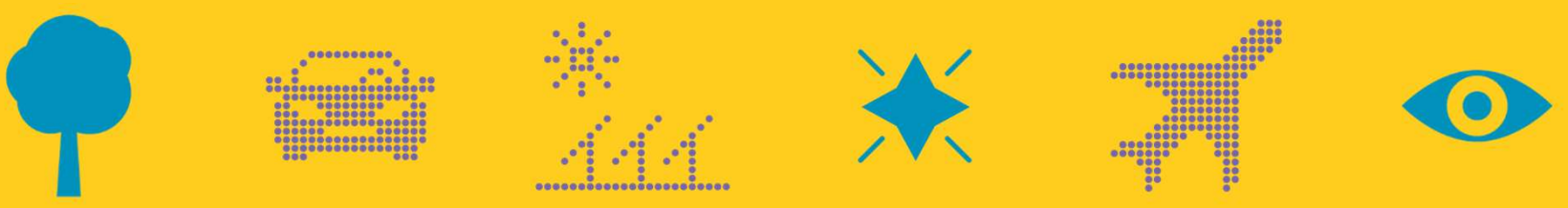
EQUIA

TRANSPARENCIA Y EQUIDAD EN LOS ALGORITMOS DE INTELIGENCIA ARTIFICIAL

Reunión de seguimiento

ALTRAN
Part of Capgemini



EQUIA

TRANSPARENCIA Y EQUIDAD EN LOS ALGORITMOS DE INTELIGENCIA ARTIFICIAL

ALTRAN
Part of Capgemini

Agenda

1 Objetivos

2 Categorizador de tickets

3 Recomendador de libros

4 Resultados y conclusiones globales

1.

Objetivos

ALTRAN
Part of Capgemini

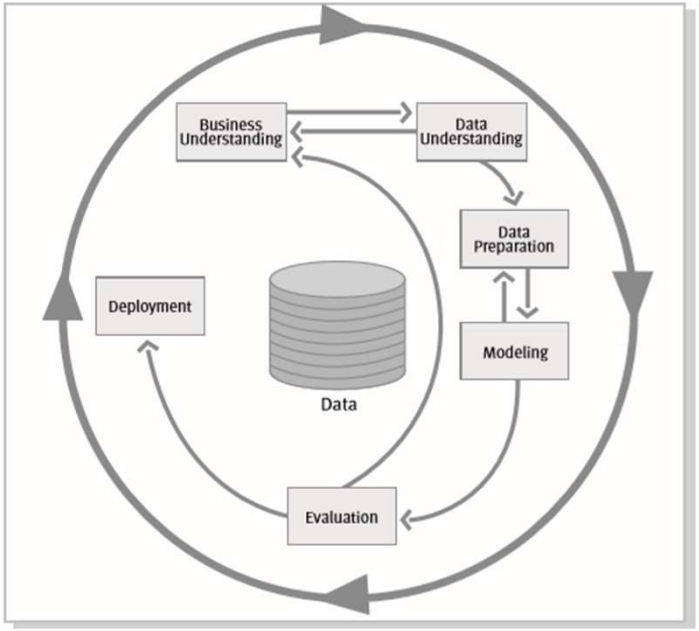
Requisitos de los sistemas de Inteligencia Artificial (IA) según la Unión Europea*:

- **IA transparente: Se debe proporcionar información contextual de cómo funciona el sistema de IA.**
- **IA explicable: Todos los sistemas de IA y sus decisiones deben poderse explicar con un lenguaje que las personas puedan entender.**
- **IA justa: Es necesario que los sistemas de IA hagan recomendaciones que no discriminen por motivos de raza, género, religión u otros factores similares para garantizar que sean representativos y logren resultados equitativos para todos.**
- **IA robusta: Dado que los sistemas de IA ya reciben un poder de decisión autónomo significativo en situaciones de alto riesgo, deben ser resistentes al riesgo, la imprevisibilidad y la volatilidad en entornos del mundo real.**
- **IA privada: los sistemas de IA deben cumplir con las leyes de privacidad que regulan la recopilación, el uso y el almacenamiento de datos y garantizar que la información personal se utilice de acuerdo con los estándares de privacidad.**

* European Commission. (2019, abril 8). Ethics guidelines for trustworthy AI [Text]. Shaping Europe's Digital Future - European Commission.

CRISP-DM (CRoss-Industry Standard Process for Data Mining)

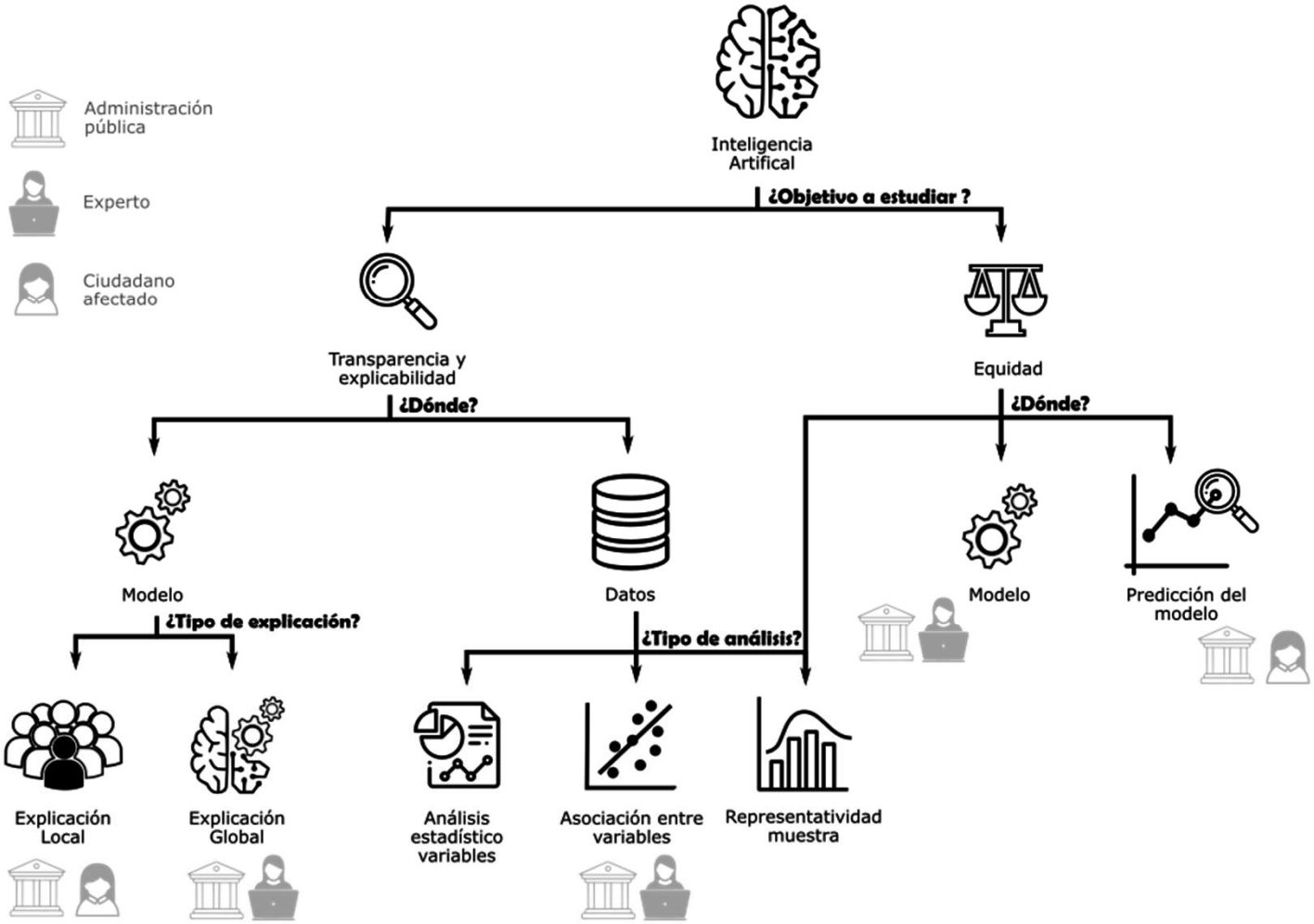
Metodología de gestión de proyectos diseñada en 1999 por un consorcio de empresas europeas: NCR(Dinamarca), AG (Alemania), SPSS (Inglaterra) y OHRA (Holanda).



Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria*	**Collect Initial Data** *Initial Data Collection Report*	**Select Data** *Rationale for Inclusion/ Exclusion*	**Select Modeling Techniques** *Modeling Technique* *Modeling Assumptions*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models*	**Plan Deployment** *Deployment Plan*
Assess Situation *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits*	**Describe Data** *Data Description Report*	**Clean Data** *Data Cleaning Report*	**Generate Test Design** *Test Design*	**Review Process** *Review of Process*	**Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan*
Determine Data Mining Goals *Data Mining Goals* *Data Mining Success Criteria*	**Explore Data** *Data Exploration Report*	**Construct Data** *Derived Attributes* *Generated Records*	**Build Model** *Parameter Settings* *Models* *Model Descriptions*	**Determine Next Steps** *List of Possible Actions* *Decision*	**Produce Final Report** *Final Report* *Final Presentation*
Produce Project Plan *Project Plan* *Initial Assessment of Tools and Techniques*	**Verify Data Quality** *Data Quality Report*	**Integrate Data** *Merged Data*	**Assess Model** *Model Assessment* *Revised Parameter Settings*		**Review Project** *Experience Documentation*
		Format Data *Reformatted Data*			
		Dataset *Dataset Description*			

ALTRAN Part of Capgemini

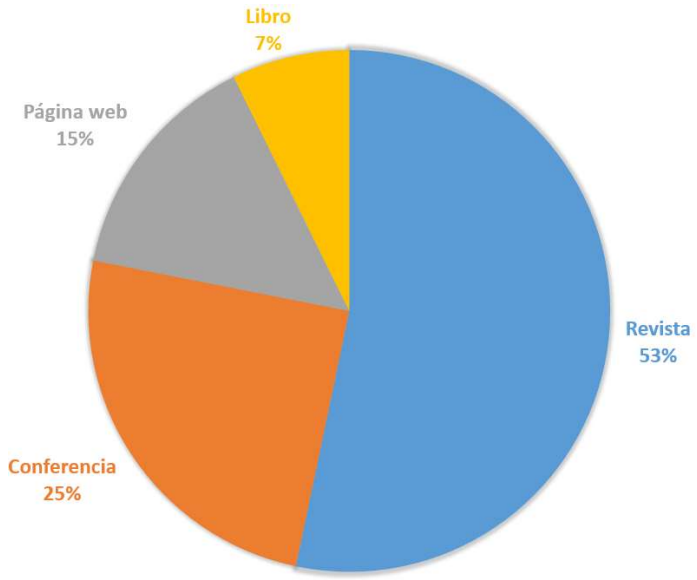
Transparencia y equidad en la Inteligencia Artificial

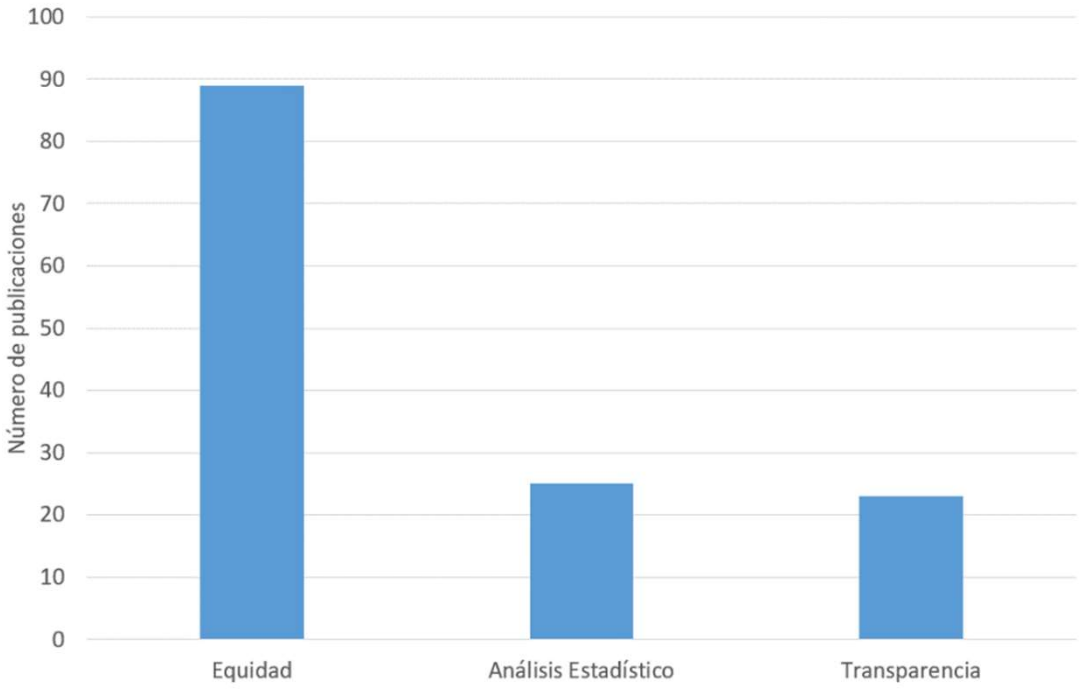


ALTRAN Part of Capgemini

Estado del arte

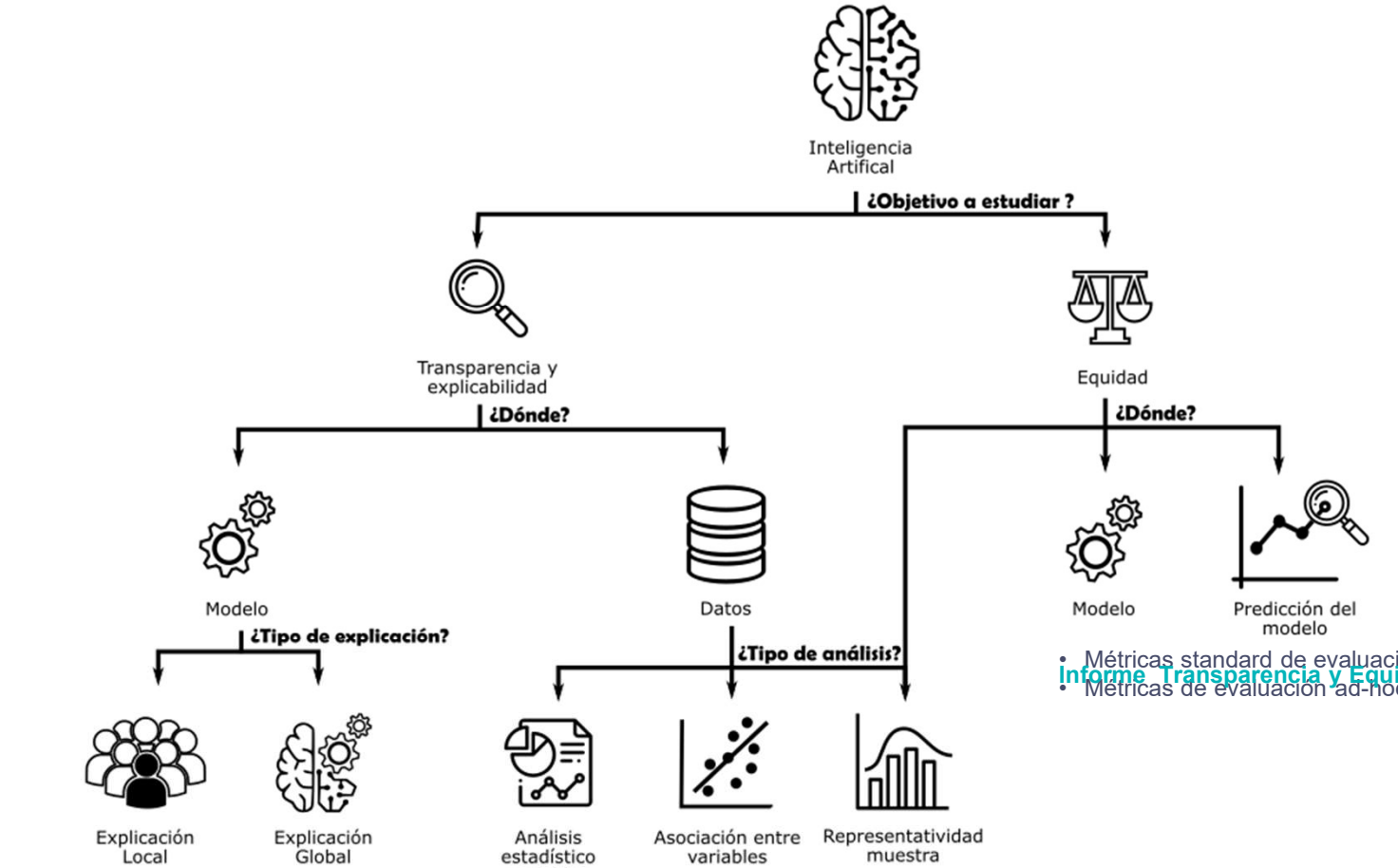
Se estudiaron un total de 134 publicaciones





ALTRAN
Part of Capgemini

Transparencia y equidad en la Inteligencia Artificial



- LIME
- Ad-hoc

Informe Transparencia y Equidad

- SHAP
- WHAT-IF
- Ad-hoc

- Análisis de diferencias
- Métricas absolutas

Informe descriptivo + Informe Exploración

- Métricas condicionales
- Métricas de situación

- Métricas standard de evaluación
- Métricas de evaluación ad-hoc

Informe Transparencia y Equidad

- Regresión
- Correlación
- Transformación de datos

ALTRAN Part of Capgemini

Casos de estudio

Se tomaron dos casos de estudio:

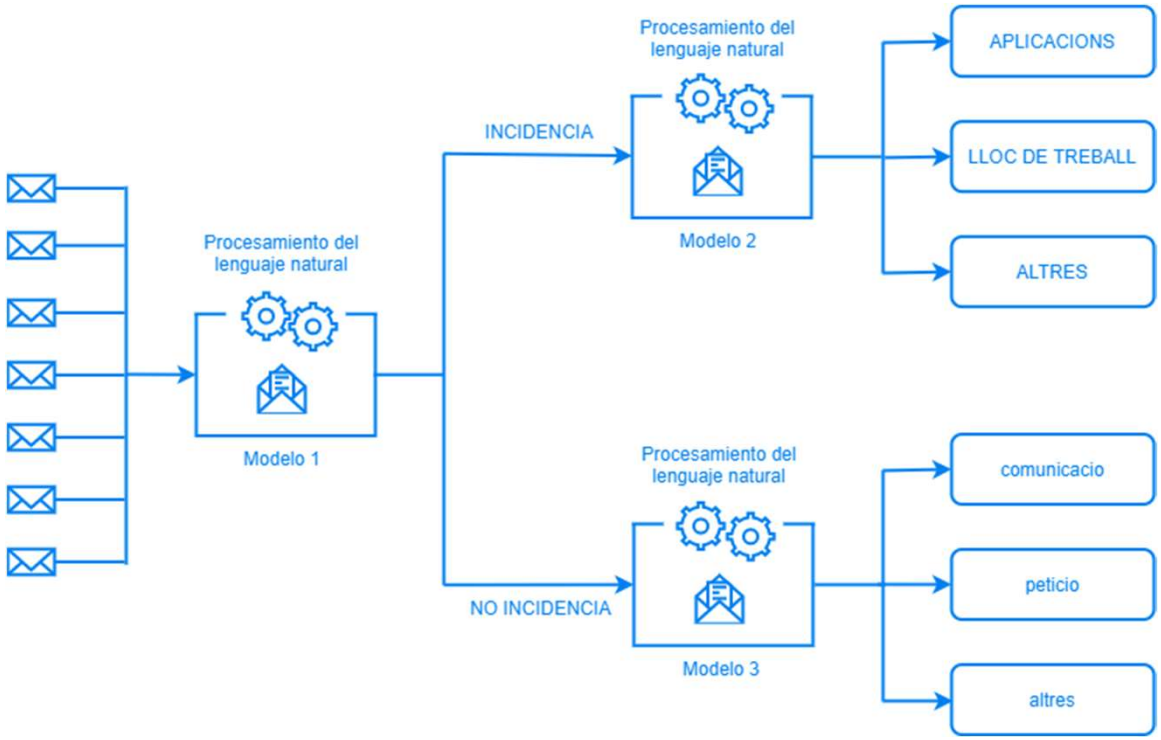
- **Categorizador de tickes.**
- **Recomendador de libros.**

2.

Categorizador de tickets

ALTRAN
Part of Capgemini

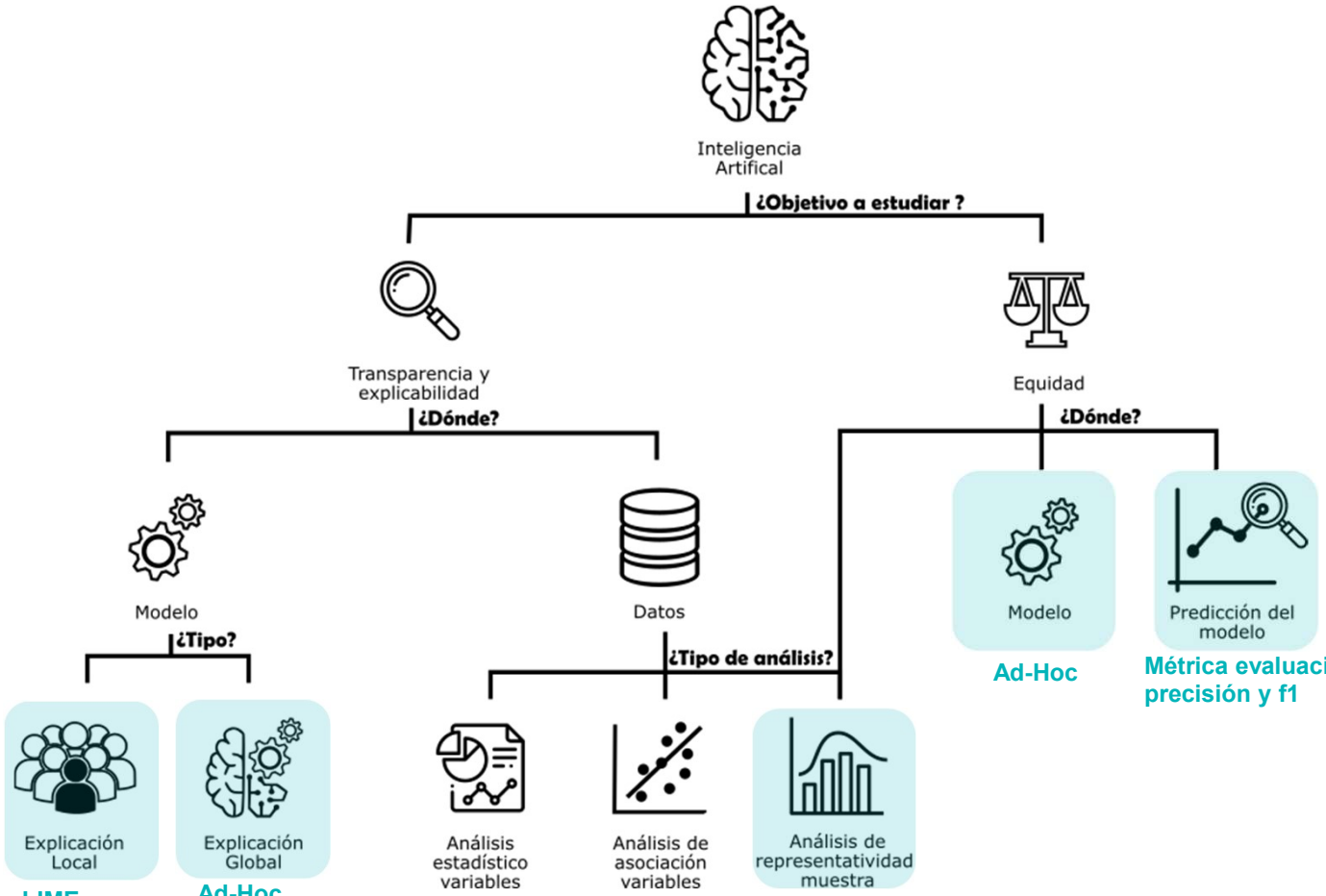
Categorizador de tickets



- **Contiene un algoritmo de machine learning autoexplicable y de muy baja complejidad.**
- **El conjunto de datos no contiene apenas variables sensibles.**

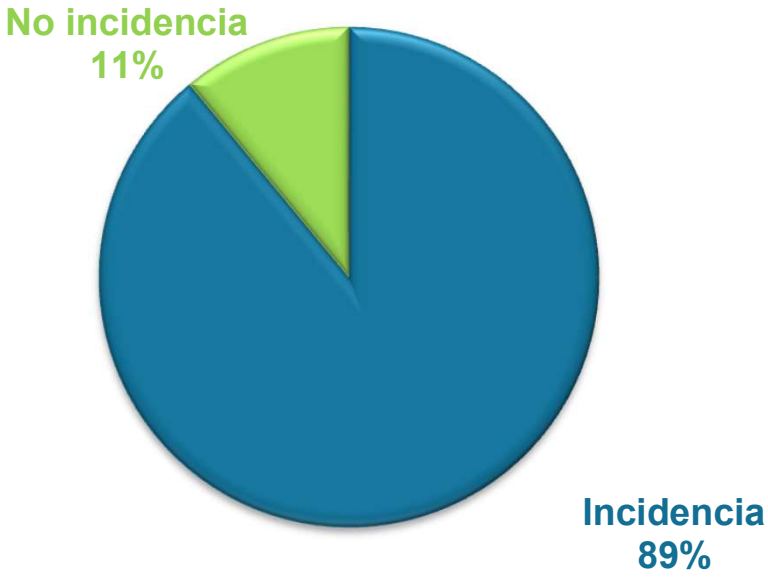
ALTRAN Part of Capgemini

Categorizador de tickets



Representatividad y Equidad: Modelo 1

Porcentaje de tickets INCIDENCIA y NO INCIDENCIA



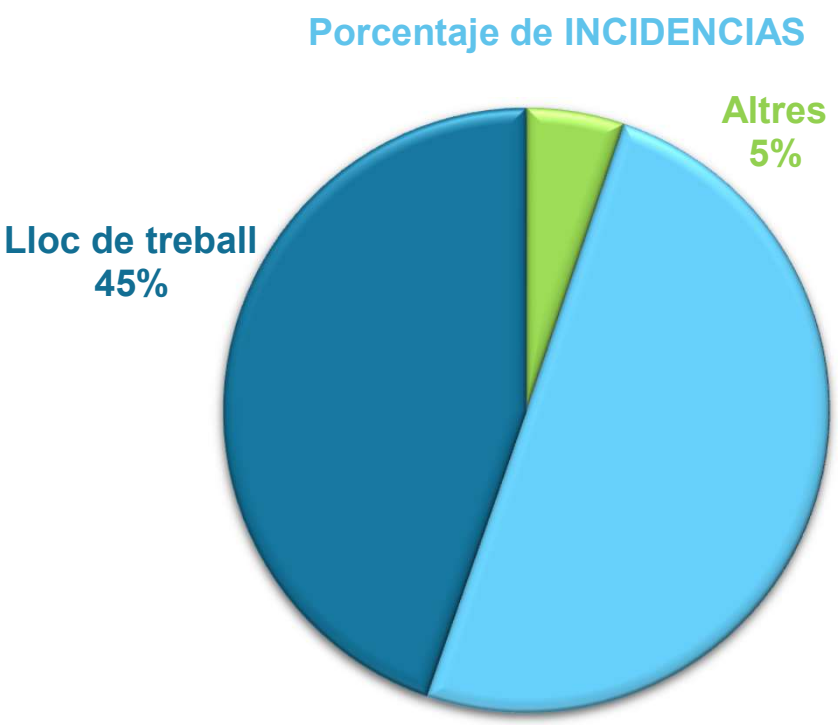
Recall

Tipo ticket	Clasificación correcta
Incidencia	99,38 %
No incidencia	20,24 %

El gran desbalance entre las INCIDENCIAS y NO INCIDENCIAS genera que las NO INCIDENCIAS sean clasificadas correctamente solo en un 20,24% de los casos.

ALTRAN Part of Capgemini

Representatividad y Equidad: Modelo 2



Recall

Tipo ticket	Clasificación correcta
Altres	45,86 %
Aplicacions	91,91 %
Lloc de treball	86,49 %

El gran desbalance entre los tipos de INCIDENCIAS genera que 1 de cada 2 incidencias Altres sean clasificadas incorrectamente.

ALTRAN Part of Capgemini

Representatividad y Equidad: Modelo 3

Recall

Tipo ticket	Clasificación correcta
Altres	4,11%
Comunicacio	40,74 %
Peticio	86,78 %

El gran desbalance entre los tipos de NO INCIDENCIAS genera que el comportamiento para cada clasificación sea desigual y especialmente malo para la categoría Altres.

ALTRAN Part of Capgemini

Representatividad y Equidad: Consecuencias de la infrarepresentación

Departamentos

El algoritmo clasificará peor las NO INCIDENCIAS para departamentos con gran desbalance, generando un tratamiento desigual y no equitativo.

Tipo ticket	Clasificación correcta	
	Departament de la Vicepresidència y d'Economia i Hisenda	Agència de l'Habitage de Catalunya (AHC)
Incidencia	99,2 %	91,2 %
No incidencia	8,2 %	65,5 %

Departament de la Vicepresidència y

Agència de l'Habitage de Catalunya

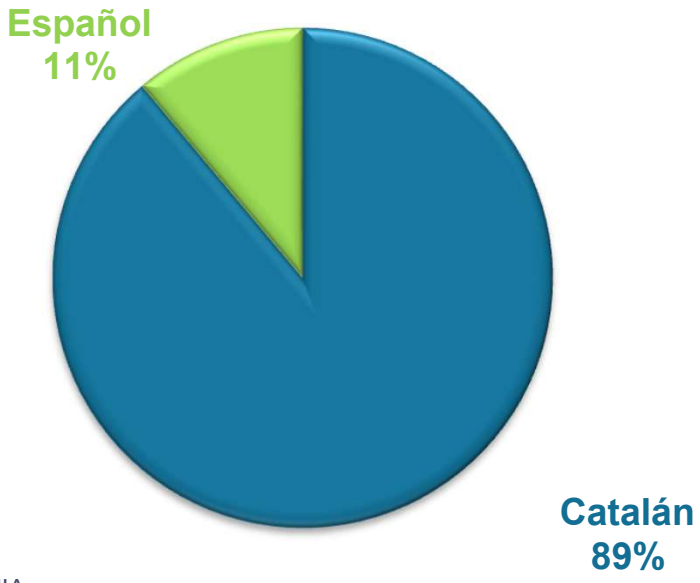
Incidencia 91%

ALTRAN Part of Capgemini

Representatividad y Equidad: Consecuencias de la infrarepresentación

Idioma

Las NO INCIDENCIAS escritas en español se clasificarán peor que las escritos en catalán. Sin embargo, al haber pocas NO INCIDENCIAS, los emails escritos en español solo se clasifican un 1% peor que los escritos en catalán.

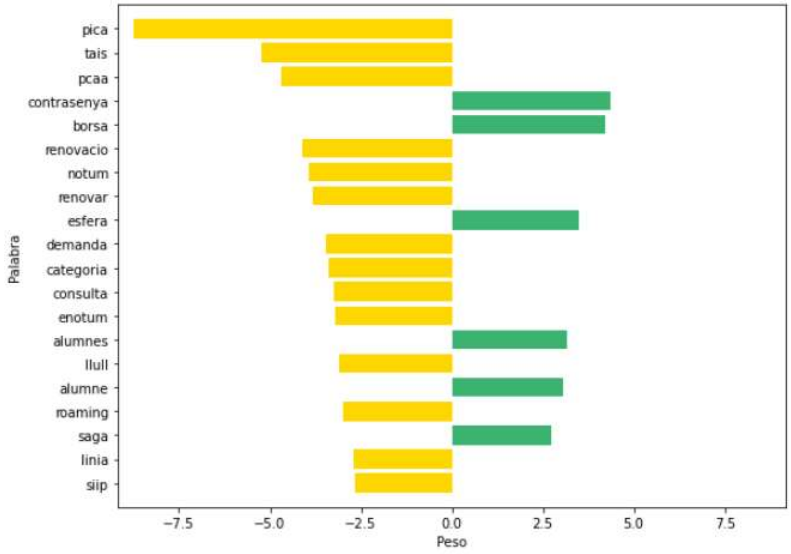


Tipo ticket	Clasificación correcta	
	Catalán	Español
Incidencia	99,4 %	99,9 %
No incidencia	19,3 %	4,1 %

Transparencia: Explicación global

Explicación del modelo 1 (¿es incidencia?): Ad-hoc

1. Calcula el peso de cada palabra a favor de INCIDENCIA o NO INCIDENCIA
2. Calcula la predisposición por defecto del modelo a hacer una u otra clasificación.
3. Suma ambas contribuciones
4. Calcula la probabilidad de ser incidencia



Transparencia: Explicación local

LIME

3.

Recomendador de libros

ALTRAN
Part of Capgemini

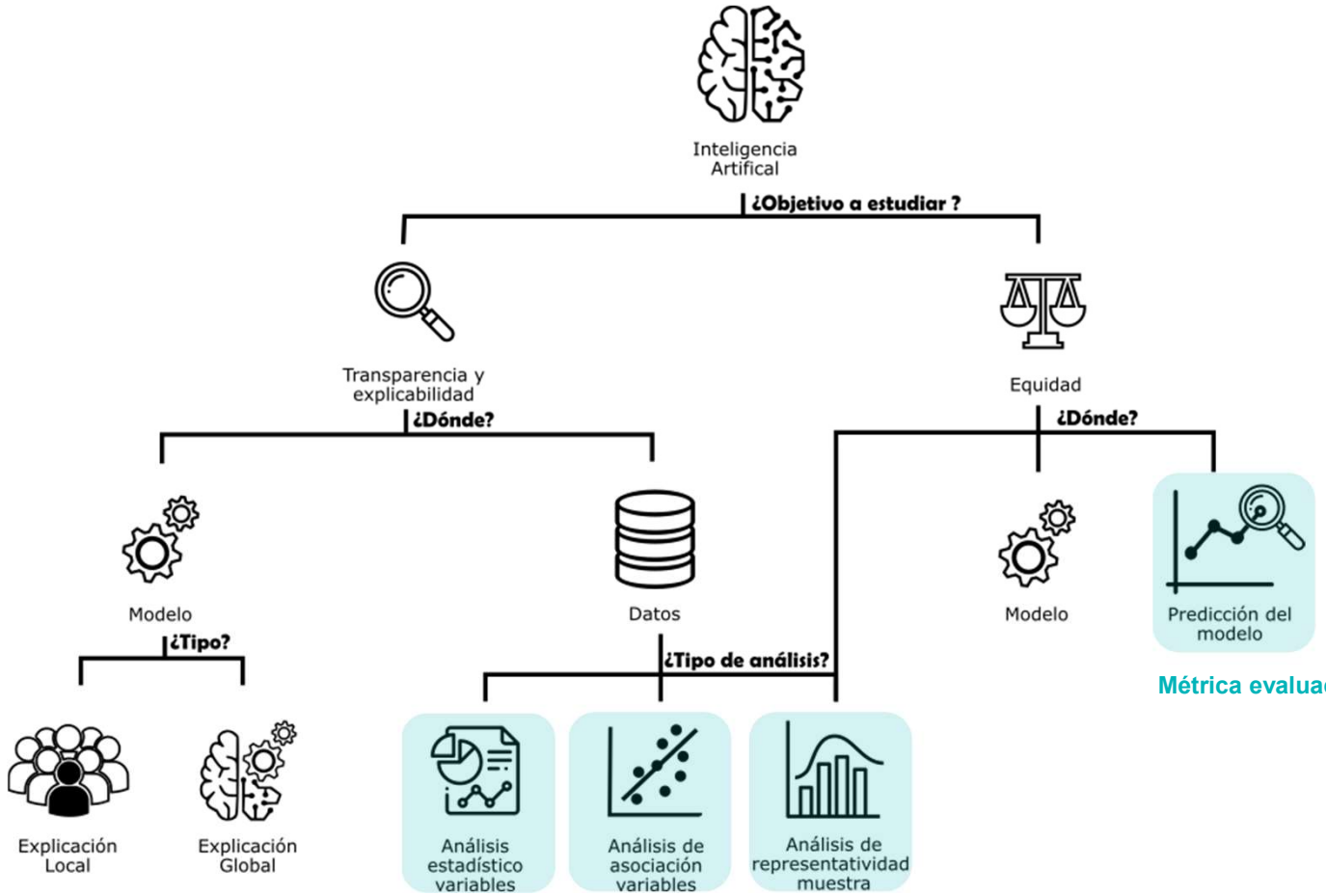
Recomendador de libros

El **modelo recomendador de libros recomienda 14 libros** al usuario referencia en base a lo que han leído los **40 usuarios más parecidos (en lecturas y no lecturas),** es decir los usuarios similares.

El algoritmo no es un algoritmo de machine learning como tal, simplemente una regla matemática simple.

A pesar de que el algoritmo no utiliza ningún dato sensible sobre los usuarios para generar las recomendaciones, el conjunto de datos contiene estas variables sensibles (edad, género, provincia, idioma, etc. de los usuarios).

Recomendador de libros



Métrica evaluación ad-hoc

Análisis de diferencias
Correlación
Regresión lineal
Regresión logistica múltiple

ALTRAN Part of Capgemini

Representatividad y Equidad

Para que el recomendador funcione eficazmente se ha de cumplir que:

- Entre todos los usuarios similares hayan leído 14 libros de más diferentes al usuario referencia, es decir que las recomendaciones no sean recomendaciones al azar o aleatorias.

- Los libros recomendados encajen con los gustos del usuario referencia.

Representatividad y Equidad: Análisis número de recomendaciones

Número de recomendaciones aleatorias, es decir recomendaciones no basadas en los libros leídos por los usuarios similares:

Número de libros leídos por el usuario referencia	Recomendaciones aleatorias
0-15	6 %
15-30	20 %
30-60	24 %
60-90	21 %
90-130	16 %
130-200	10 %
Más de 200	5 %

Representatividad y Equidad para usuarios:

Únicamente se ha detectado un trato desigual por edad y por género de usuario

Edad de usuarios

Debido a que el intervalo de edad de 12 a 20 años está sobrerrepresentado, las recomendaciones a usuarios de estas edades son más precisas.

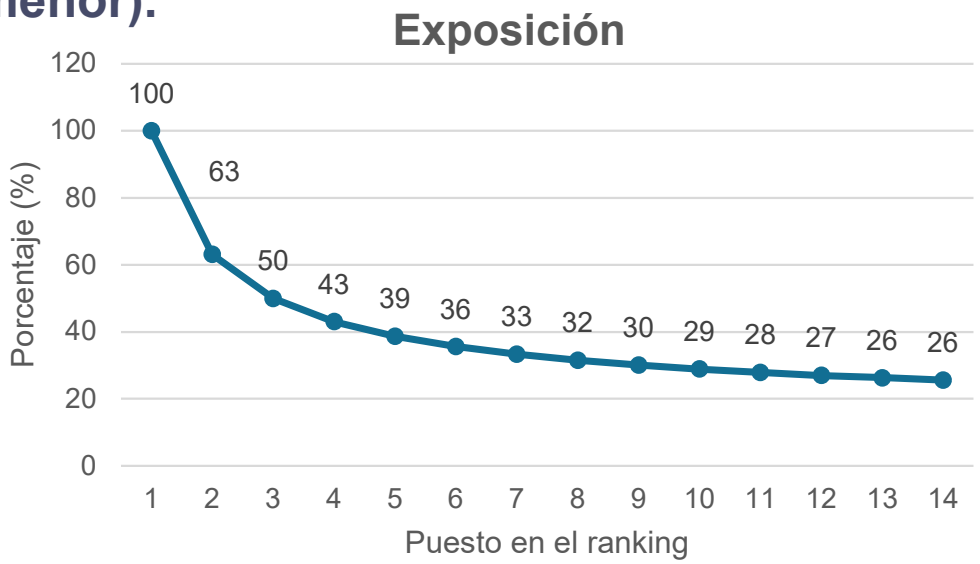
Género de usuarios

Las recomendaciones a mujeres son un 10 % menos precisas.

Representatividad y Equidad para autoras/autores:

El recomendador de libros promueve un trato desigual para los libros escritos por mujeres y hombres:

- **Los libros escritos por mujeres se recomiendan un 27,05 % menos que los escritos por hombres.**
- **En un 70,10 % de los rankings de recomendación las autoras reciben una menor exposición (22,48 % menor).**



4.

Resultados y conclusiones globales

ALTRAN
Part of Capgemini

Resultados y conclusiones globales

- Se ha realizado un estudio exhaustivo de las metodologías de transparencia, explicabilidad y equidad de los sistemas de inteligencia artificial.
- Se ha experimentado y validado la versatilidad y adecuación de las técnicas halladas en el estudio del arte con dos casos de estudio.
- Ambos casos de estudio han podido ser desarrollados utilizando una versión del modelo que no atenta contra la propiedad industrial de los proveedores y se han establecido los mecanismos de comunicación con el proveedor para intercambiar información.
- Se han descrito los múltiples pasos seguidos en el análisis y sus resultados en diferentes entregables que siguen la metodología de trabajo estandarizada CRISP-DM para un buen desarrollo de un proyecto de Inteligencia Artificial.
- En ambos casos de estudio los algoritmos son de baja complejidad y autoexplicables.
- Se ha confirmado la viabilidad de la técnicas seleccionadas y del framework.


ALTRAN
Part of Capgemini