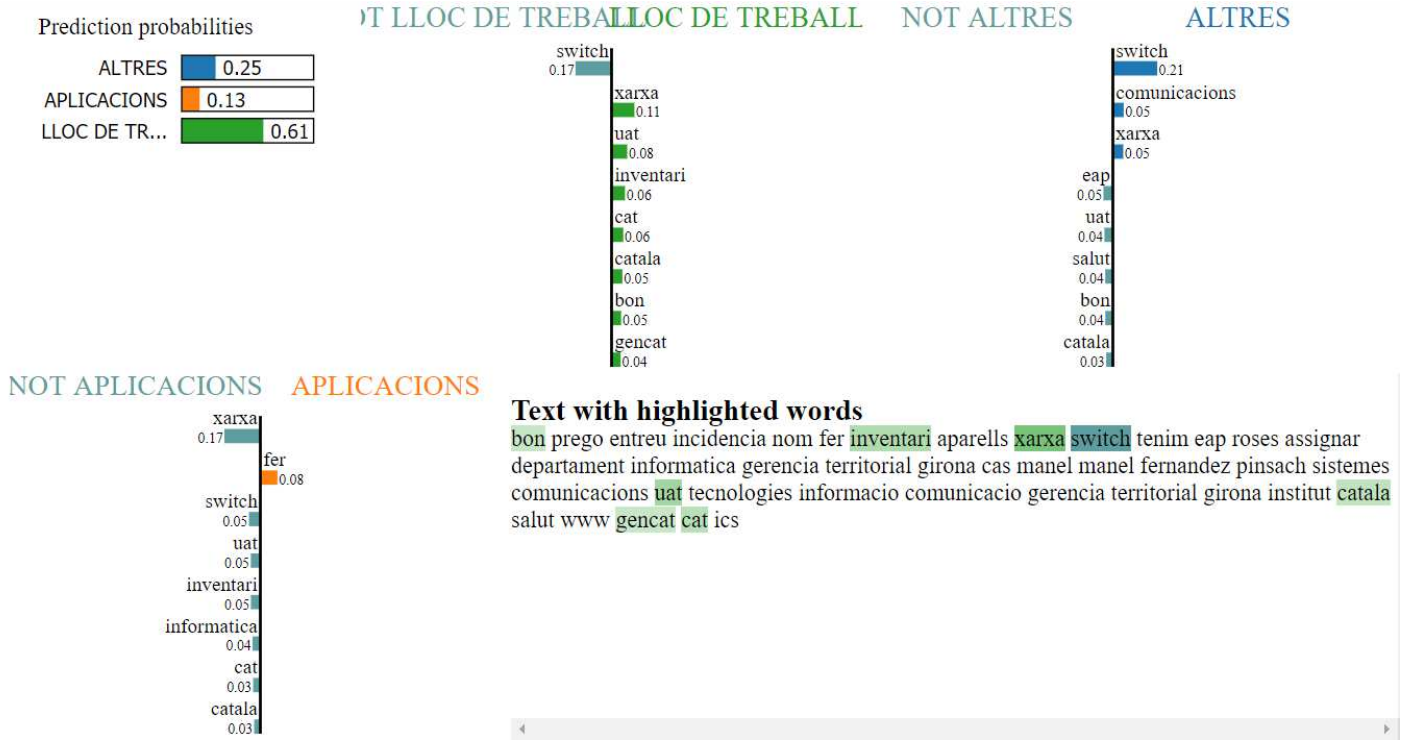


Figura 10: Ejemplo de la explicación de un correo que se clasifica como *Incidencia* en una primera etapa, y a continuación la predicción del modelo es que pertenece a la clase *LLOC DE TREBALL* con un 61% de probabilidad.

La información que nos aporta se interpreta de la siguiente forma.

Por ejemplo, la palabra ***switch*** **contribuye negativamente a** la categoría ***LLOC DE TREBALL*** **con un peso de** ***0.17*** **y a** la categoría ***Aplicacions*** **con un** ***0.05***; es decir, si se eliminase del texto la palabra *switch*, es de esperar que la probabilidad de predecir estas clases aumentase en las cantidades indicadas (entendiendo que la probabilidad varía entre *0* y *1*).

Por otra parte, la palabra ***switch*** **contribuye con un** ***0.21*** **de forma positiva a la categoría** ***ALTRES***, por lo que si se eliminase del texto de la descripción la probabilidad de predecir el correo dentro de esta clase se reduciría aproximadamente esa cantidad asociada.

Otra palabra con relevancia para el modelo es *xarxa*, ya que si se eliminase del texto de entrada se esperaría que la probabilidad de predecir el correo dentro de las clases *LLOC DE TREBALL* y *ALTRES* descendería un *0.11* y un *0.05* en cada una de ellas, mientras que la predicción de la clase *APLICACIONS* aumentaría aproximadamente un *0.17*.

Para comprobar si la interpretación de LIME es un buen reflejo de cómo funciona el modelo, en la Figura 11 **se elimina del texto de entrada la palabra** ***switch*** **y se comprueba** si **la predicción** de cada clase es el resultado esperado.

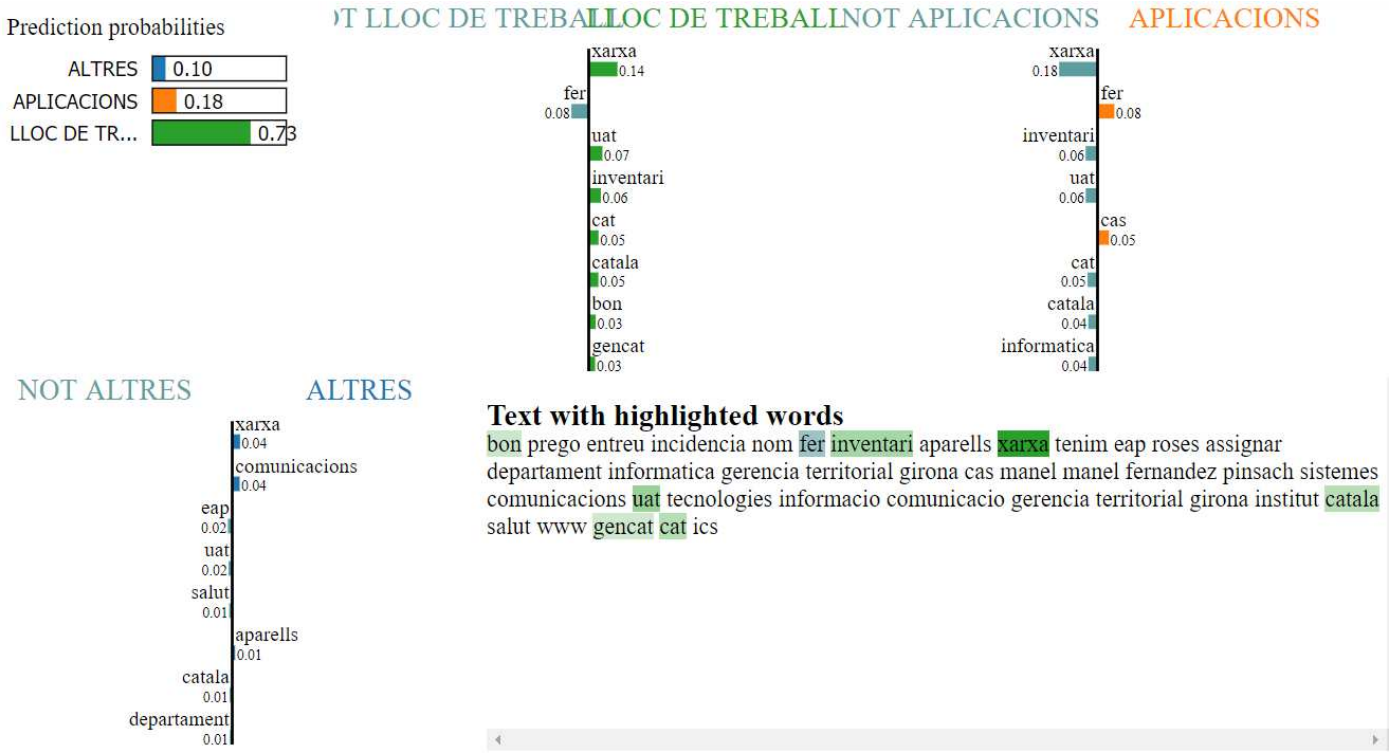


Figura 11: Mismo ejemplo que en la figura anterior pero eliminando del texto de entrada la palabra *switch*. Aunque los resultados no coinciden exactamente con los pesos que asignó LIME originalmente a esta palabra, si que apuntaron en la dirección correcta.

Comparando la Figura 10 y la Figura 11 se observa que la categoría *ALTRES* ha reducido su probabilidad de *0.25* a *0.10*, es decir, un *0.15* y no un *0.21* como apuntaba LIME en su explicación de la predicción.

Por otra parte, *APLICACIONS* ha aumentado de *0.13* a *0.18*, es decir, exactamente *0.05* tal y como LIME indicaba.

Finalmente, la probabilidad de predecir la categoría *LLOC DE TREBALL* ha aumentado de *0.61* a *0.73*, es decir, un *0.12* mientras que LIME informaba de que aumentaría la cantidad de *0.17*.

Además, para comprobar la influencia de la palabra *xarxa* se prueba a eliminar del texto original y ver los resultados.

Comparando la Figura 10 y la Figura 12, de nuevo se observa que la predicción de las probabilidades se aproximan bastante a lo que LIME nos informaba en su explicación original de la Figura 10.

Más allá de que los valores en los que cambia la predicción de cada categoría no sean exactamente los pesos que LIME asigna a cada palabra en su explicación, sí que se puede afirmar que apuntan en la dirección correcta.

Por tanto, **LIME se pueden tomar como una orientación suficientemente aproximada de cómo se alterarán las predicciones del modelo si alguna de las palabras no apareciese en el texto original**.

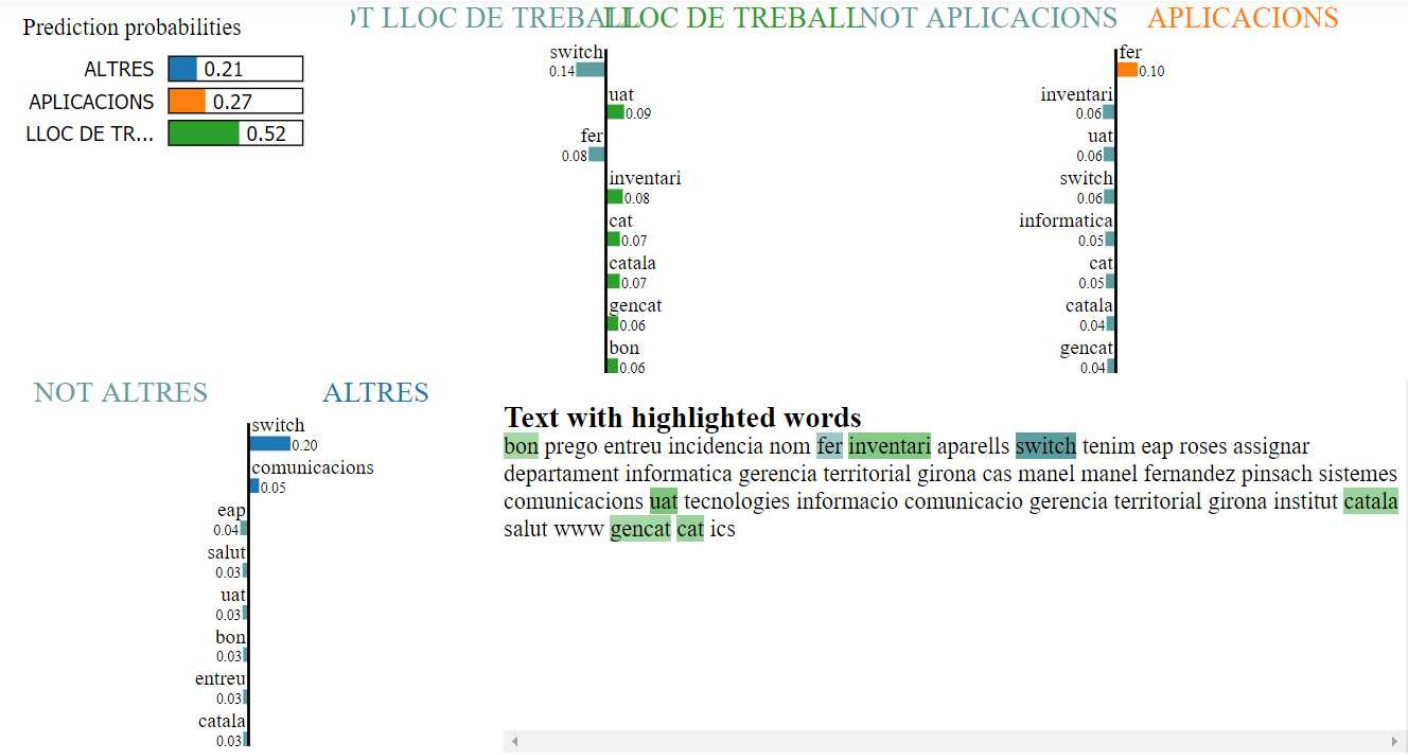


Figura 12: Mismo ejemplo que en la figura anterior pero eliminando en este caso del texto de entrada la palabra *xarxa*. De nuevo, aunque los valores no son exactos, LIME supone una buena aproximación a los resultados que se obtienen.

Un ejemplo del segundo caso, en el que un correo se clasifica como *NO-INCIDENCIA* en la primera etapa, se muestra en la Figura 13. La interpretación es semejante: en este caso concreto, las palabras más relevantes para el modelo son el término *ingresos,* que contribuye positivamente (*0.17*) a la categoría *PETICIÓ* pero negativamente (*0.20*) a la categoría *ALTRES*, y la palabra *youtube,* que si se decidiese eliminar del texto de entrada se esperaría que aumentase la probabilidad de predecir la clase *PETICIÓ* en un *0.14* pero reduciría la probabilidad de predecir *ALTRES* aproximadamente en *0.12*.

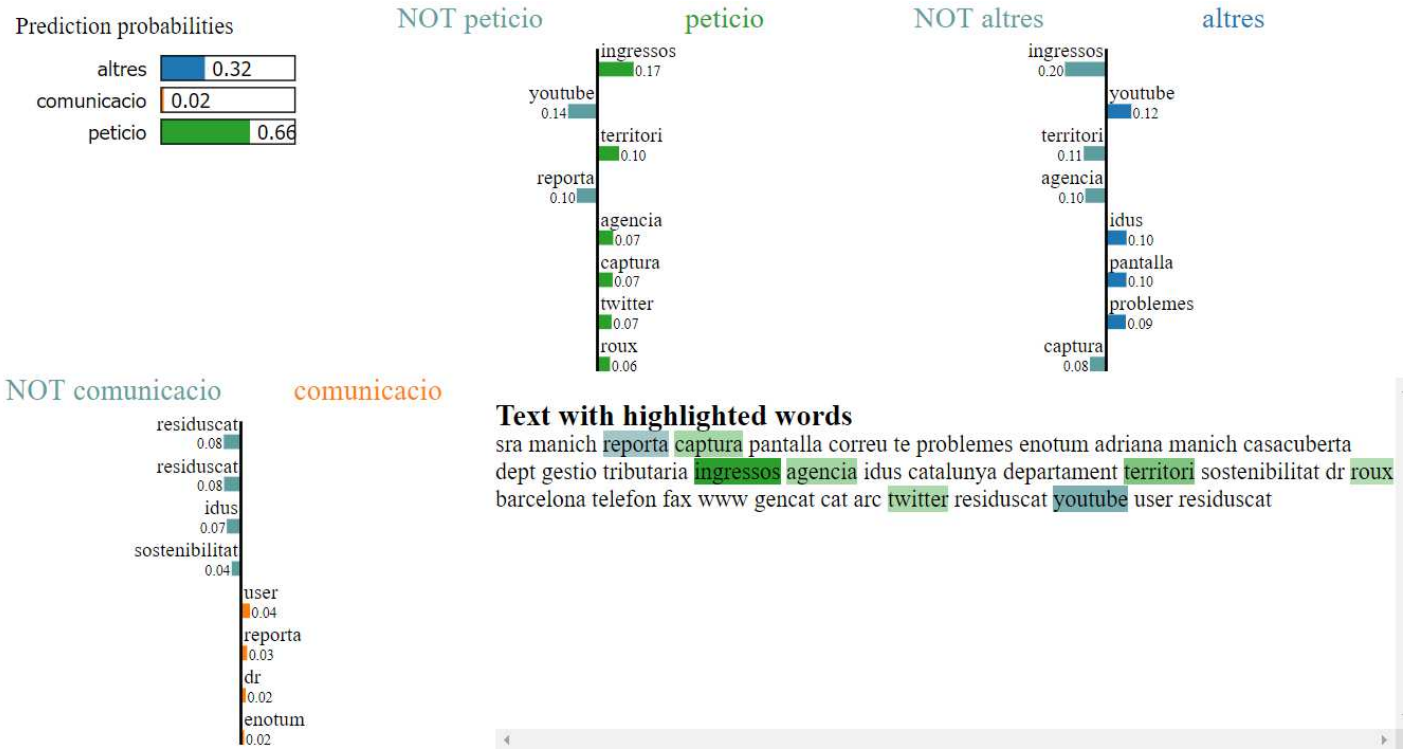


Figura 13: Ejemplo de un correo clasificado como *NO-INCIDENCIA* por la primera parte del modelo y, más adelante, clasificado con un *66%* de probabilidad dentro de la categoría *PETICIÓ*.

La principal diferencia entre LIME y la visualización Ad-Hoc descrita en el anterior apartado, es que **LIME aproxima de forma local el modelo**, y muestra por tanto la contribución de cada palabra en ese modelo aproximado, mientras **la visualización Ad-Hoc muestra la contribución exacta de cada palabra**, y no es una aproximación del modelo.

En el caso de un modelo auto explicable podría construirse un visualizador Ad-Hoc como se ha hecho para este caso, pero para el caso de modelos de tipo caja negra, esa aproximación no sería válida y de ahí la importancia de LIME a la hora de explicar de forma local ese tipo de modelos, aunque en el caso actual parezca que su aportación no ha sido grande ya que se disponía de una visualización Ad-Hoc.

5.2.1 Aplicación web para la experimentación con LIME

Con el objetivo de que los usuarios del servicio de atención al usuario SAU del CTTI puedan contrastar las explicaciones ofrecidas por LIME, con su propio conocimiento de los problemas o necesidades reflejados en los correos enviados a dicho servicio, y de cómo estos se clasifican, se ha desarrollado un aplicación web mínima, que permite **explorar el conjunto de datos** utilizado para entrenar y probar los distintos modelos seleccionados para el proyecto piloto del categorizador de tickets, y que permite **seleccionar el texto** correspondiente a un ticket y **visualizar la explicación** que LIME ofrece para las clasificaciones que los modelos han realizado.

En la siguiente figura pueden verse cómo a través de esta aplicación puede explorarse el conjunto de datos, ofreciéndose la posibilidad de ordenar las filas por el valor de las columnas, y de realizar filtros en las distintas columnas.

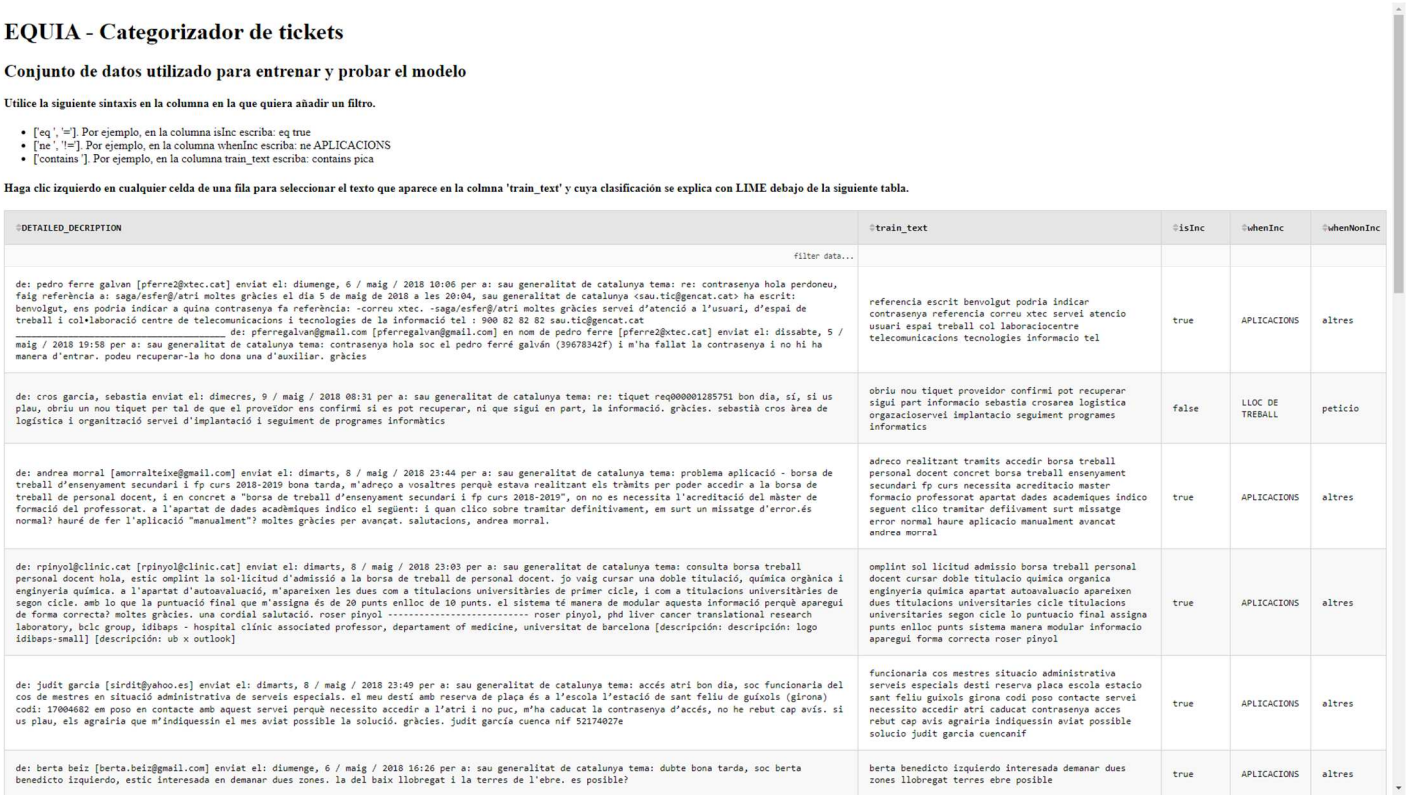


Figura 14 Aplicación web para la experimentación con LIME.

En la siguiente figura puede verse cómo se ha seleccionado el texto existente en la columna *train_text* haciendo clic izquierdo sobre la celda cuyo texto se quiere seleccionar, aunque esta selección se puede realizar haciendo clic izquierdo en cualquier celda de la fila en la que se encuentra el texto que se quiere seleccionar.

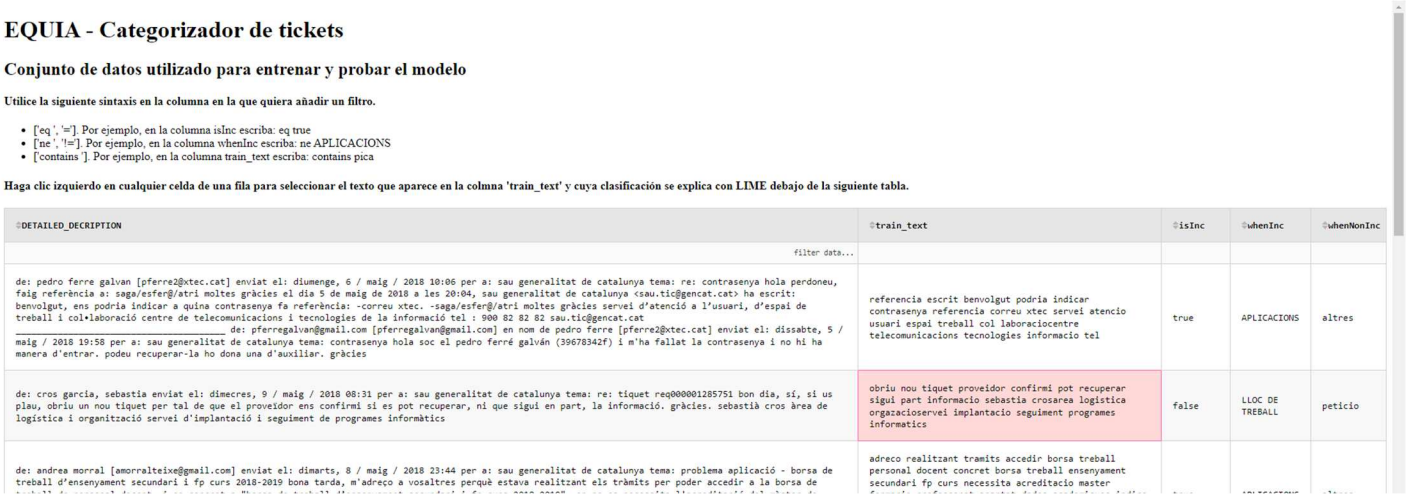


Figura 15 Selección de texto de un ticket.

En la siguiente figura puede verse el resultado de haber filtrado la columna *whenInc* por el valor ALTRES.

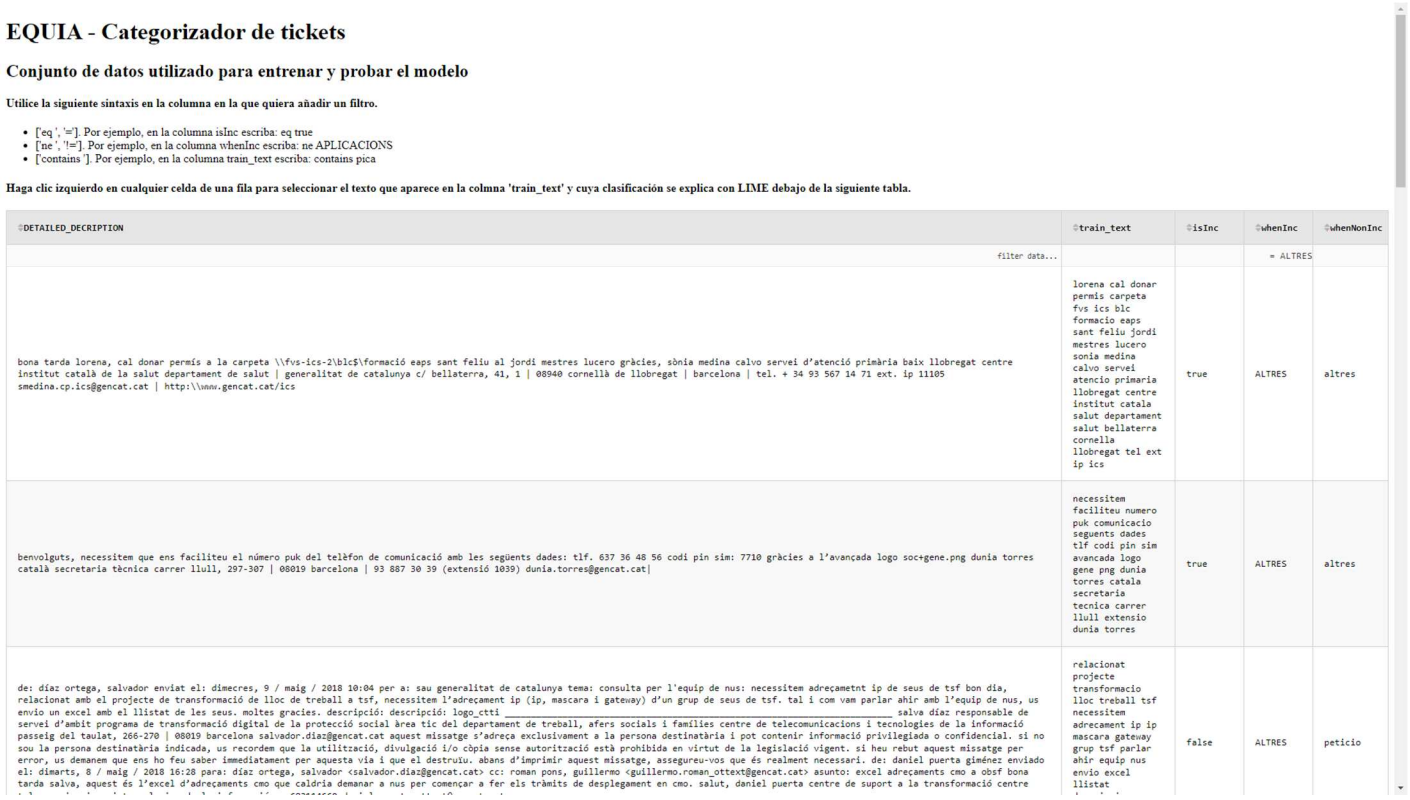
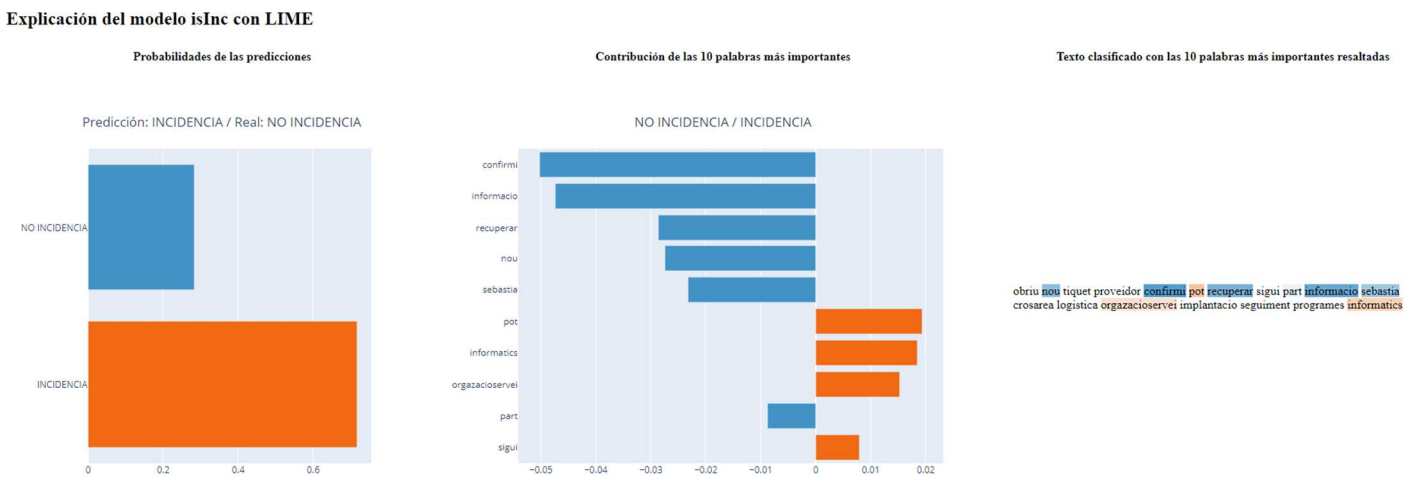


Figura 16 Filtrado de la columna whenInc por el valor ALTRES.

En la siguiente figura se pueden ver las explicaciones proporcionadas por LIME para las clasificaciones realizadas por los modelos para el texto seleccionado.



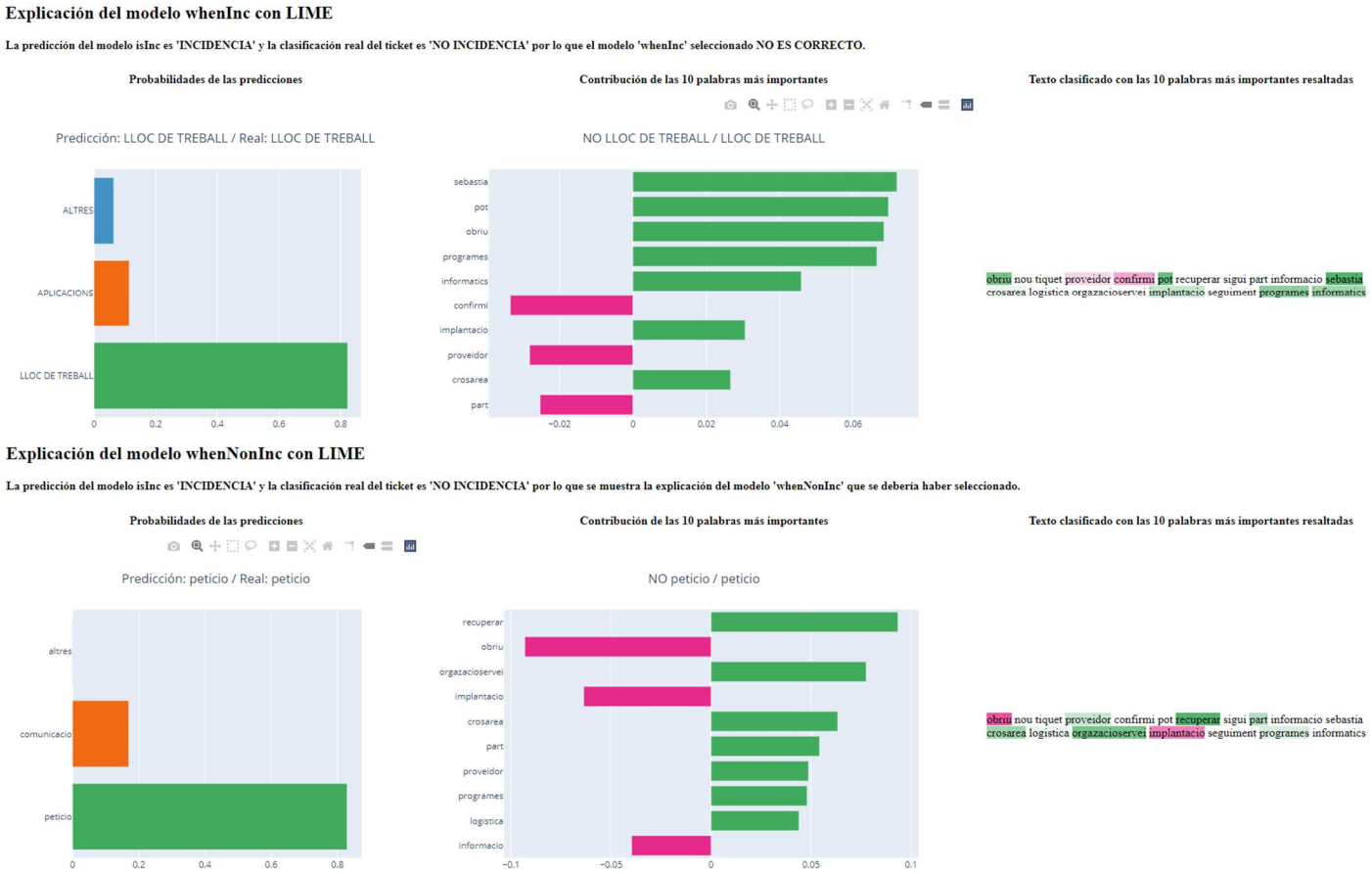


Figura 17 Explicaciones proporcionadas por LIME.

Tal y como puede verse en la figura anterior, el primer modelo *isInc* clasifica de forma errónea el ticket como INCIDENCIA, por lo que el segundo modelo *whenInc* se selecciona de forma errónea y por tanto la clasificación proporcionada por el categorizador de tickets no será correcta.

La aplicación detecta esta situación, y muestra cuál habría sido la clasificación, y su explicación, en caso de que el primer modelo *isInc* hubiera clasificado correctamente el ticket como NO INCIDENCIA y se hubiera seleccionado el model *whenNonInc*.

5.3 SHAP

Para visualizar el modelo de Regresión Logística, IsInc_v2.pkl, se ha aplicado la técnica de visualización **SHAP**. SHAP es un método diseñado para modelos de Machine Learning que explica la predicción de una observación, mostrando la **contribución de cada característica al total de la predicción**. Aplicando esta técnica al modelo de Incidencia/No incidencia, se podrá ver qué **palabras son más influyentes** en el modelo final, es decir, qué palabras son las que empujan a que un ticket sea incidencia y cuales a que no sean incidencia.

SHAP dispone de varios explicadores dependiendo del tipo del modelo. Al ser un modelo de Regresión Lineal, el explicador utilizado es **LinearExplainer.** Este explicador aprende con los datos de entrenamiento que en este caso es

el cuerpo del ticket, sin embargo, para que la máquina comprenda las palabras, éstas tienen que ser **vectorizadas,** es decir, deben codificarse como un vector de números enteros, donde la longitud del vector comprenderá el número de palabras que contenga el vocabulario. El vocabulario del modelo isInc comprende ***749.130*** **palabras**, por lo que el vector tiene esta longitud. Además, este explicador utiliza la Regresión Logística existente en el fichero IsInc_v2.pkl.

Para el cálculo de los shap values, que son los valores con la importancia de cada palabra, se tiene en cuenta el explicador mencionado anteriormente (LinearExplainer), y los valores de *X_test* vectorizados. LinearExplainer calcula los shap values restando a cada elemento del vector la media de los datos y multiplicándolo por cada coeficiente de la Regresión Lineal. Tanto la longitud del vector como el número de coeficientes es de ***749.130***, por lo tanto, **debido al alto coste computacional que supone este cálculo, solo se ha podido explicar conjuntamente hasta *4000* registros de *X_test*** de los casi *20.000* (*4000* tickets – es un *20%* de los registros de X_test y un *3%* de los datos finales).

Los resultados que se muestran a continuación **no** están **basados** en **todos los registros** de *X_test* necesarios para la explicación global del modelo, sino solamente en *4.000*, por lo tanto, **no se pueden sacar conclusiones finales.**

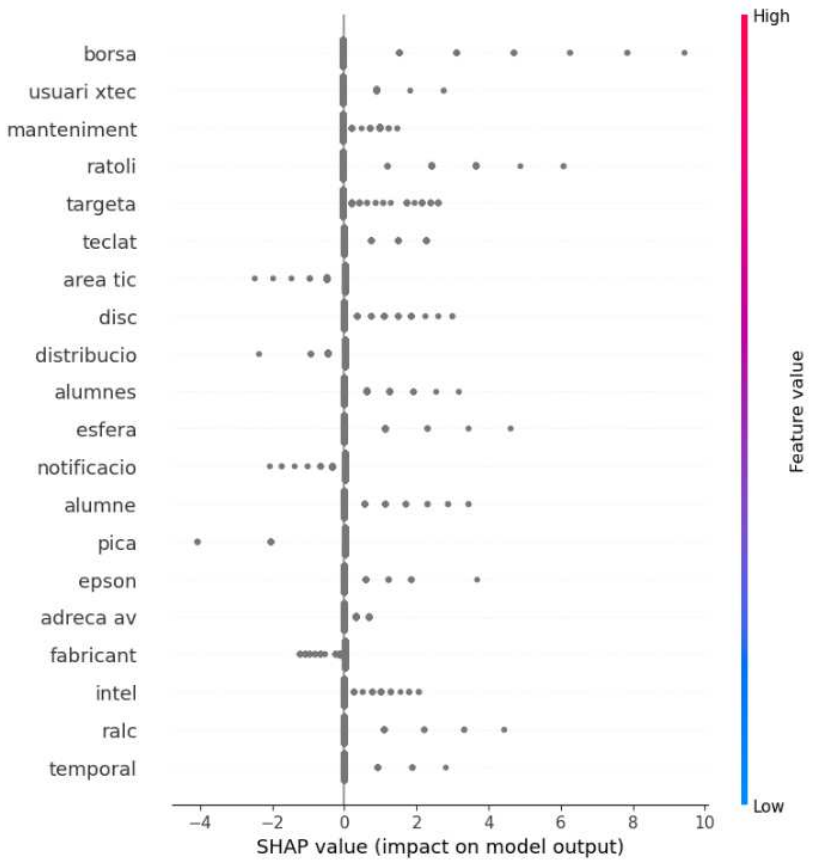


Figura 18: Summary Plot de Shap Values con 4.000 tickets.

Explicando *4.000* tickets de *X_test*, las palabras que más influyen para que un ticket sea incidencia es **"borsa", "usuari xtec", "manteniment" y "ratoli".** Por el contrario, las palabras que más influyen para que un ticket sea no incidencia son, "area tic", "distribucio", "notificacio", "pica" y "fabricant".

5.4 WHAT-IF

WHAT IF es una técnica de visualización interactiva que permite interactuar con los datos, modificándolos y así poder ver como esa modificación afecta a la decisión tomada por el modelo. Esta herramienta ha sido utilizada para visualizar y evaluar los resultados de la clasificación ofrecida por los modelos: *isInc, whenInc* y *whenNonInc*.

WHAT IF muestra para todos los modelos la **clasificación** de cada ticket presente en los datos de test, el **rendimiento** del modelo utilizando **matrices** de **confusión**, **gráficas *precisión-recall***, curvas **ROC** y las métricas ***accuracy*** y ***F1 score***. Debido a que el proveedor no determinó cuales eran los datos test, primeramente se procedió a reentrenar **cada uno** de los **modelos**, utilizando el ***80%*** de los datos para el **entrenamiento** y el ***20%*** restante para el **test**, y así obtener medidas de rendimiento lo más veraces posible.

Al abrir la herramienta se muestra un panel inicial *Datapoint editor,* en el que aparecen puntos que representan un ticket y la clasificación del mismo (Figura 19).

Para el modelo *isInc*, los resultados de clasificación que aparecen en el primer panel son los mostrados en la Figura 19. La clasificación será incidencia, en azul o no incidencia, en rojo.

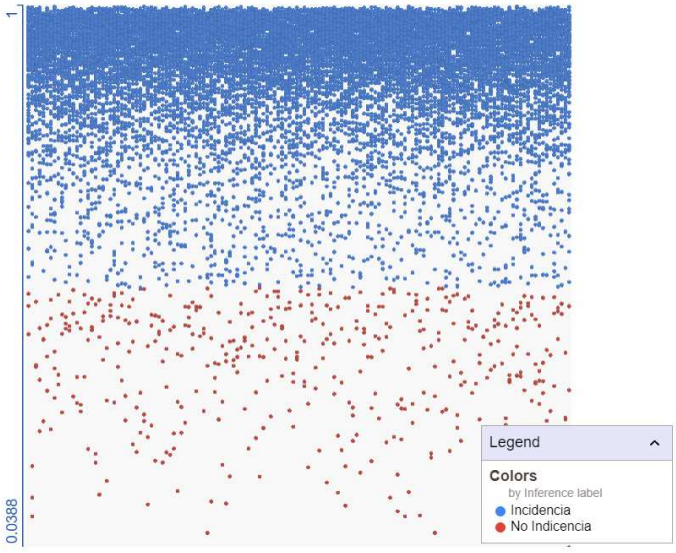


Figura 19: Resultados de clasificación modelo isInc.

Este panel permite además mostrar lo que se conoce como *nearest counterfactual* que, en este caso, sería el ticket más similar al ticket seleccionado pero que pertenece a otra de las posibles clases. La Figura 20 muestra el *nearest counterfactual* para un ticket clasificado como incidencia.

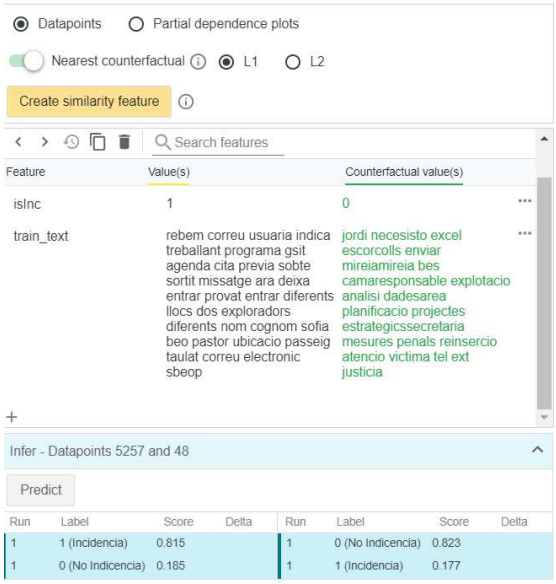


Figura 20: Nearest counterfactual de una incidencia.

También permite cambiar palabras del ticket y realizar una predicción para ese ticket, facilitando la comprensión del modelo al mostrar las probabilidades que tiene el ticket de pertenecer a una clase u otra, en función de las palabras que se hayan cambiado. Cuando en el ejemplo de la Figura 20 se eliminan las palabras *correu* y *electronic* la probabilidade de que el ticket pertenezca a la clase incidencia se incrementa en un *0.008058* (Figura 21), por lo que eliminar esas palabras, hace que la clasificación del ticket tienda aún más a la clase incidencia disminuyendo las probabilidades de pertenecer a la clase no incidencia.

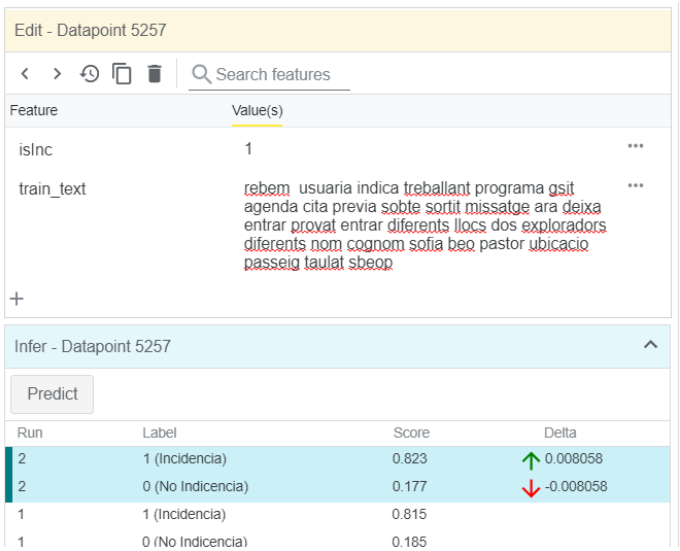


Figura 21: Nueva predicción para una incidencia.

Un segundo panel llamado *Performance & Fairness* muestra las matrices de confusión, las curvas *precision-recall* y ROC, y el *accuracy* y el *F1 score* del modelo, pudiendo desglosar incluso estos resultados clase a clase (Figura 22). Por

último, el panel *Features* muestra una descripción estadística de los datos, mostrando la media, la desviación estándar, el número de ceros, etc.

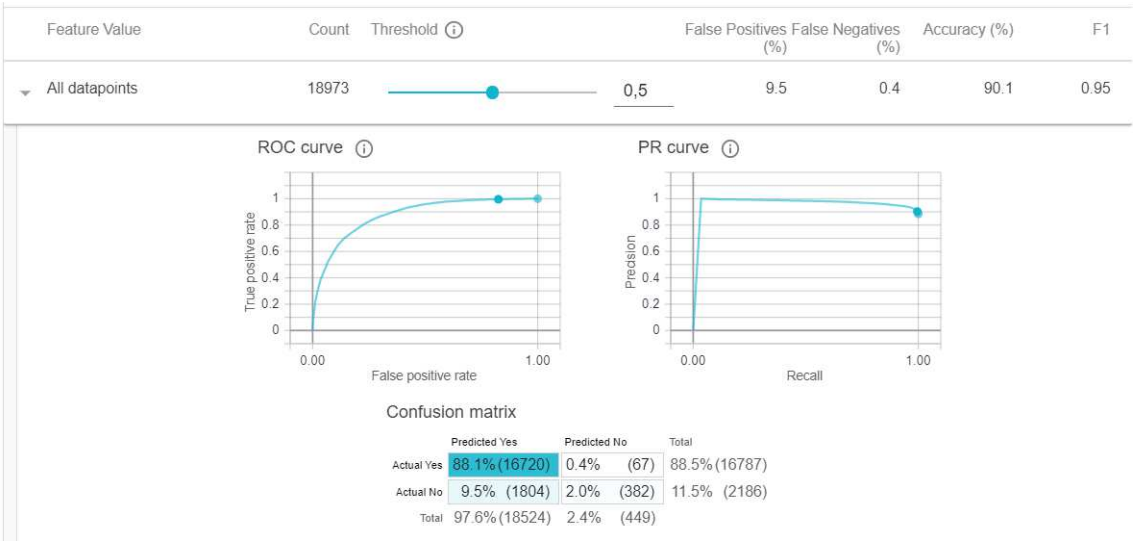


Figura 22: Resultados para el modelo isInc.

Los resultados de comportamiento del modelo son los representados en la Figura 22. Se observa un ***accuracy*** del ***90,1%*** y un ***F1 Score*** del ***0,95*** con un desequilibrio claro entre las incidencias y no incidencias que puede observarse en la matriz de confusión, *16.720 Incidencias* frente a *2.186 No Incidencias*.

Si se realiza un desglose del comportamiento por clase, se observa que la ***accuracy*** obtenida para la clase ***Incidencia (1)*** es del ***99,6%*** con un *F1 Score* de *1*. Para la clase **no incidencia (0)**, la ***accuracy*** es del ***17,5%*** y el ***F1 Score*** de ***0***.

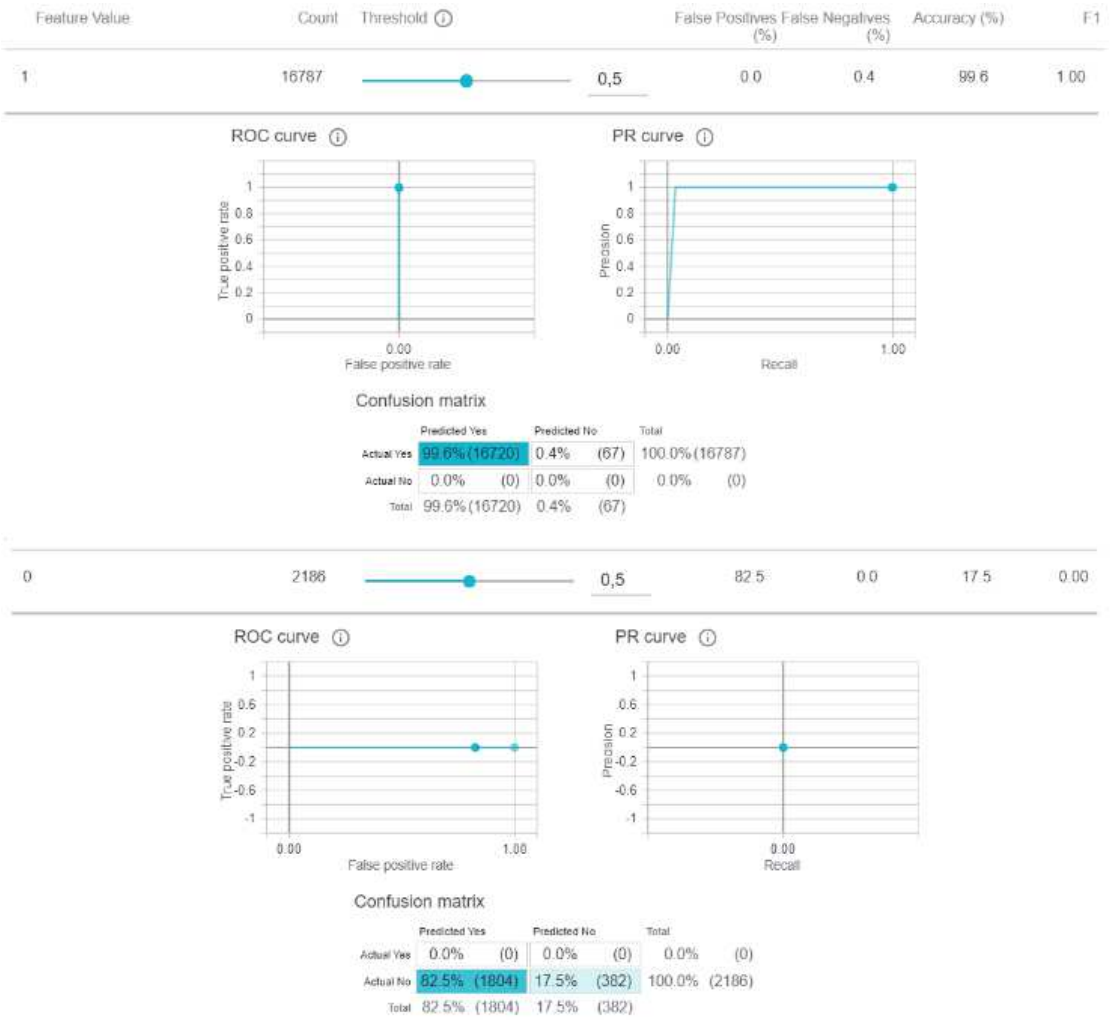


Figura 23: Comportamiento del modelo desglosado por clase.

En el caso del modelo ***whenInc***, los resultados de clasificación obtenidos son los mostrados en la Figura 24. Se realiza una clasificación de las incidencias en las categorías: *ALTRES, APLICACIONS* y *LLOC DE TREBALL*.

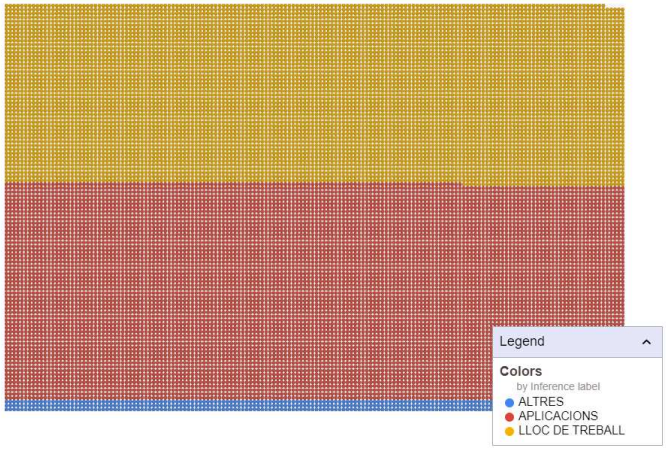


Figura 24: Resultados clasificación modelo whenInc.

El comportamiento del modelo es el mostrado en la Figura 25. Presenta una *accuracy* del 88,2%. En la matriz de confusión se observa que la clase que mejor predice es *APLICACIONS* (1), seguida de *LLOC DE TREBALL* (3) y, por último, para la que se comporta peor, *ALTRES* (0). Esto también puede verse en la Figura 26, donde se muestran los resultados por clase.

Figura 25: Comportamiento modelo whenInc.

Para la clase ***APLICACIONS*** se obtiene un ***accuracy*** del ***93,7%***, para ***LLOC DE TREBALL*** el ***87,1%*** y, por último, para ***ALTRES*** el ***47,1%*** siendo esta última la que peores resultados obtiene.

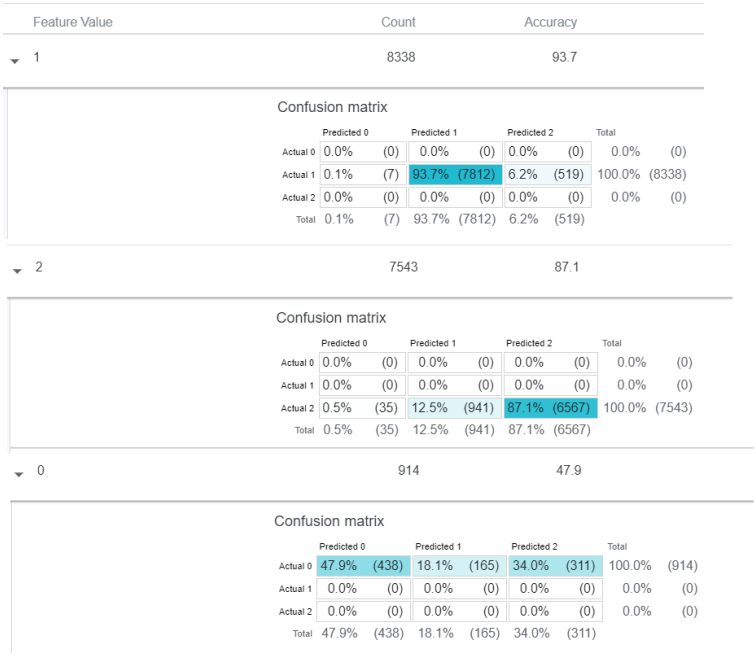


Figura 26: Resultados del modelo whenInc desglosados por clase.

En el caso del modelo ***whenNonInc***, los resultados de clasificación mostrados en la pestaña *datapoint editor* son los representados en la Figura 27. Aquí la clasificación se realiza entre las clases *ALTRES, COMUNICACIÓ* y *PETICIÓ.*

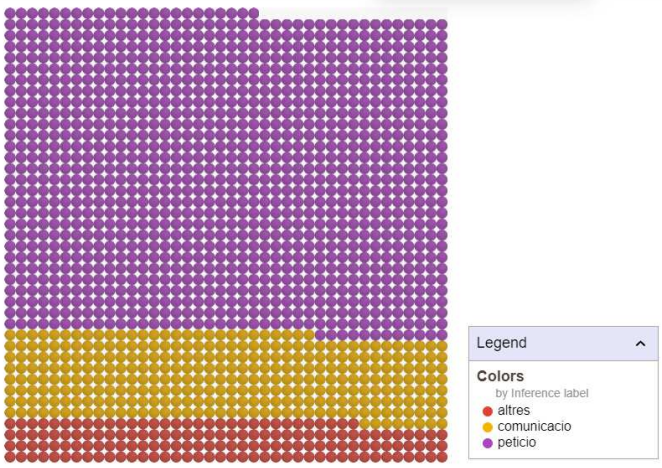


Figura 27: Resultados de clasificación del modelo whenNonInc.

Los resultados en cuanto al comportamiento del modelo son los siguientes. Presenta una *accuracy* del *83%* y las clases para las que se comporta mejor son *PETICIÓ* seguida de *COMUNICACIÓ*. La clase que peores resultados obtiene es *ALTRES*, donde se predice correctamente el *1%* de los *tickets*.

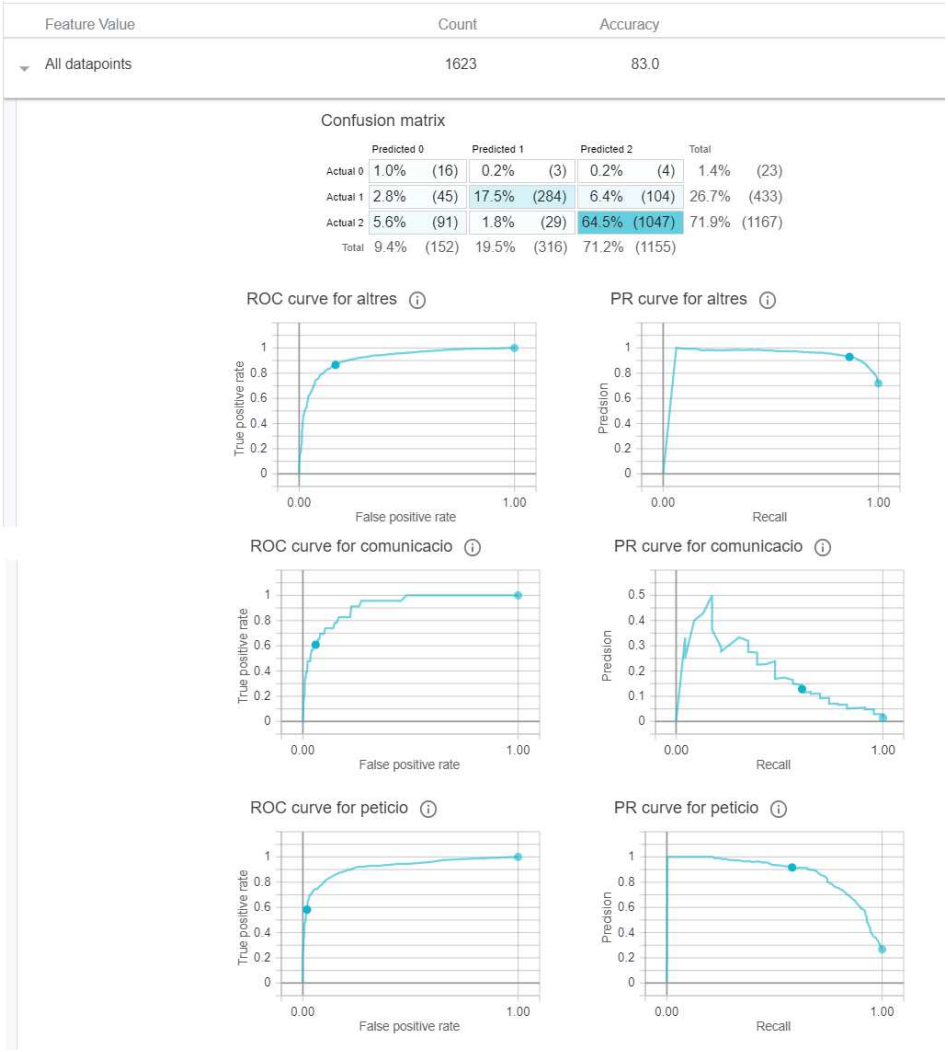


Figura 28: Comportamiento del modelo whenNonInc.

La Figura 29 muestra el comportamiento del modelo para cada una de las tres clases a clasificar. Al desglosar estos resultados por clases, la que mejor resultado obtiene es *PETICIÓ* (2), con un *89,7% accuracy*. La siguiente clase con mejor resultado es *COMUNICACIÓ* (1) con un *65,6%* de *accuracy*, mientras que la que peores resultados obtiene es *ALTRES* (0), con un *69,6%.*

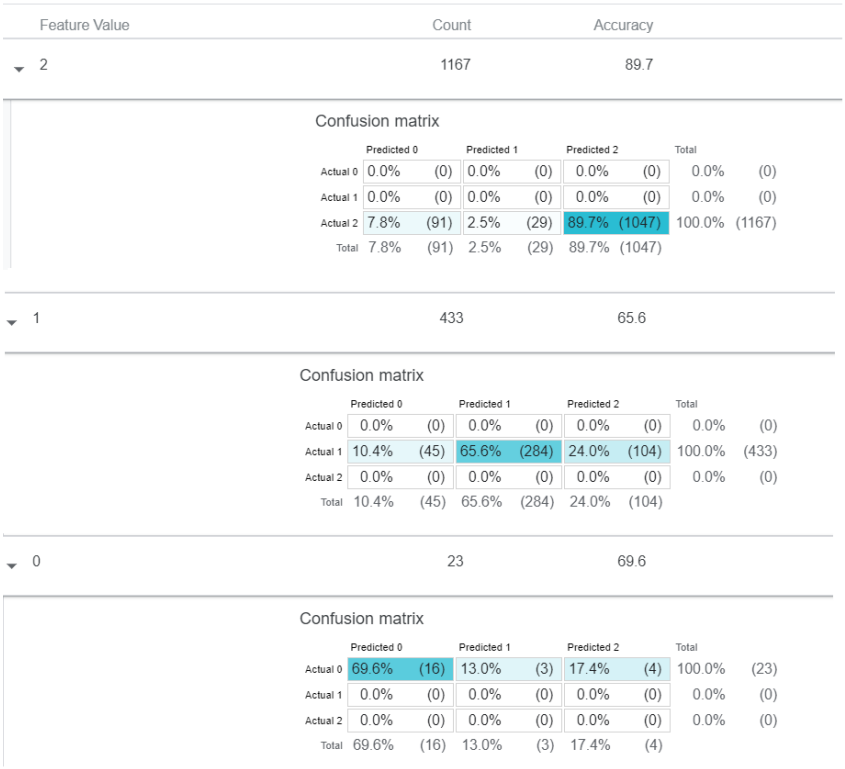


Figura 29: Resultados del modelo whenNonInc desglosados por clase.

Como se ha observado, esta herramienta permite visualizar tanto los resultados de clasificación como el rendimiento del modelo de forma rápida y sencilla y permite realizar nuevas predicciones sobre un ticket cambiando algunas palabras. Esto ayudará al usuario a comprender el funcionamiento interno del modelo ya que, al realizar un cambio y generar una nueva predicción, se mostrarán tanto las probabilidades de la incidencia original como las de esta nueva predicción.

En cuanto al rendimiento del modelo, se muestran las matrices de confusión que ayudan a comprender los resultados de clasificación y como consecuencia la validez del modelo. No sólo se muestra el rendimiento general del modelo, que normalmente se ofrece facilitando dos métricas: *accuracy* y *F1 Score*, sino que también muestra el rendimiento del modelo para cada clase particular.

Además, permite interactuar con el propio modelo cambiando sus umbrales de clasificación y alguno de sus parámetros, presentando los nuevos resultados de manera inmediata.

Una de las funcionalidades más potentes proporcionadas por What-If es la de calcular el *nearest counterfactual* de una predicción, con lo que se podrá explicar dicha predicción basándose en el cambio más pequeño en los valores de una o varias características, que hace que cambie la predicción. Este tipo de **explicación basada en el *contrafactual*** es independiente del modelo, ya que solo utiliza las entradas y salidas del modelo para calcular el *contractual,* y para contrastarlo con la entrada actual.

Si por ejemplo, se quiere utilizar una explicación basada en el *contrafactual*, para explicar por qué una solicitud de préstamo es rechazada por un modelo de Machine Learning de un banco, y cómo podrían mejorarse las posibilidades de obtener un préstamo, se puede formular la pregunta:

¿Cuál es el cambio más pequeño en las características (ingresos, número de tarjetas de crédito, edad, etc.) que cambiaría la predicción de rechazada a aprobada?

Algunas de las posibles respuestas podrían ser:

- Si el solicitante o la solicitante, ganara 10.000 € más al año, obtendría el préstamo.
- Si el solicitante o la solicitante, tuviera menos tarjetas de crédito y no hubiera incumplido con un préstamo hace 5 años, obtendría el préstamo.

Puede verse por tanto, como mediante el *contrafactual* puede proporcionarse información útil al usuario para saber **por qué** se ha tomado una decisión, y **qué puede hacer** para cambiarla.

6 Equidad

Este apartado trata de evaluar el modelo con el objetivo de concluir si se ha detectado sesgo en él. Por una parte, se evaluará el modelo general entre incidencias y no incidencias, y después éstas, por los campos más susceptibles de discriminación como es el caso del idioma del ticket y la compañía que interpone el ticket.

6.1 Análisis general del modelo

En los siguientes apartados, se mostrarán los resultados obtenidos en la evaluación del rendimiento general para cada modelo del categorizador de tickets en sus diferentes etapas.

Debido a que el proveedor no especificó cuál había sido el set de datos de entrenamiento específico se reentreno el modelo utilizando un *80%* de los datos y evaluando en el *20%* restante. Este análisis se repitió *1.000 veces* utilizando cada vez un *80%* de datos diferente y después se promedio entre todas las repeticiones, es decir se utilizó la metodología bootstrap sin remplazamiento, con el fin de que el resultado no dependiese de un modelo concreto y fuese lo más general y veraz posible.

Para cada uno de los modelos se mostrará: (1) Una matriz de confusión con el promedio de los resultados de las *1.000* repeticiones y su desviación estándar, (2) la versión normalizada de esta misma matriz de confusión, donde el valor de la suma de todas sus celdas es *100,* es decir, se ofrecen los valores como porcentaje del total, y (3) una gráfica de barras que resume la información de la matriz de confusión y presenta de forma visual algunas métricas calculadas (*precision*, *recall* y *f1-score)* para cada una de las clases. Al igual que con las matrices de confusión, como se calculan *1.000* iteraciones, la altura de la barra se corresponde con el valor medio y el intervalo de confianza es la desviación estándar.

El conjunto de datos aquí utilizado ha sido *preprocessed_ground_truth_filtered_train_text.csv,* del cual se emplea la columna *train_text* como texto de entrada para alimentar cada uno de los modelos siguientes.

6.1.1 Modelo isInc

Esta parte del modelo se encarga, en una primera etapa, de identificar si un correo es *Incidencia* o *No-incidencia*. Este modelo se carga con el fichero pickle *isInc_v2.pkl* y se entrena con la columna *IsInc*.

La matriz de confusión expone claramente que **el modelo tiende a clasificar como *Incidencia* la mayoría de los correos** (Figura 30), es decir, un ***97.66%*** (*90.64 + 7.02 = 97.66,* Figura 30 derecha), como consecuencia las métricas *precisión*, *recall* y *f1-score* son altas para la clase *Incidencia* (ver barras de color verde de la Figura 31).

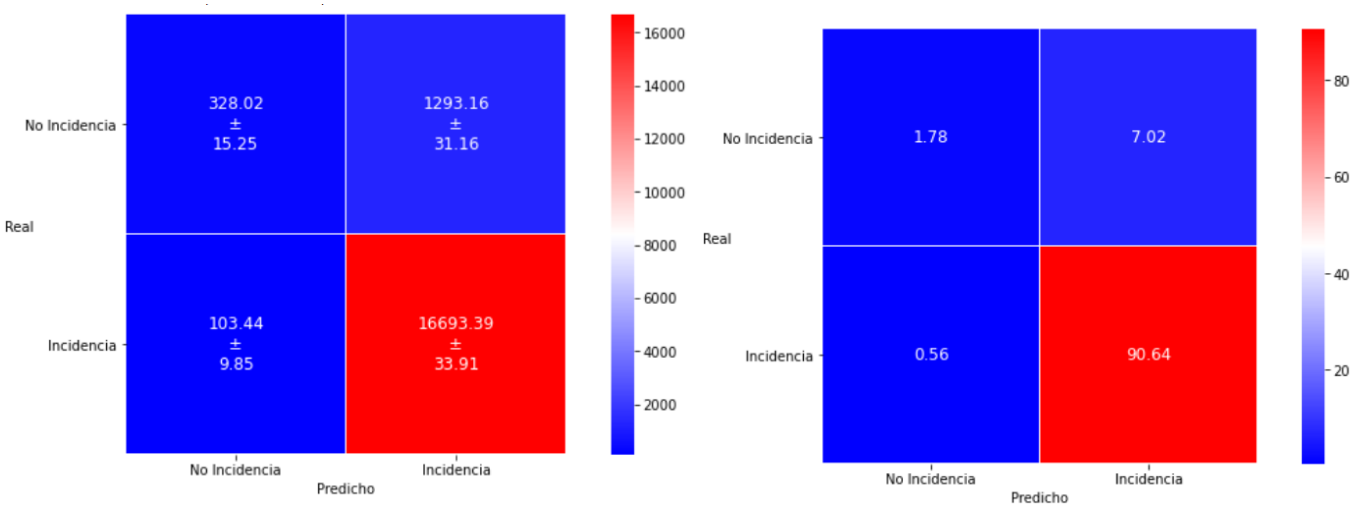


Figura 30: Matrices de confusión absoluta (izquierda) y normalizada (derecha) del modelo isInc.

Además, también se comprueba que **el modelo comete un error importante al clasificar correctamente la clase *No-incidencia*** (Figura 30 derecha); en concreto, **tan solo clasifica correctamente el *20%*** de las *No-incidencias* (*1.78 / (1.78 + 7.02) = 0.20*), lo cual también se refleja en *recall* dentro de la clase *No-incidencia* (Figura 31).

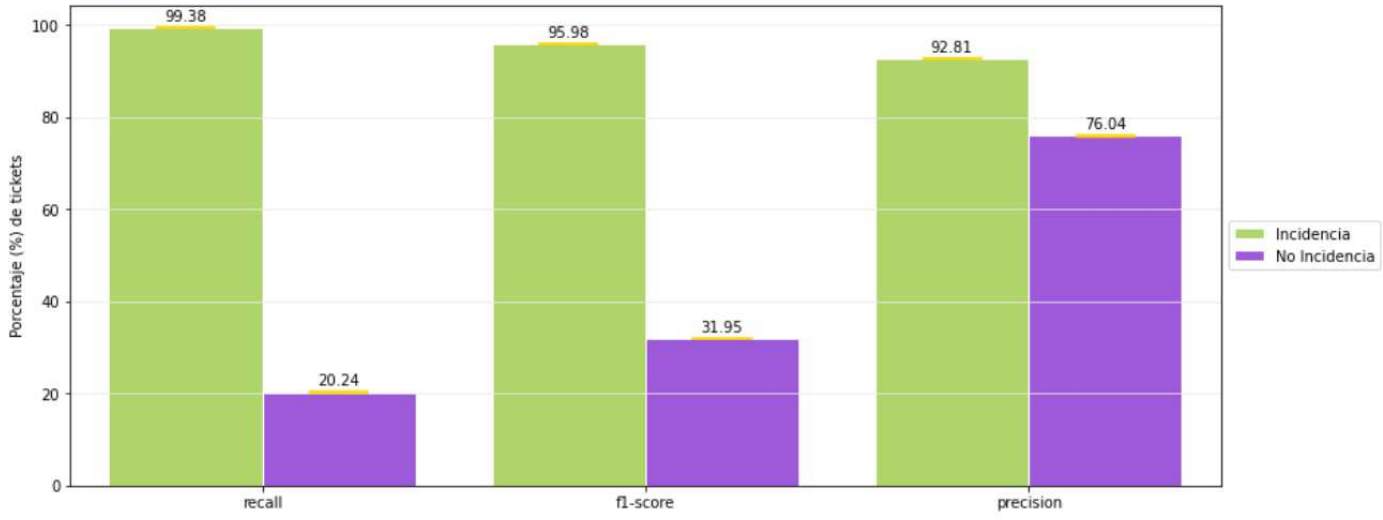


Figura 31: Cálculo de las métricas para el modelo isInc.

Finalmente, los intervalos de confianza (Figura 30 y Figura 31) tienen una amplitud mínima, lo cual indica que el modelo se muestra estable frente a las diferentes particiones de datos realizadas (entrenamiento y test) durante la metodología bootstrap y por lo tanto los resultados son extrapolables al modelo desarrollado por el proveedor.

En definitiva, las figuras nos informan de que **el modelo en cuestión está sesgado hacia la clase mayoritaria, *Incidencia***, y por esta razón se consiguen buenos resultados dentro de ella descuidándose en rendimiento del modelo en la otra clase (véase un *95.98* de F1-score de la clase *Incidencia* frente a un *31.95* dentro de la clase *No-incidencia*).

6.1.2 Modelo whenInc

Este modelo se carga con el fichero pickle *whenInc_v2.pkl* y se entrena con la columna *whenInc*. En esta segunda parte, una vez que el correo se ha clasificado como *incidencia*, el modelo trata de clasificarlo dentro de alguna de las tres siguientes categorías: *ALTRES*, *APLICACIONS* o *LLOC DE TREBALL*.

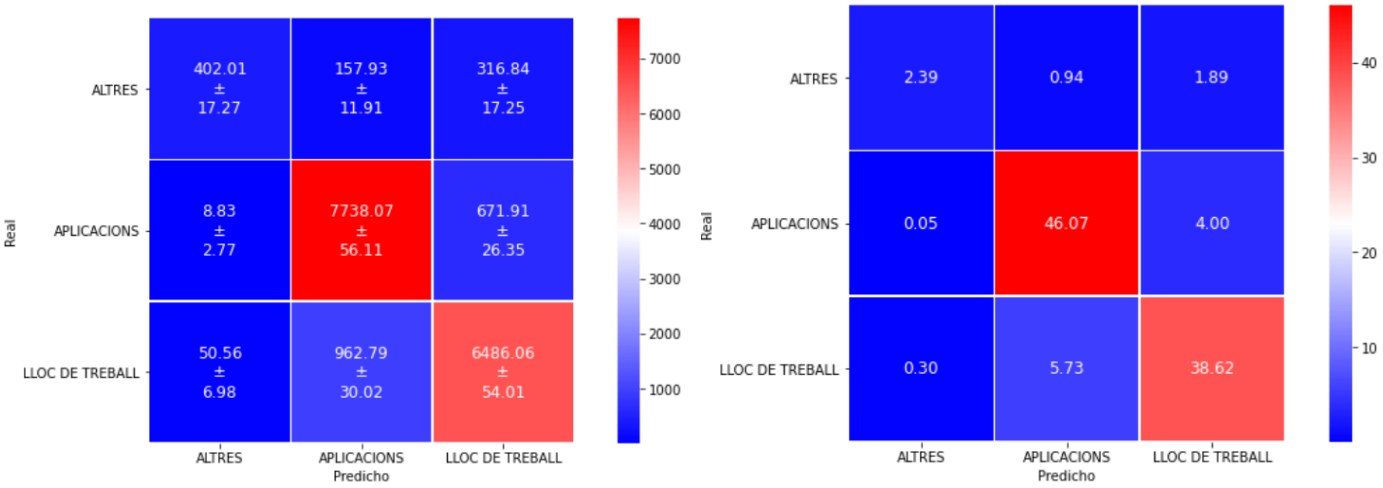


Figura 32: Matrices de confusión absoluta (izquierda) y normalizada (derecha) del modelo whenInc.

A partir de la matriz de confusión de valores absolutos (Figura 32 gráfica izquierda), y al igual que se informaba en el análisis exploratorio, se observa que la clase *ALTRES* es minoritaria respecto al resto. Este menor número genera que el modelo tienda a confundir tickets que originalmente pertenecen a esta clase y las clasifique en otra, dando lugar así a un valor bajo del recall (*45.86%*), es decir, algo más de la mitad de tickets que originalmente pertenecen a la clase *ALTRES* se clasificarán mal en otras categorías.

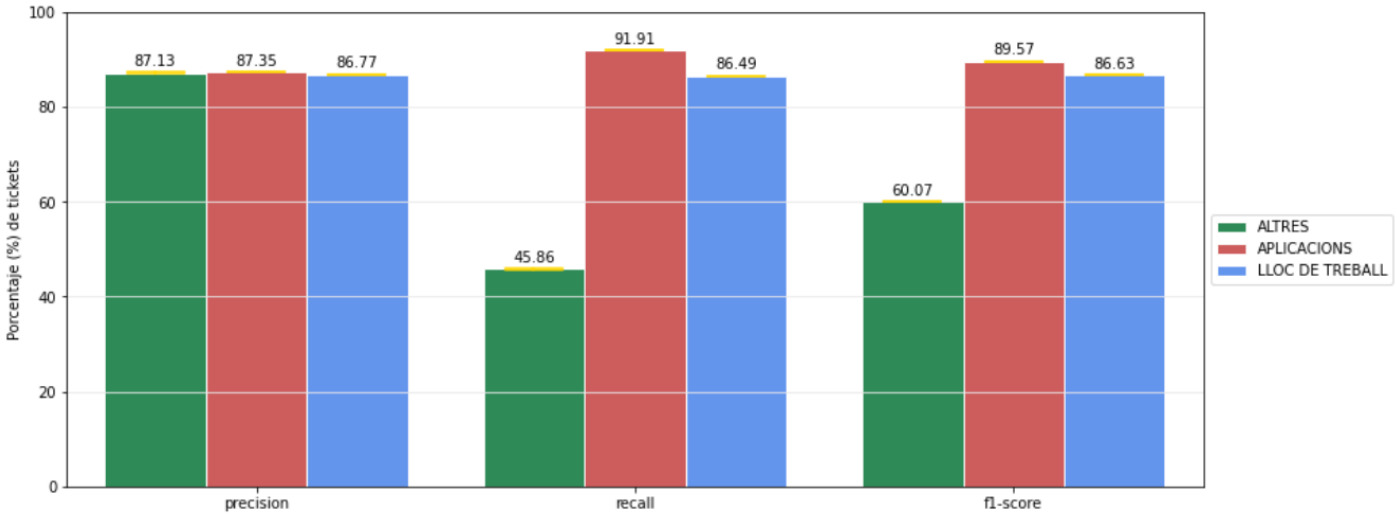


Figura 33: Cálculo de las métricas para el modelo whenInc.

La ligera mejora en el recall de la clase *APLICACIONS* respecto a la clase *LLOC DE TREBALL* no parece que tenga la suficiente relevancia como para constituir una situación de tratamiento desigual (Figura 32). Por lo tanto, **la clase minoritaria *ALTRES* recibe un tratamiento desigual** por parte del modelo **en comparación con el resto de clases.**

6.1.3 Modelo whenNonInc

Este modelo se carga con el fichero pickle *whenNonInc_v2.pkl* y se entrena con la columna w*henNonInc*. En esta segunda etapa, una vez que el correo se ha clasificado como *No-incidencia*, el modelo trata de clasificarlo dentro de alguna de las siguientes categorías: *ALTRES*, *COMUNICACIÓ* o *PETICIÓ*.

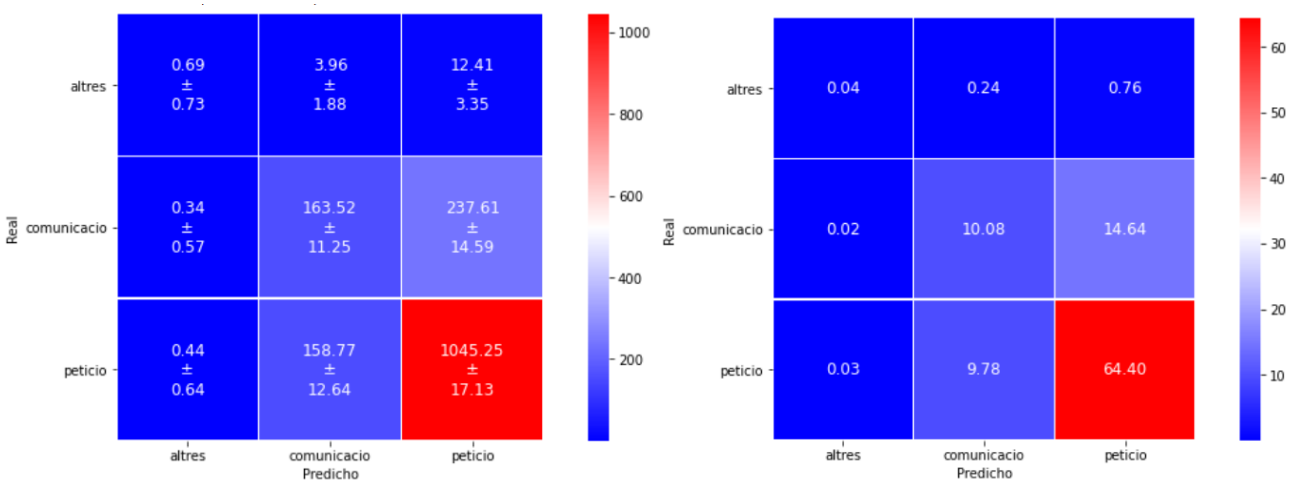


Figura 34: Matrices de confusión absoluta (izquierda) y normalizada (derecha) del modelo whenNonInc.

A diferencia del modelo anterior, en este caso los resultados solo se conservan para la categoría *PETICIÓ*. Las causas de este comportamiento son dos: en primer lugar, el conjunto de *No-incidencias* representa solo en torno al *10%* del total; en segundo lugar y dentro de la clase *No-incidencia*, la categoría *PETICIÓ* destaca principalmente, mientras que la categoría *ALTRES* es casi residual.

Como consecuencia de lo anterior, las métricas de las categorías *COMUNICACIÓ* y *ALTRES* empeoran significativamente, y especialmente el recall de la clase *ALTRES* que indica que tan solo el *4%* de los tickets serán bien clasificados.

Además, como apunte, se puede observar que la métrica *precision* de la categoría *ALTRES* sufre una mayor variabilidad en sus resultados, es decir, la elección del conjunto de datos influirá en este resultado.

Por tanto, aquí se puede afirmar que el modelo favorece a la clase *PETICIÓ* al compararse con el resto puesto que correos que pertenecen a las clases *COMUNICIÓ* y *ALTRES* tienen una probabilidad mucho mayor de ser incorrectamente clasificados respecto a correos procedentes de la clase *PETICIÓ,* los cuales tendrán, en general, una buena clasificación.

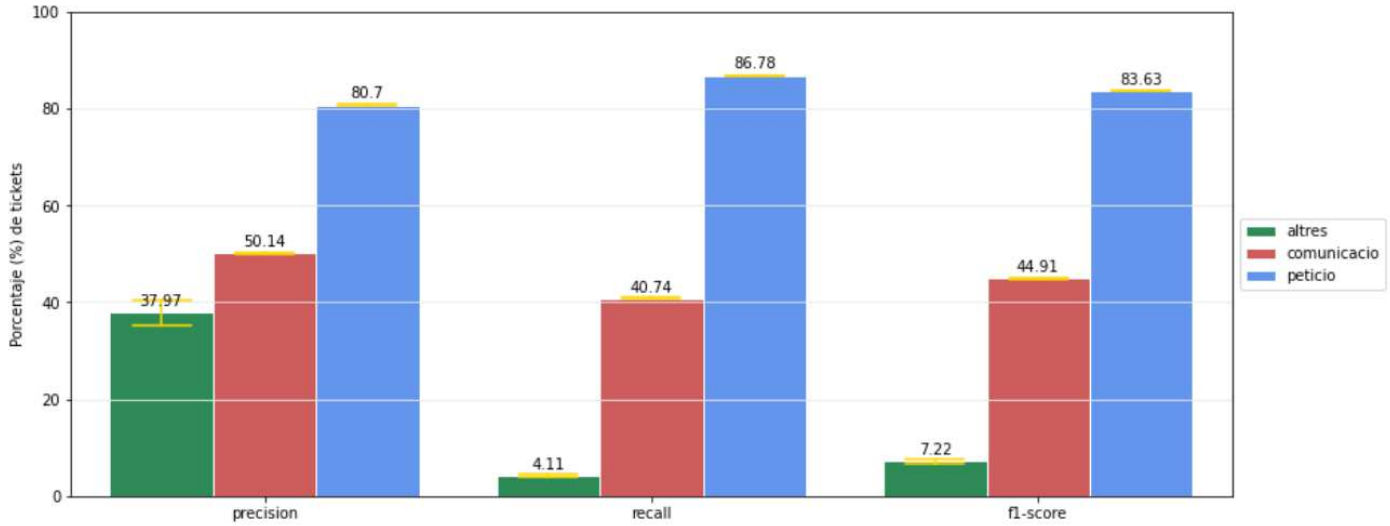


Figura 35: Cálculo de las métricas para el modelo whenNonInc.

6.2 Análisis del modelo por compañías

Para observar el comportamiento para las distintas compañías y comprobar si existe o no un trato desigual hacia alguna de ellas, se ha utilizado el modelo isInc.pk y el conjunto de datos obtenido del fichero *preprocessed_ground_truth_filtered_train_text.csv* al que se ha aplicado el filtro *Filtered_yes_no* igual a *yes*.

El modelo ***IsInc_v2.pkl*** clasifica los *tickets* como **incidencia** o **no incidencia**. Debido a que no se consiguió obtener el conjunto de datos utilizado en el entrenamiento y el testeo del modelo, y para **comprobar** el **funcionamiento** del mismo, *se* ha realizado un **reentrenamiento** de dicho modelo utilizando el método ***k-fold***.: Se ha **reentrenado** el modelo ***10 veces***, dividiendo el conjunto de datos en subconjuntos de *train (80%)* y *test* (*20%)* de manera aleatoria utilizando *kfold* y fijando una semilla para poder replicar los resultados obtenidos.

En este caso no se pudo utilizar la metodología más general *bootstrap* ya que al tener unas compañías menos tickets el análisis quedaba comprometido.

A continuación, se ha calculado el **porcentaje** de **incidencias** y **no incidencias** y el número total de tickets **para todas** las **compañías**. Una vez obtenido este cálculo, el subconjunto de test se ha filtrado por las compañías que aparecen en todos los entrenamientos y cuya distribución de tickets cumpliera **tres criterios**:

- Distribución con un **porcentaje** de **incidencias** mucho **mayor** que de **no incidencias**.
- Distribución con **porcentajes** de **incidencias** y **no incidencias balanceados**.
- Distribución con **porcentaje** de **no incidencias** mucho **mayor** que de **incidencias**.

Se ha seleccionado una compañía por criterio, para los dos primeros criterios, y **no** ha sido **posible encontrar** una **compañía** que **cumpliera** el **último criterio**. Esto se debe a que el **número total** de **tickets** asociado a las compañías que lo cumplen tiene un **valor** de ***1 a 5 tickets***, y realizar una prueba con un **número** tan **reducido** de **entradas** al modelo ofrece **resultados** poco **fiables**.

Las compañías elegidas han sido ***Departament de la Vicepresidència i d'Economia i Hisenda*** para el **primer criterio** y ***Agència de l'Habitatge de Catalunya (AHC)*** para el **segundo**.

A continuación, la Figura 36 muestra la distribución del porcentaje de incidencias/no incidencias para cada compañía.

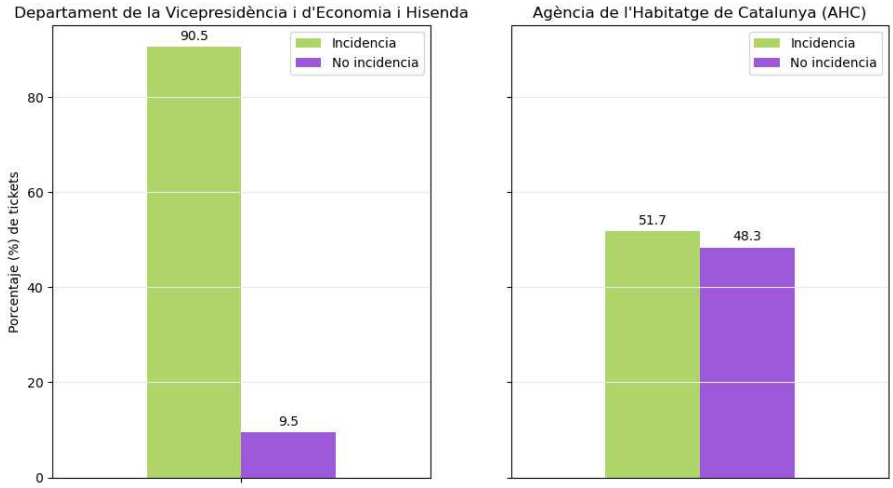


Figura 36: Distribución, en tanto por ciento de las incidencias/no incidencias de las compañías seleccionadas.

Se observa que, la compañía *Departament de la Vicepresidència i d'Economia i Hisenda* tiene un *90,5%* de incidencias y un *5%* de no incidencias cumpliendo así el primer criterio explicado. Para la compañía *Agència de l'Habitatge de Catalunya (AHC)* estos porcentajes están balanceados tal y como exige el segundo criterio, con un *51,7%* de incidencias y un *48,3%* de no incidencias.

Una vez seleccionadas las compañías, se obtiene una matriz de confusión y distintas métricas en cada entrenamiento, y se ha obtenido la matriz de confusión promedio de todos los entrenamientos, con valores tanto numéricos como porcentuales. Además de obtener estas matrices de confusión, se ha calculado la media y la desviación estándar de las métricas ofrecidas en cada entrenamiento: *precision*, *recall* y *F1 Score*.

La primera matriz de confusión obtenida es la observada en la Figura 37. Muestra la **media** de las **matrices** de **confusión obtenidas** a lo largo de los diez **entrenamientos** realizados. En ella se **incluye** la **desviación típica** para que se observe el posible **error introducido** en el cálculo de la **media**.

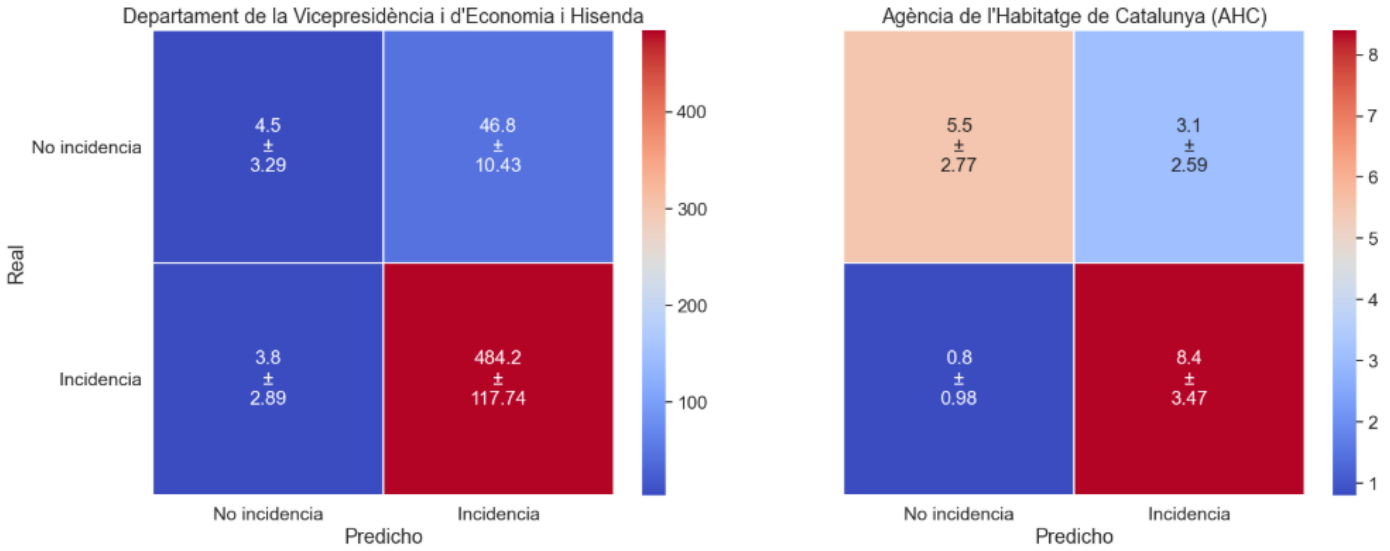


Figura 37: Matriz de confusión promedio de los 10 entrenamientos realizados.

En esta matriz, para la compañía *Departament de la Vicepresidenència i d'Economia i Hisenda*, se observa que los *tickets* que son **incidencias** se predicen con **bastante precisión**, sin embargo, las **no incidencias** a menudo son **confundidas** con **incidencias** obteniéndose una **predicción errónea** para esta clase en la mayoría de los casos.

En el caso de *Agència de l'Habitatge de Catalunya (AHC)*, al tener un **porcentaje** de **clases balanceado**, la **predicción** es **mejor** en general para las dos clases, obteniendo una predicción correcta de ambas clases en la mayoría de los casos.

A continuación, se muestran las **matrices** de **confusión** en **tanto por ciento**, esto servirá para poder calcular las métricas obtenidas en cada entrenamiento. En ella se observa lo que se ha explicado en cuanto a la predicción por clases. Aunque en el caso del *Departament de la Vicepresidenència i d'Economia i Hisenda* el porcentaje de No incidencias clasificadas como incidencias sea menor que en el caso de *AHC*, un *8.68%* frente un *17,42%*, también lo es el de no incidencias clasificadas como no incidencia, un 0,83% frente un *30,9%*. Esto implica lo que ya se ha explicado, **si existe un equilibrio entre las clases, el modelo funcionará mucho mejor que si las clases están desequilibradas**.

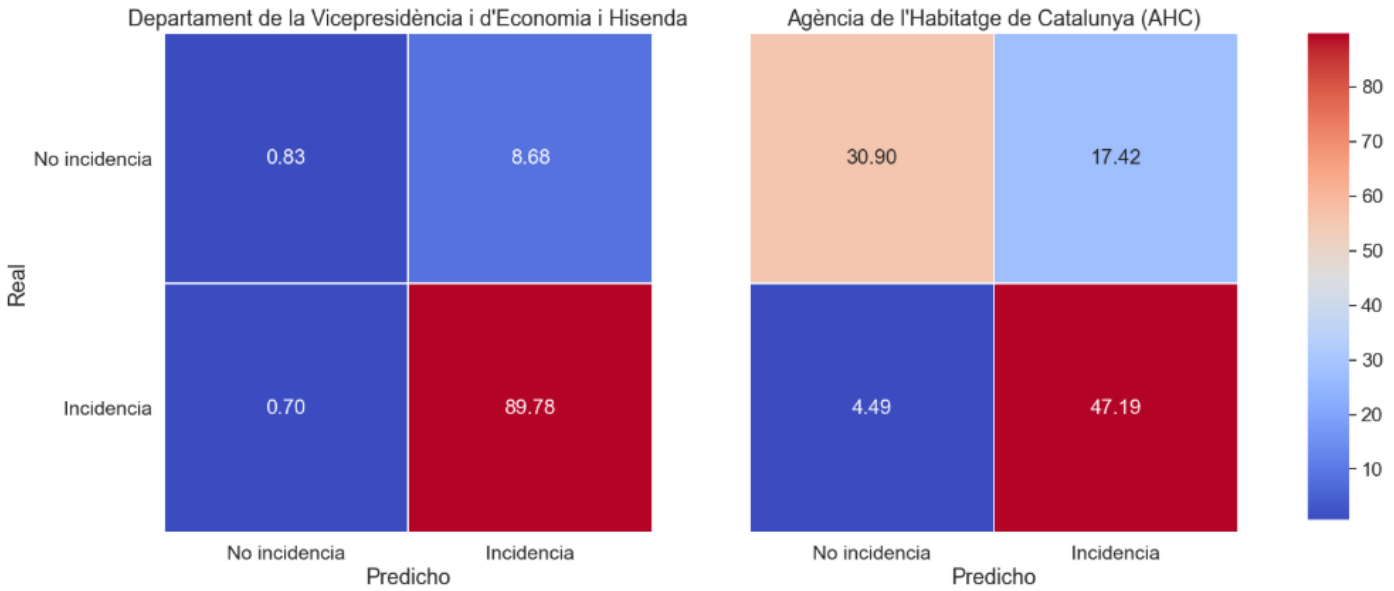


Figura 38: Matrices de confusión promedio de los 10 entrenamientos realizados, en tanto por ciento.

En cuanto a las métricas mencionadas, la Figura 39 muestra un histograma con la media de las métricas y su desviación típica por clases.

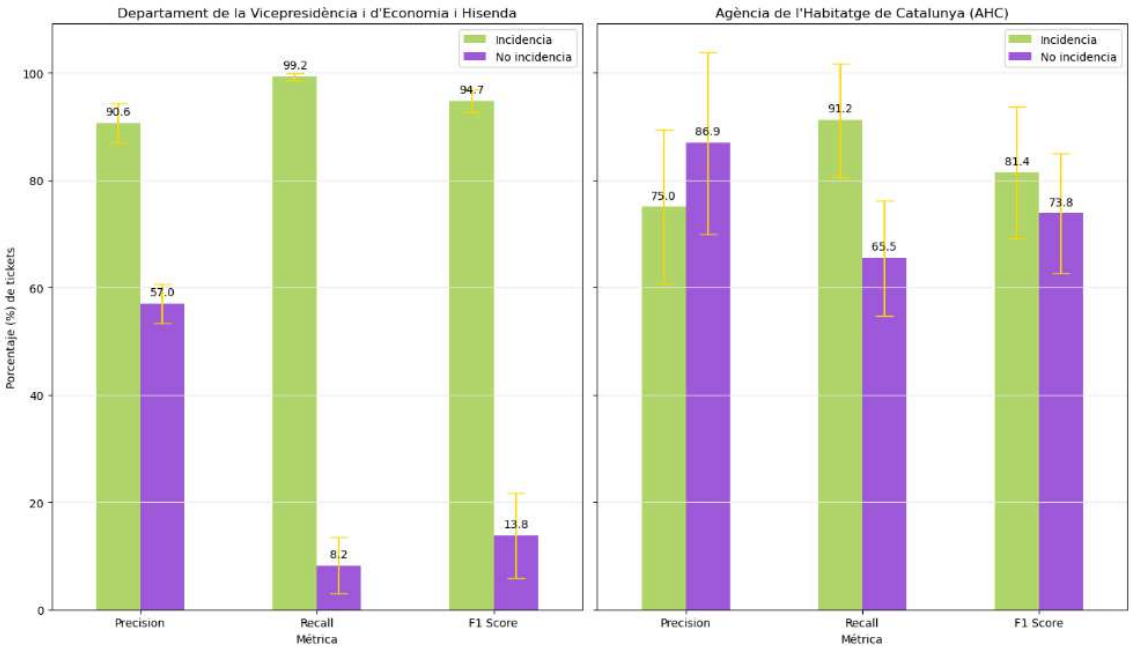


Figura 39: Media de las métricas precisión, recall y F1 Score a lo largo de los diez entrenamientos.

En la compañía ***Departament de la Vicepresidència i d'Economia i Hisenda*** la métrica ***precision*** para las **incidencias** es mucho **mayor** que para las **no incidencias**, sin embargo, para ***AHC*** ocurre lo **contrario**, la **precisión** para las **no incidencias** es **mayor** que para las **incidencias**. Aparentemente, esto puede contradecir la explicación

dada anteriormente sobre que el modelo funciona peor para la clasificación de no incidencias, pero no es así ya que, la ***precision*** también **tiene** en **cuenta** los **tickets** que han sido **clasificados** de **manera incorrecta (falsos positivos)**. Esto hace que la **precisión** de la **clase Incidencia disminuya**, porque el número de **no incidencias clasificadas** como **incidencias** es **mayor** que el número de **Incidencias clasificadas** como **no incidencias**. En este ejemplo, hay *46*,8 no incidencias clasificadas como incidencias lo que repercutirá de manera negativa en el cálculo de la *precision* para la clase *Incidencia*. Sin embargo, para el caso de la clase *No incidencia*, el número de incidencias clasificadas como no incidencias es de *3,8* por lo que el cálculo de la precisión para esta clase no se verá tan afectado.

En cuanto al *Recall, p*ara el caso de ***Departament de la Vicepresidència i d'Economia i Hisenda*** el ***Recall*** es mucho **mayor** para **incidencias** que para **no incidencias**, esto se debe a los **Falsos negativos**, en el caso de la **clase *Incidencia*** el **número** de **Falsos negativos** es muy **pequeño** comparado con los existentes para el caso no incidencia. En el caso de *AHC*, el *Recall* está más equilibrado, no existe una diferencia tan grande como la existente en el caso anterior. Esto se debe a que el funcionamiento del modelo es mejor, existen menos Falsos negativos para ambas clases.

Por último, para la **métrica** *F1 Score* ocurre lo mismo que para el *Recall,* en el **primer caso** existirá una **diferencia significativa entre** el *F1 Score* de las **clases** *Incidencia* y *No incidencia* mientras que, para el **segundo caso**, para *AHC*, el **comportamiento** de esta **métrica** será **similar** ya que el **modelo funcionará mejor** por el **equilibrio entre** las **clases** ya explicado.

Como **conclusión, no** puede **afirmarse** que **exista** un **trato desigual entre compañías** ya que, **existen** numerosas **compañías** que reportan **únicamente no incidencias** y cuyo **número** de *tickets* es **muy pequeño**. Para estos casos los **resultados** del **modelo no serán buenos,** ya que el modelo ha sido **entrenado** con **datos desbalanceados** que reúnen un número de incidencias muy superior al de no incidencias. Se ha demostrado que **cuanto mayor** sea el **equilibrio entre clases**, **mejor funcionará** el modelo **para ambas,** mientras que **cuanto mayor** sea el **desequilibrio entre** las dos **clases**, la clase ***No incidencia*** se verá **afectada** de manera **negativa** obteniendo predicciones pobres.

6.3 Análisis del modelo por idioma

Una vez comprobado el rendimiento del modelo en cada una de sus etapas, se decide incorporar información acerca del idioma con el que se redacta el correo, con el objetivo de comprobar si el idioma tiene algún tipo de impacto en los resultados obtenidos.

Para ello, se aplica la función *detect* de la librería *langdetect* para identificar el idioma, y se distingue entre *catalán*, *español* y o*tros*.

En este apartado se mostrarán las gráficas obtenidas evaluando el rendimiento de los modelos recopilados en sus diferentes etapas en función del idioma, y al igual que en los análisis anteriores, los resultados se obtienen a partir de un *bootstrap* sin reemplazamiento de *1.000* repeticiones, empleando en cada una de ellas el método *hold-out* con un *80%* de datos destinados al entrenamiento y el *20%* restante a la evaluación de cada uno de los modelos.

Con esto, se intentará detectar si el idioma en el que está redactado el correo tiene influencia en el modelo, es decir, si el idioma del correo influye de alguna forma en cómo de bien clasifica el modelo a dicho correo.

6.3.1 Modelo isInc

Esta primera parte del modelo se encarga de identificar, en una primera etapa, si un correo es *Incidencia* o *No-incidencia*.

Tal y como se puede observar en la **¡Error! No se encuentra el origen de la referencia.**, los resultados para la clase *Incidencia* se mantienen estables sin importar el idioma seleccionado, pero dentro de la clase *No-incidencia* sí que se observa una importante reducción del *recall* (a la mitad en el caso del idioma *otros* y a una cuarta parte para el *español*),

respecto al *20%* observado en la Figura 31, donde no se tenía en cuenta el idioma. Por otra parte, se aprecia una mejora significativa en la métrica *precision* del idioma *otros*.

Por tanto, y al igual que se constató en el apartado **6.1.1 Modelo isInc**, para **correos de la clase minoritaria *No-incidencia* se genera una situación de tratamiento desigual** como consecuencia del *recall* tan bajo **del *español* respecto al *catalán***, *y* que se traduce en que las probabilidades de un correo de ser incorrectamente clasificado aumentan considerablemente cuando está redactado en *español*.

Este tratamiento desigual, se ve agravado por el hecho de que los datos están desbalanceados a favor del idioma *catalán* (que representa, al menos, el *80%* respecto del total), tal y como se puso de manifiesto en el apartado **5 Análisis del idioma existente en la columna train_text**, del documento **EQUIA_Categorizador de tickets_Informe de exploración de datos_v1.pdf**, y por tanto el carácter minoritario de los correos electrónicos escritos en *español* es muy elevado ya de por sí mismo, sin tener en cuenta además el hecho de estar tratando correos de la clase minoritaria No-incidencia.

En menor medida, sucede algo parecido con la clase *otros*, a pesar de su ligera mejora en la métrica *precision*.

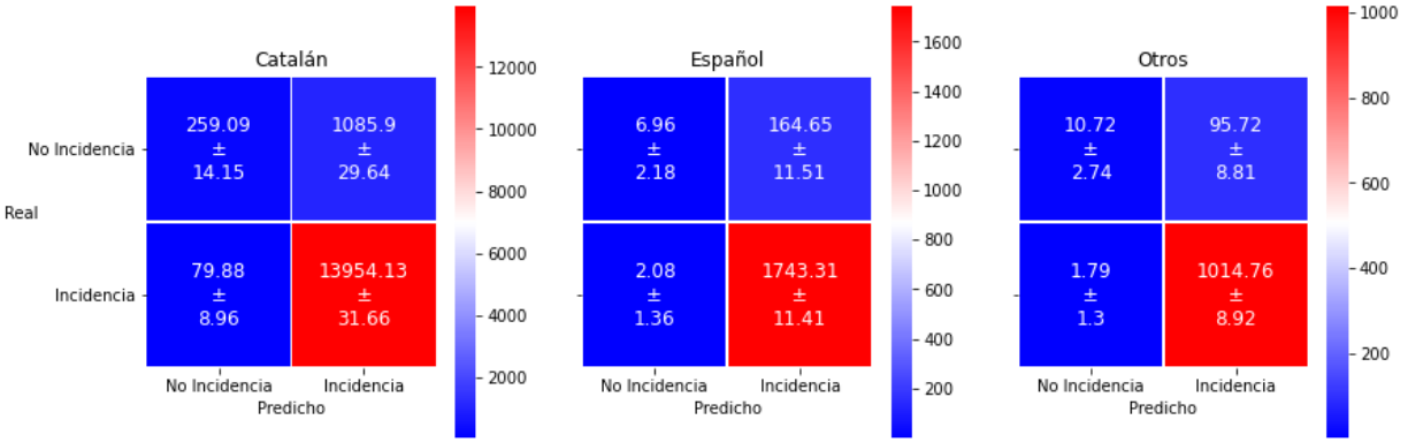


Figura 40: Matrices de confusión en valores absolutos del modelo isInc en función del idioma del correo.

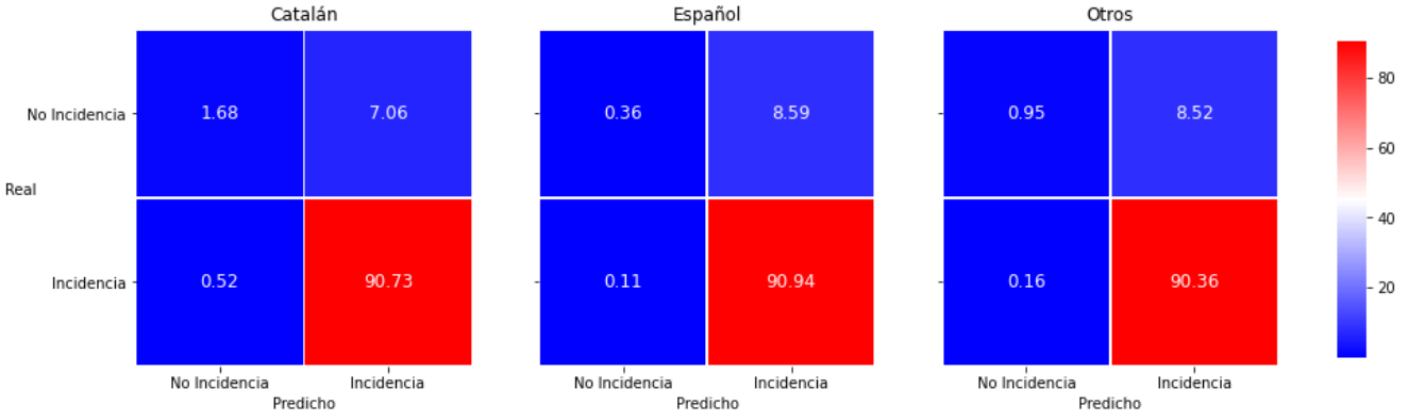


Figura 41: Matrices de confusión en valores porcentuales del modelo isInc en función del idioma del correo.

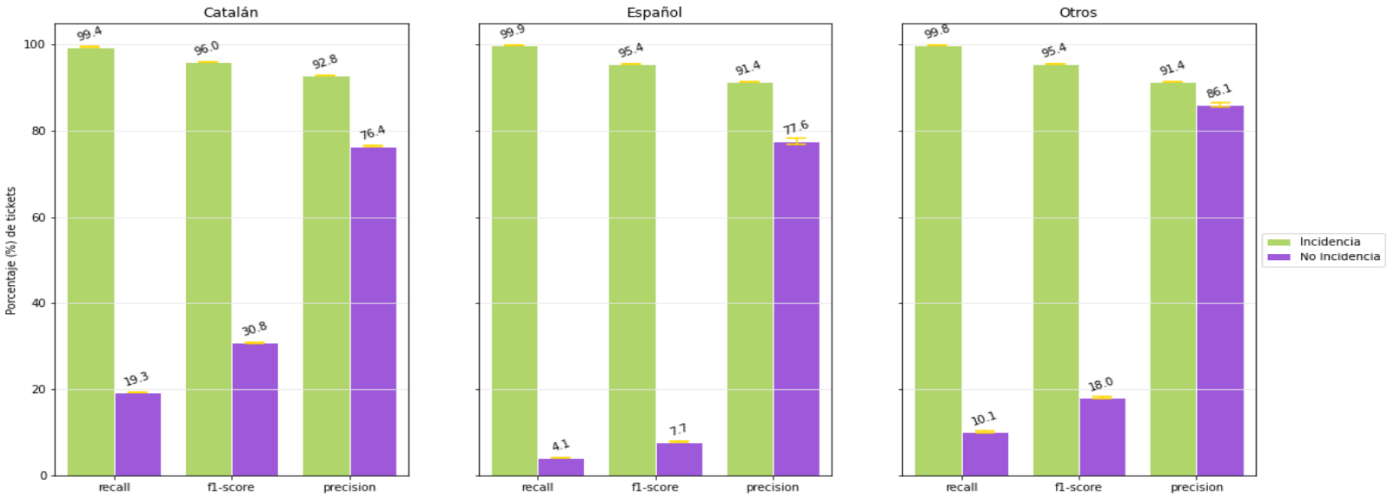


Figura 42: Cálculo de las métricas para el modelo isInc en función del idioma.

6.3.2 Modelo whenInc

Una vez que el correo se ha clasificado como *Incidencia*, el modelo trata de clasificarlo dentro de alguna de las siguientes categorías: *ALTRES*, *APLICACIONS* o *LLOC DE TREBALL*.

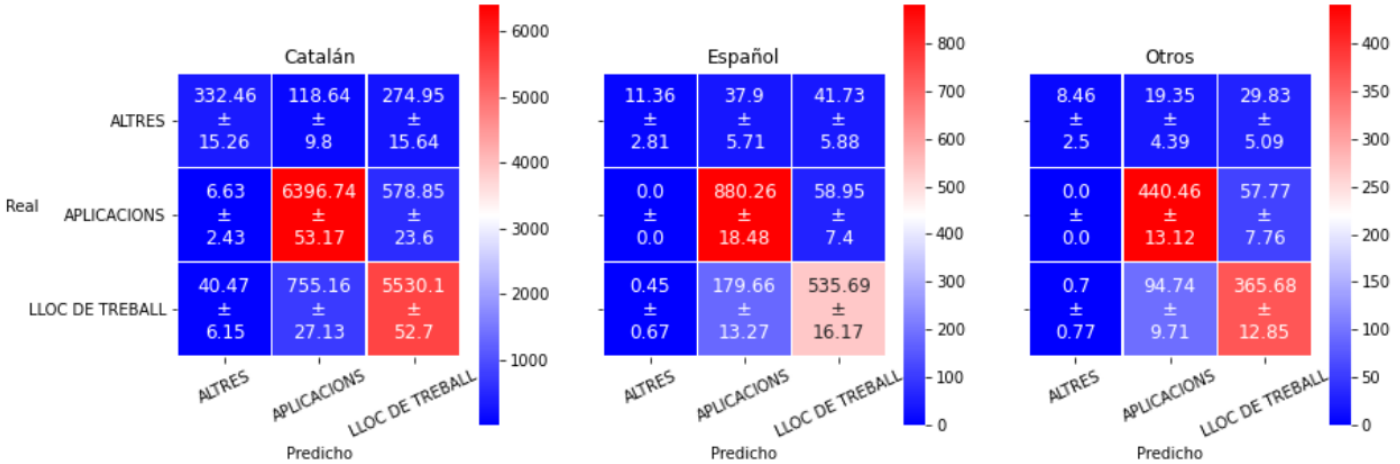


Figura 43: Matrices de confusión en valores absolutos del modelo whenInc en función del idioma del correo.

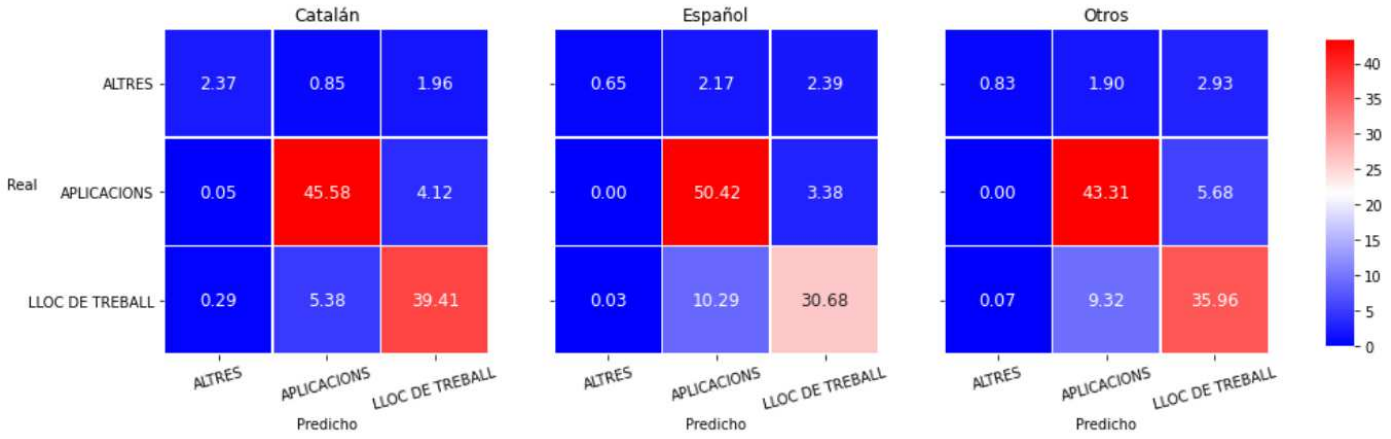


Figura 44: Matrices de confusión en valores porcentuales del modelo whenInc en función del idioma del correo.

Observando las tres clases por separado, los resultados de sus métricas se pueden interpretar de la forma siguiente.

Por una parte, *APLIACIONS* muestra un aumento significativo, del *8%*, en la *precision* para el idioma c*atalán*.

A su vez, para la clase *LLOC DE TREBALL* se reducen un *8%* los valores del *F1-socre* en las clases *español* y *otros* (y que representa el promedio del descenso de entre un *5%* y un *12%* de los valores de *precision* y *recall* correspondientes).

Finalmente, aunque para la clase *ALTRES la precision* del *catalán* sea significativamente inferior (entre un *5%* y un *10%*) respecto al resto, su *F1-score* asociado es sustancialmente superior (al menos un *30%*) debido a que el *recall* se reduce drásticamente para las clases *español* y *otros*.

En resumen, las clases *APLICACIONS* y *LLOC DE TREBALL* se tratan ligeramente mejor si el correo se redacta en *catalán* aunque no parece que sea suficiente como para afirmar que esto supone una situación desigual. Sin embargo, sí que **se puede concluir que cuando el correo pertenece a la clase minoritaria *ALTRES*** el idioma del mismo influye claramente ya que la probabilidad de ser clasificado correctamente es *3* veces superior para el *catalán* respecto al resto, por lo que aquí **el modelo presenta un sesgo en contra de la clase *ALTRES***.

Al igual que en el apartado anterior, este tratamiento desigual, se ve agravado por el hecho de que los datos están desbalanceados a favor del idioma *catalán* (que representa, al menos, el *80%* respecto del total), tal y como se puso de manifiesto en el apartado **5 Análisis del idioma existente en la columna train_text**, del documento **EQUIA_Categorizador de tickets_Informe de exploración de datos_v1.pdf**, y por tanto el carácter minoritario de los correos electrónicos no escritos en *catalán* es muy elevado ya de por sí mismo, sin tener en cuenta además el hecho de estar tratando correos de la clase minoritaria *ALTRES*.

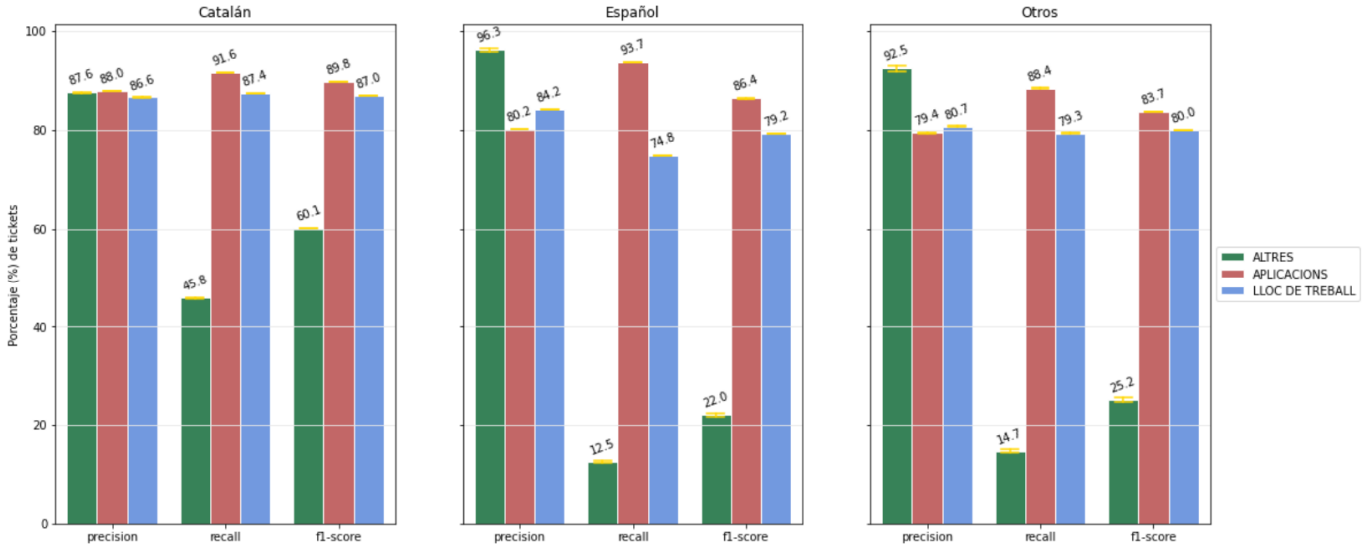


Figura 45: Cálculo de las métricas para el modelo whenInc en función del idioma.

6.3.3 Modelo whenNonInc

Una vez que el correo se ha clasificado como *No-incidencia*, el modelo trata de clasificarlo dentro de alguna de las siguientes categorías: *ALTRES*, *COMUNICACIÓ* o *PETICIÓ*.

Debido a que la clase *ALTRES* tiene escasas observaciones tanto en los idiomas *Castellano* como sobre todo en *otros* (que debido a la aleatoriedad provoca que no aparezcan observaciones de esta clase en el entrenamiento y el modelo sea incapaz de generar predicciones dentro de esta categoría, dando lugar en ocasiones a matrices de confusión de tamaño 2x2 en lugar de 3x3 y que, por tanto, no pueden promediarse), se ha obligado a que en cada una de las particiones de entrenamiento y test y para cada idioma, al menos exista una observación de la categoría ALTRES. Esta solución se ha considerado mejor que completar con ceros la matriz de tamaño 2x2 para conseguir el tamaño 3x3 deseado porque, de algún modo, se estarían *falseando* los resultados ya que, por ejemplo, se asignaría una precisión de *0* en la categoría *ALTRES* cuando en realidad no se podía clasificar ninguna observación dentro de esta categoría porque el modelo no la conocía.

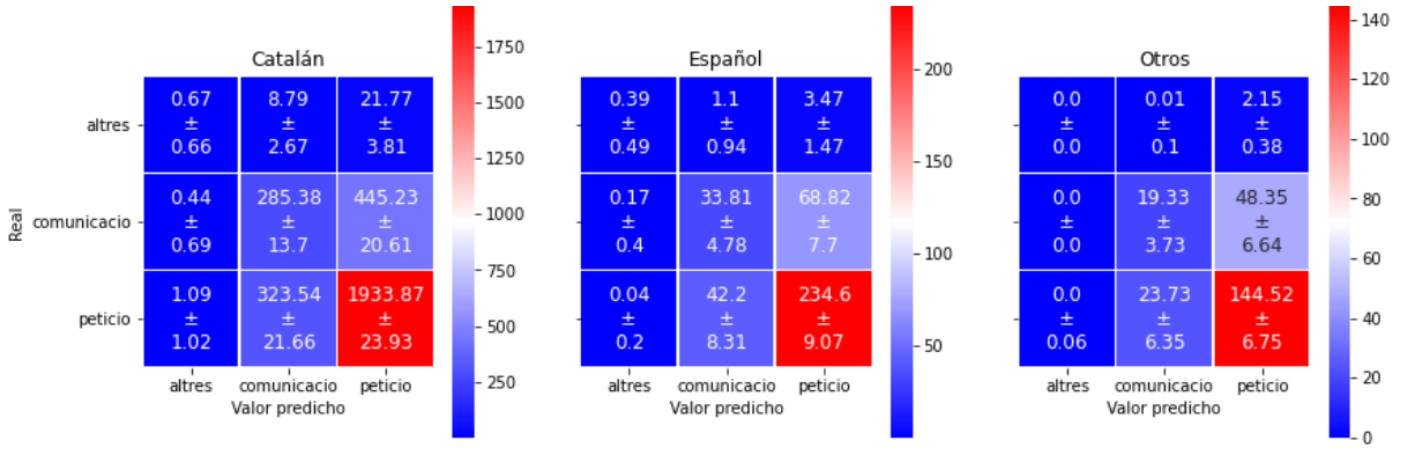


Figura 46: Matrices de confusión en valores absolutos del modelo whenNonInc en función del idioma del correo.

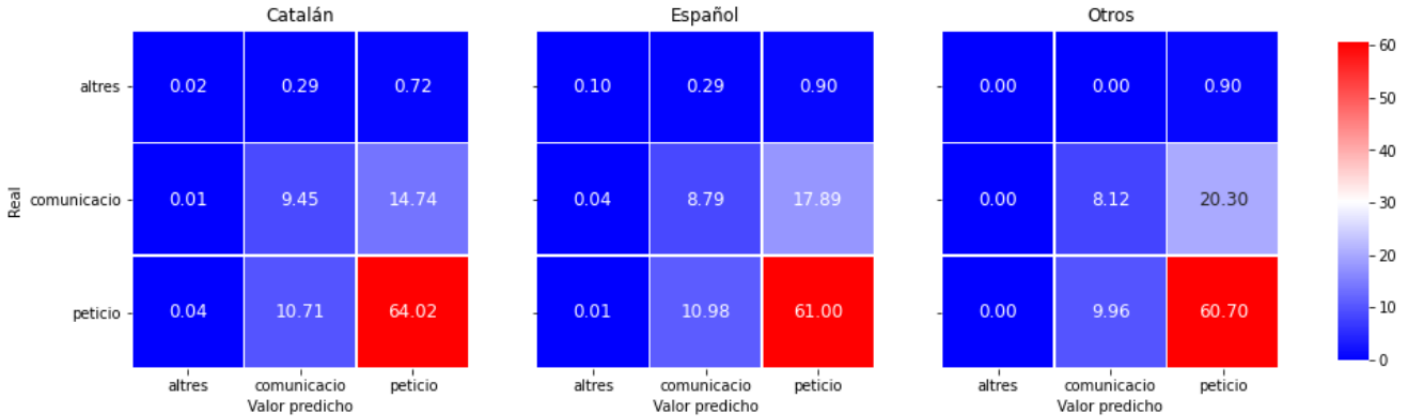


Figura 47: Matrices de confusión en valores porcentuales del modelo whenNonInc en función del idioma del correo.

Observando las tres clases por separado, los resultados de sus métricas se pueden interpretar del siguiente modo.

Por una parte, la clase *PETICIÓ* tiene resultados parecidos para todos los idiomas y solo se aprecia una leve mejora de la *precision* del modelo para el *catalán*.

A su vez, para la clase *COMUNICACIÓ* se observa que, por este orden, *español* y *otros* empeoran significativamente su valor de *recall*, mientras que el valor de la *precision* se mantiene estable sin importar el idioma (y por esta razón, el valor del *F1-score* sigue siendo mejor para el *catalán*).

Finalmente, **la clase *ALTRES* muestra el comportamiento más impredecible**, seguramente causado por el bajo número de muestras disponibles (*70* del *catalán*, *11* del *español* y *4* de *Otro*). Por ello, a pesar de que predomina la categoría *catalán,* el modelo muestra su mejor rendimiento en la clase *español* para todas las métricas.

Llama la atención que en la categoría *ALTRES* no haya valores disponibles de las métricas, pero esto se explica atendiendo a la matriz de confusión: si nos fijamos en la matriz normalizada de la categoría *otros*, la *precision* es *0* porque no se ha predicho ningún ticket dentro de la categoría *ALTRES* (aquí se puede ver cómo los *3* valores de la primera columna en la matriz de confusión son cero, por lo que no es posible acertar ningún ticket puesto que no se ha incluido ninguno en la predicción). Por su parte, el *recall* es *0* porque, aunque como se puede ver en la matriz absoluta

en el conjunto de test se han incluido en promedio unas *2.15* muestras, ninguna de ellas se ha identificado correctamente dentro de la clase y se han asignado erróneamente casi en su totalidad a la clase *PETICIÓ*.

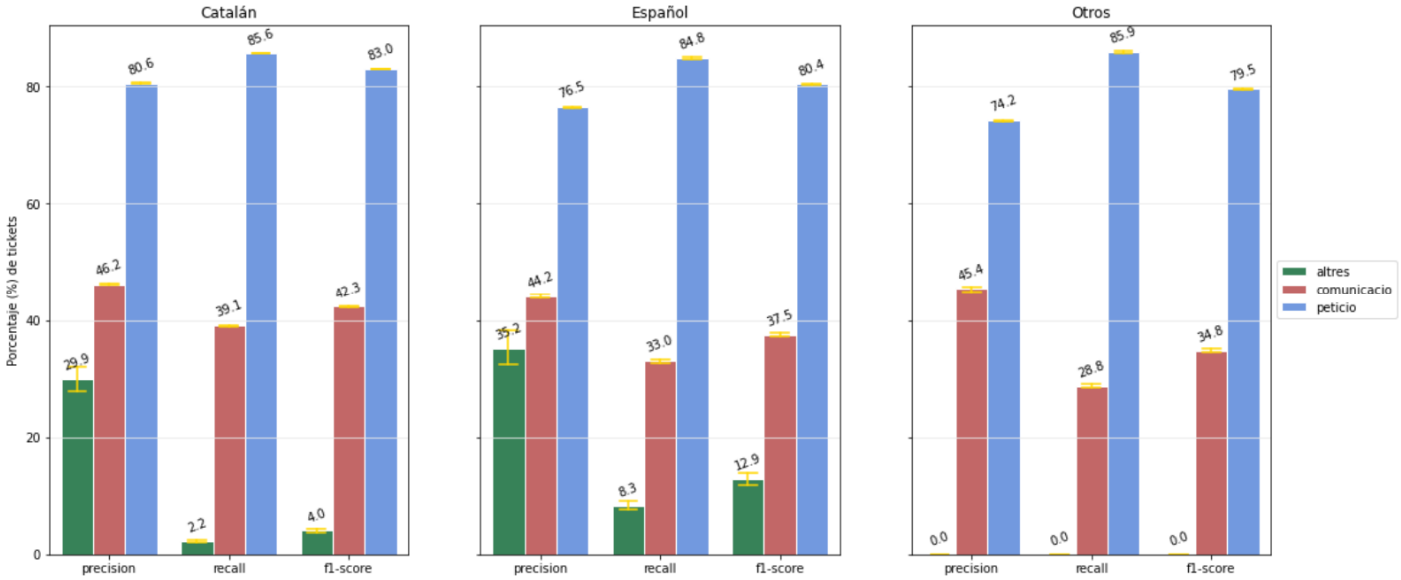


Figura 48: Cálculo de las métricas para el modelo whenNonInc en función del idioma.

Como conclusión de este apartado, **se puede afirmar que el modelo no reconoce la categoría *ALTRES* cuando uno de sus tickets ha sido redactado en otro idioma distinto al *español* o *catalán*** (y es una consecuencia directa de que solo haya *4* tickets de esta clase en este idioma, por lo que el modelo no llega a identificar características diferentes en ellos como para separar la clase puesto que la función de pérdida del modelo asigna la misma importancia a todos los tickets sin importar la clase o idioma al que pertenezcan).

Además, dentro de la clase *COMUNICACIÓ* se puede afirmar que si el idioma del ticket es *catalán* facilitará su clasificación y que al menos el idioma *otros* sí que se puede ver suficientemente perjudicado en comparación con el *catalán*.

6.4 Análisis de la longitud del texto de los correos

Este apartado adicional se dedica a comparar la longitud en número de palabras de los correos redactados en idioma catalán frente al resto, estableciendo como límite las *69* palabras y su propósito es el siguiente.

Las métricas de precisión de los modelos calculados en los apartados anteriores pueden verse afectadas por la decisión tomada acerca de truncar los correos a una longitud de *69* palabras o directamente prescindir de aquellos correos cuya longitud sea superior. Según el criterio escogido, los resultados de los apartados anteriores pueden verse afectados.

Por esta razón, a continuación, se ofrece una comparativa del *catalán* frente al resto de idiomas, para comprobar el número y proporción de correos que se modificarían o eliminarían en caso de seguir el criterio expuesto anteriormente.

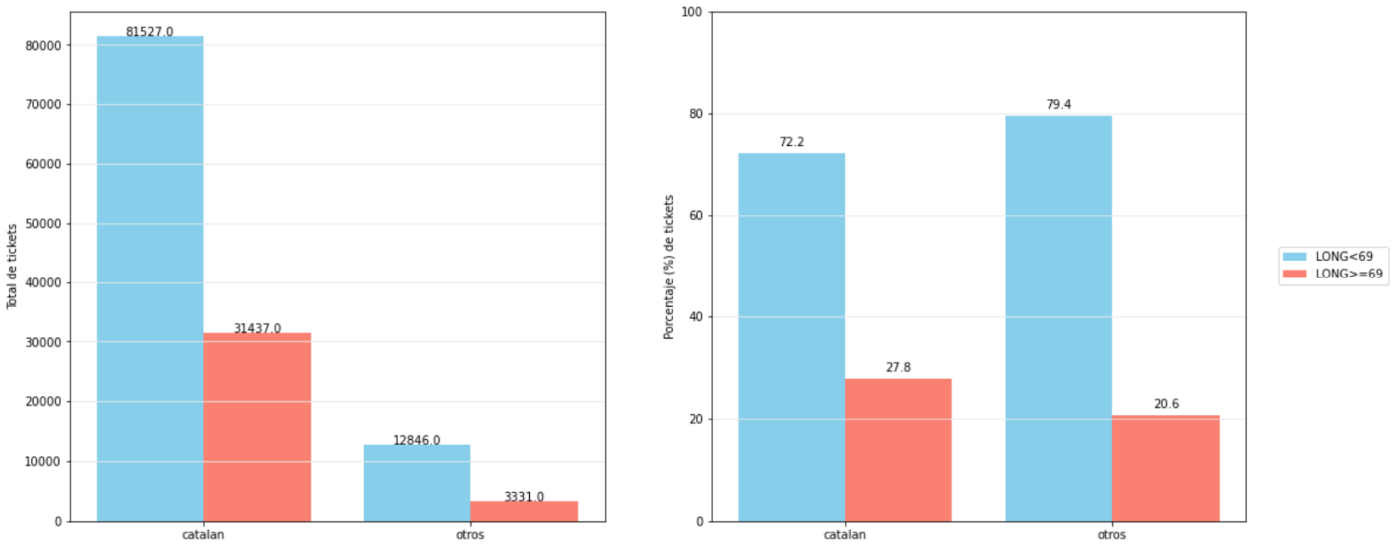


Figura 49: Distribución absoluta y en porcentaje de los tickets en función de su longitud inferior o superior a las *69* palabras.

Si se decidiese aceptar como válido el criterio que afirma prescindir de aquellos correos con longitudes que superan las 69 palabras, entonces a partir de la *Figura 49* se comprueba que en términos absolutos el número de correos en catalán que se eliminarían sería muy superior respecto al resto de idiomas (básicamente porque es el idioma más representativo). Si se comparan en términos de porcentajes, la proporción de correos en *catalán* que se eliminan sigue siendo significativamente superior.

Por lo tanto, en caso de aplicarse algún tipo de reducción de la información en base a la longitud del texto de los correos, ya sea eliminando los correos que superen una longitud determinada, truncando el texto o sustituyendo el texto del correo por el título del correo, se estaría generando un escenario en el que l**os correos redactados en catalán podrían recibir un tratamiento desigual** en la fase de limpieza de datos, ya que se eliminación de información afectaría en mayor medida a estos correos.


ALTRAN
Part of Capgemini