Predictive modeling of materials: EDA, evaluation, simulation, and physical explanation

Harold Sebastián Rodríguez
UAX

November 16, 2025

Contents

1 EDA (Exploratory Data Analysis)

1.1 Semiconductor (Classification)

We use the mass range, which is a variable we created from the maximum and minimum of the mass.

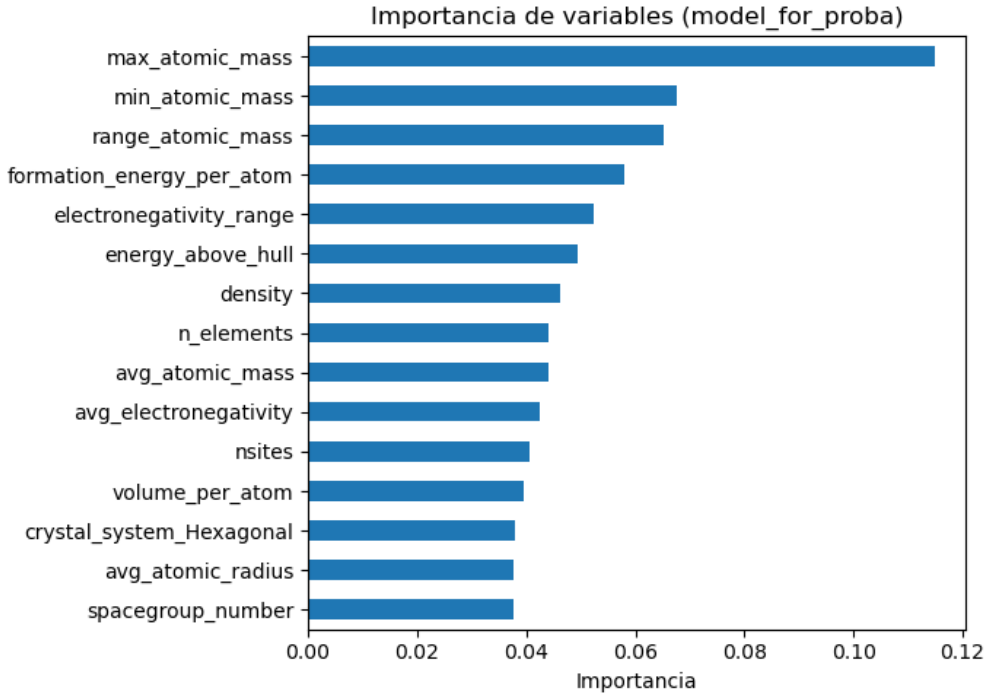


Figure 1: Model importances

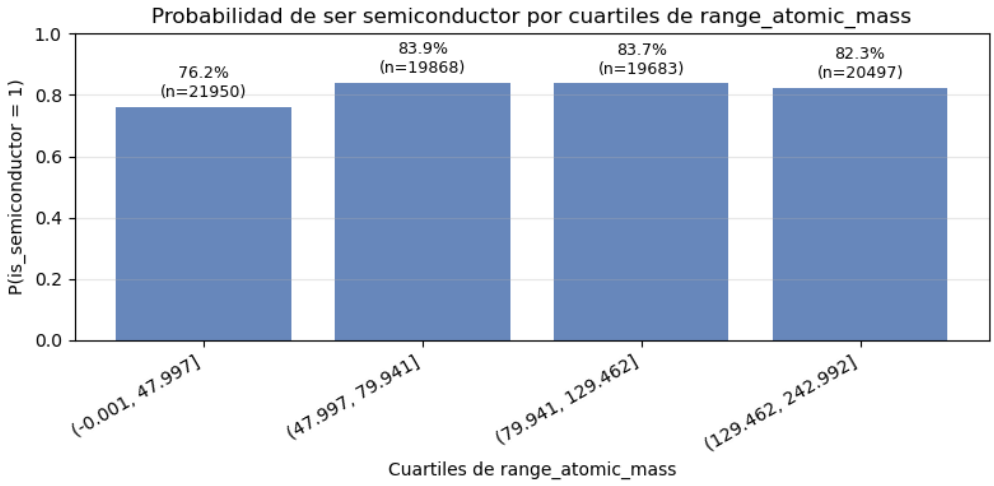


Figure 2: Mass vs semiconductor

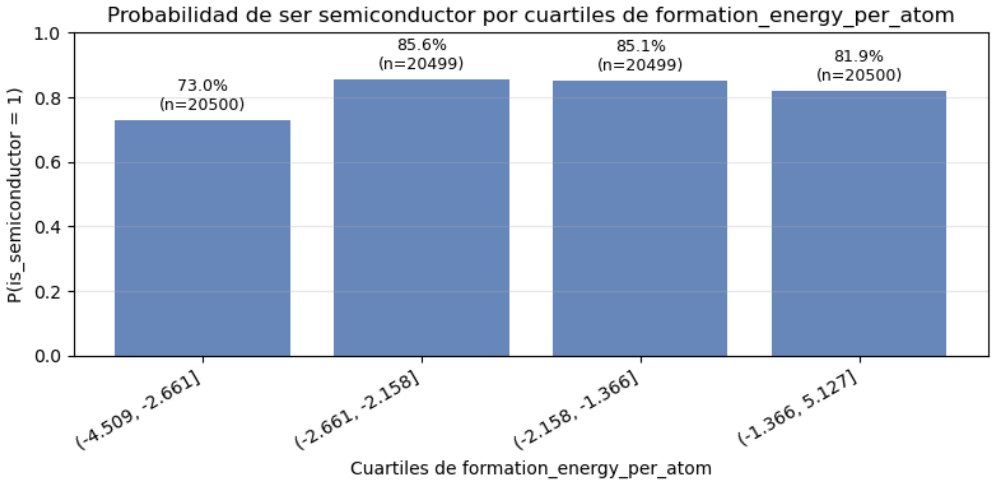


Figure 3: Energy vs semiconductor

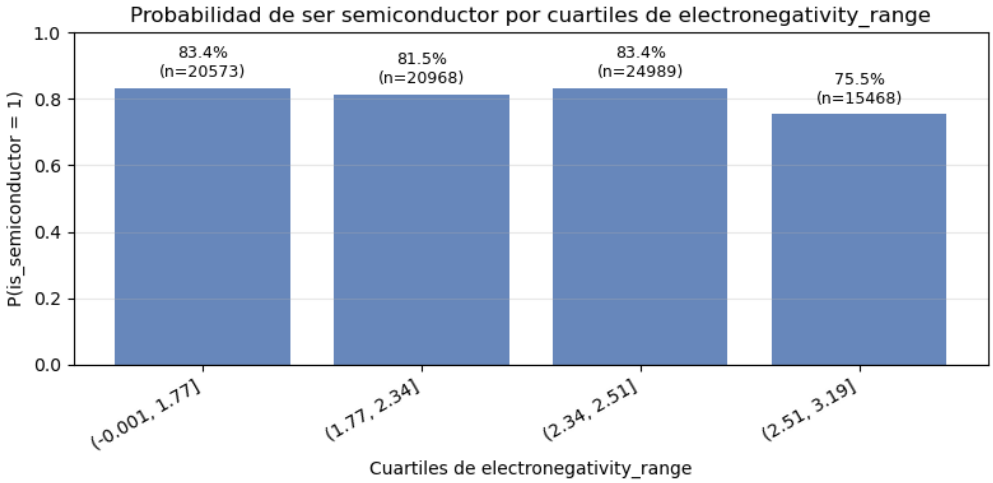


Figure 4: Electronegativity vs semiconductor

The probability that a material is a semiconductor can be explained from the interaction between three fundamental properties: **electronegativity**, **atomic mass**, and **formation energy per atom**.

The **electronegativity range** among the elements of a compound determines the predominant type of chemical bond. When this range is very high, the bonds tend to be ionic, generating a large separation between the valence and conduction bands, which leads to insulating materials. On the other hand, a very low range favors metallic bonds, with band overlap and the absence of a gap. Materials with an *intermediate* range present partially covalent bonds, opening a finite *band gap* characteristic of semiconductors.

The **atomic mass** influences the width of the electronic bands and the spin–orbit coupling. Heavier elements tend to narrow or split the bands, facilitating the appearance of a gap, while lighter ones tend to keep them wide, promoting metallic behavior. A mixture of moderate or high masses can therefore favor semiconducting behavior.

Finally, the **formation energy per atom** constrains the thermodynamic viability of the material. Only compounds with sufficiently negative formation energies can stabilize the crystal structure that gives rise to the gap.

Altogether, we hypothesize that the combination of an *intermediate electronegativity range*, *moderate or high atomic masses*, and a *sufficiently negative formation energy* defines a stability region where the material presents a suitable band gap, i.e., a **semiconducting** behavior.

1.2 Stability (Classification)

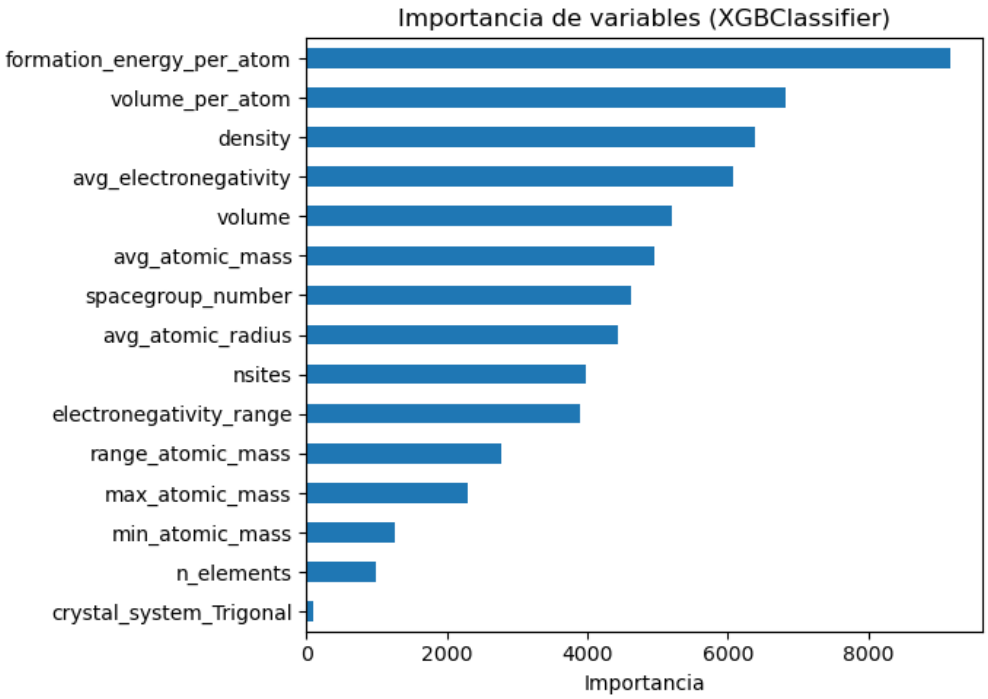


Figure 5: Model importances

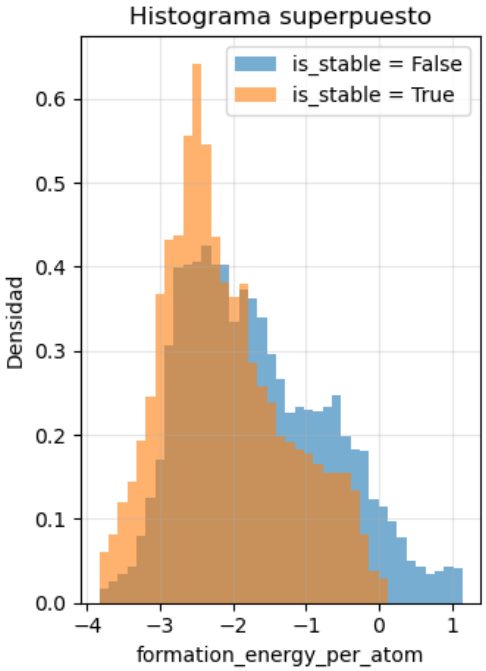


Figure 6: Energy vs stability

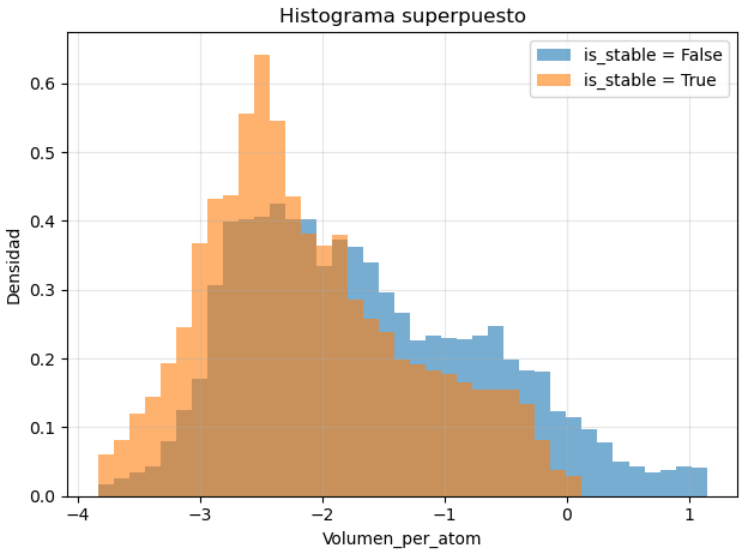


Figure 7: Volume vs stability

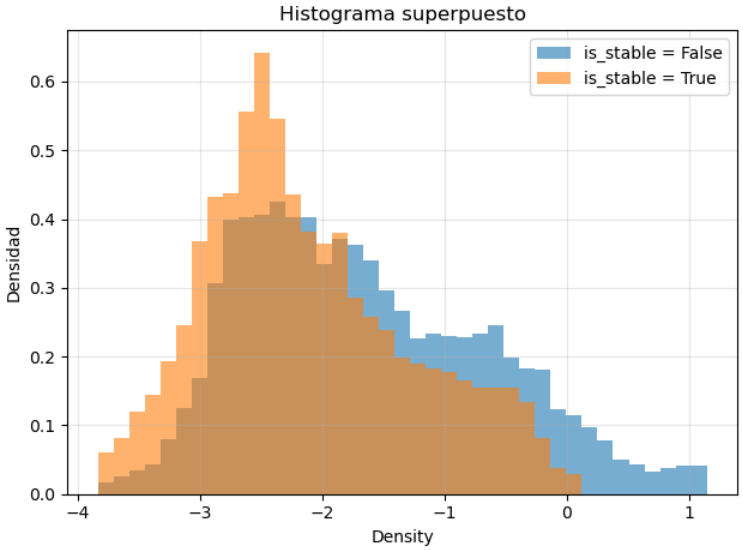


Figure 8: Density vs stability

Formation energy Most direct indicator of intrinsic stability: low values indicate more favorable phases that are more resistant to decomposition.

Volume Measures how "compressed" the structure is. At high pressures, phases that occupy less volume tend to be preferred; useful for comparing polymorphs.

Density Reflects the packing of the crystal lattice. Denser materials tend to be favored as pressure increases; it provides information similar to that of volume.

Joint interpretation The formation energy sets the baseline of how stable a phase is; the volume (and, in a correlated way, the density) indicates how this stability changes with pressure and which structure is more likely to dominate under different conditions.

1.3 Photovoltaic (Classification)

In the plots shown below, we compare the symmetry of the atom, since it seems to be an important variable in the model; later we will see if it makes physical sense together with the others.

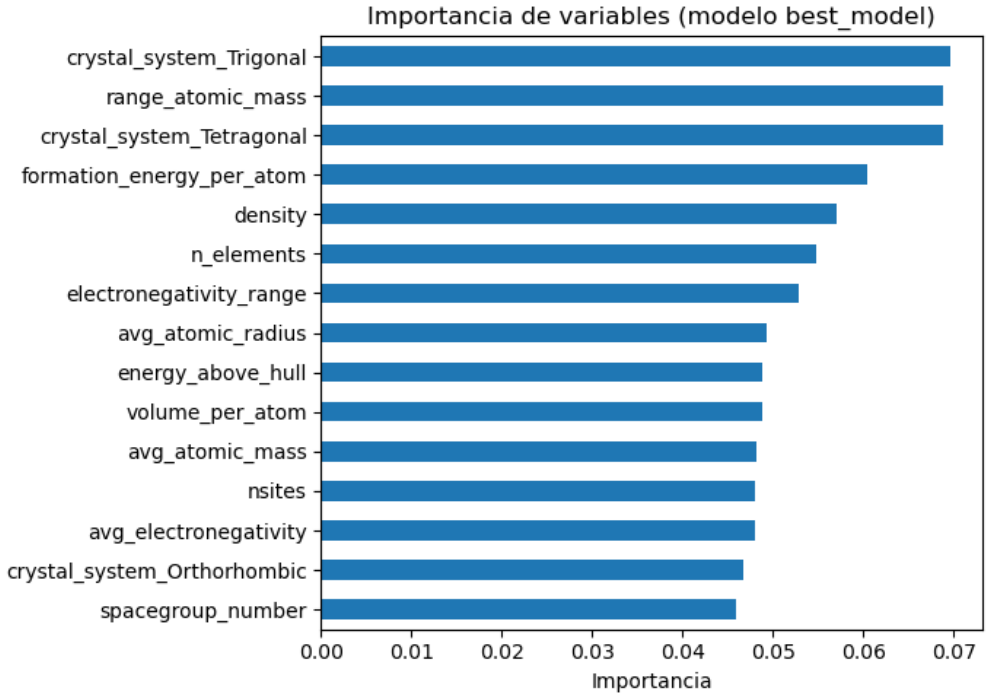


Figure 9: Model importances

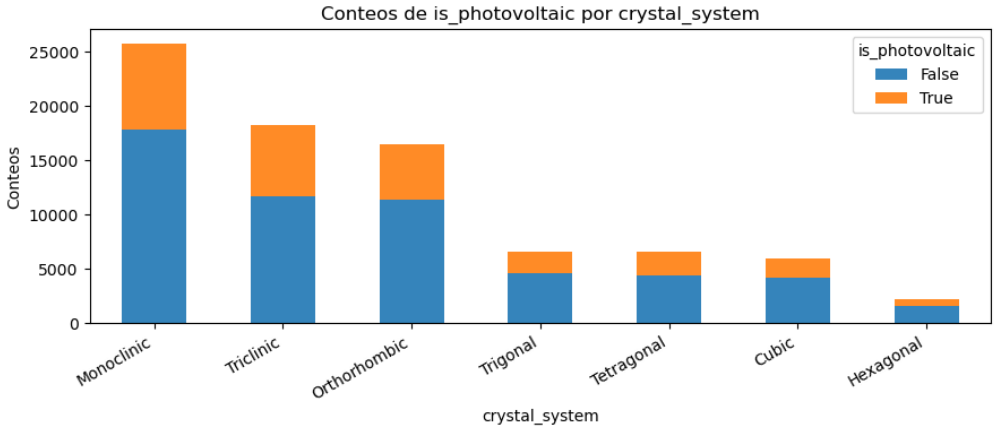


Figure 10: Symmetry vs photovoltaic

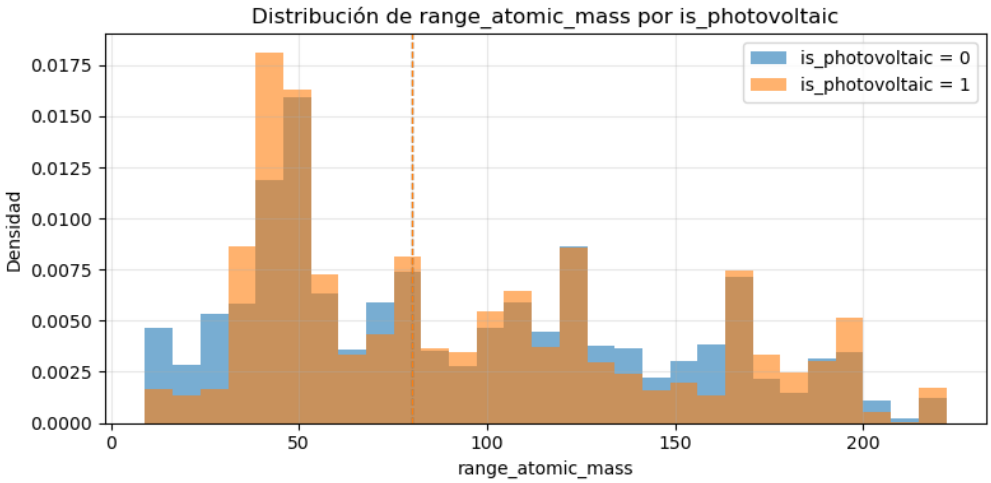


Figure 11: Mass vs photovoltaic

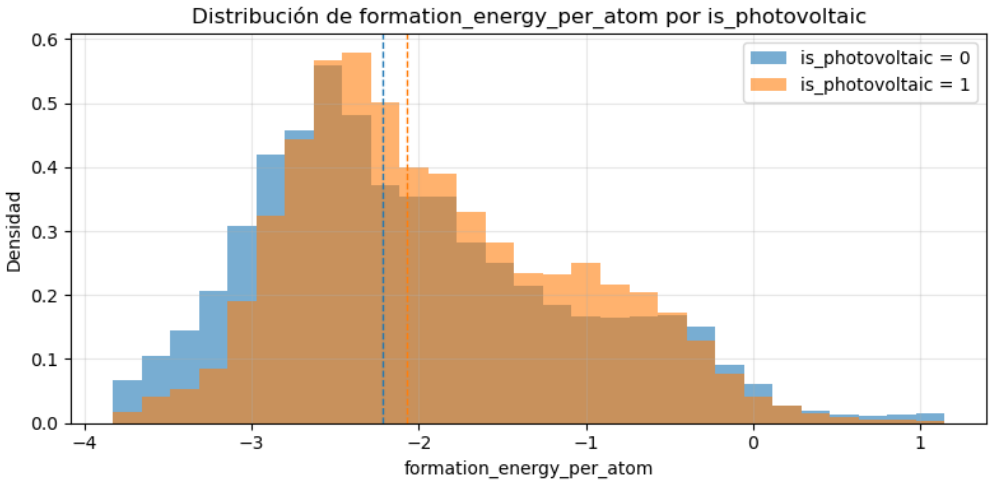


Figure 12: Energy vs photovoltaic

Symmetry. Determines whether the material has a **direct gap** and which optical transitions are **allowed**.

Mass range. Serves as an indicator of **heavy elements** and, therefore, of **spin–orbit coupling (SOC)**. High SOC can tune the **band gap** into the PV window.

Formation energy. Quantifies the **thermodynamic stability** and synthesizability of the phase. It correlates with **defect formation energies**: if deep traps are energetically costly, the material shows **defect tolerance**, which favors higher V_{oc} and long lifetimes.

Relationship. A good PV material combines: (i) **symmetry** that allows **efficient absorption** (direct gap and optically active transitions), (ii) **presence of heavy elements** (high SOC, dielectric screening, and favorable phonons) reflected in a high *mass range*, and (iii) sufficient **stability** (low *formation energy*) that limits the formation of **deep defects**. These three characteristics together provide coherent information to predict whether a material is PV.

1.4 Band gap (Regression)

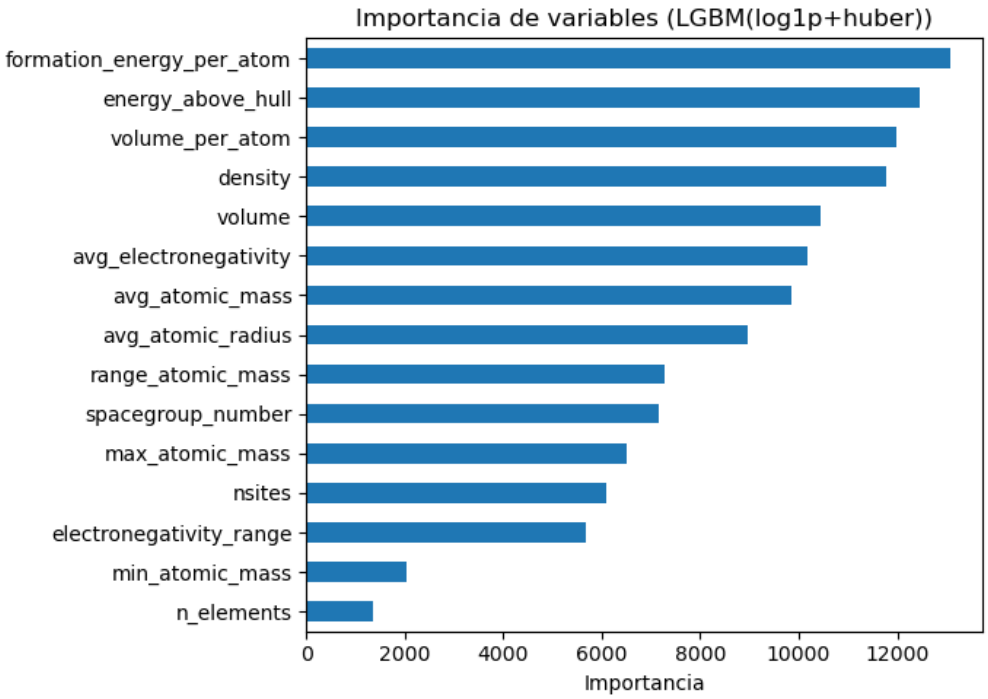


Figure 13: Model importances

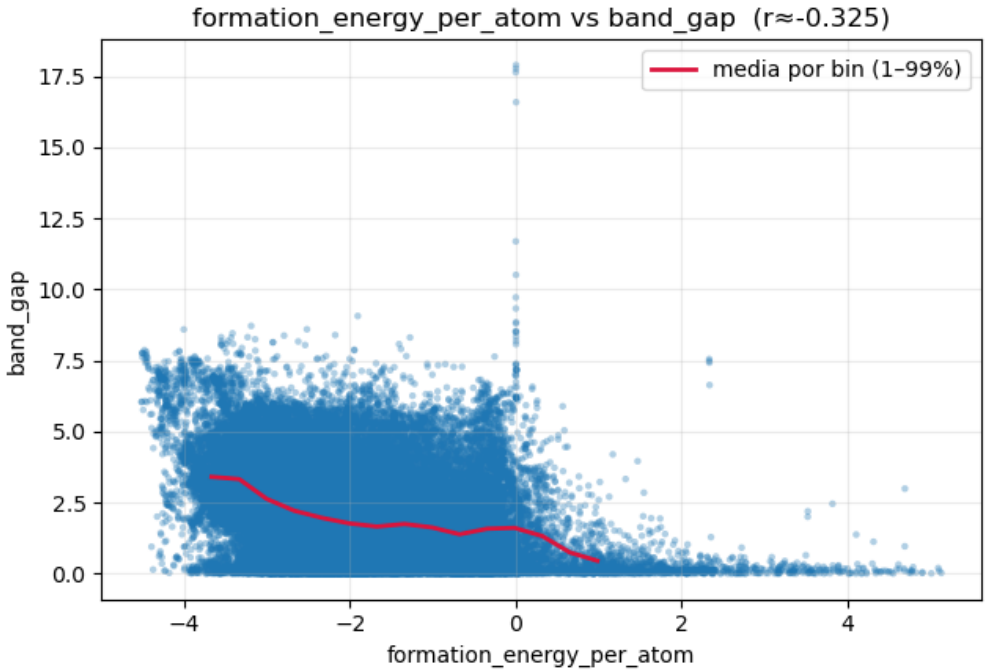


Figure 14: Formation energy vs band gap

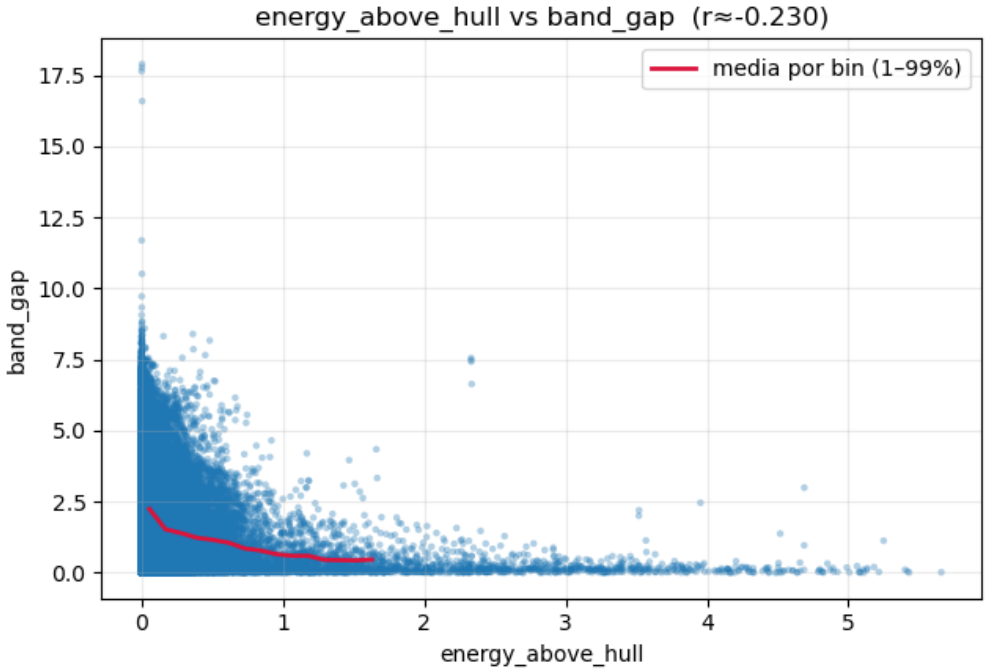


Figure 15: Energy above hull vs band gap

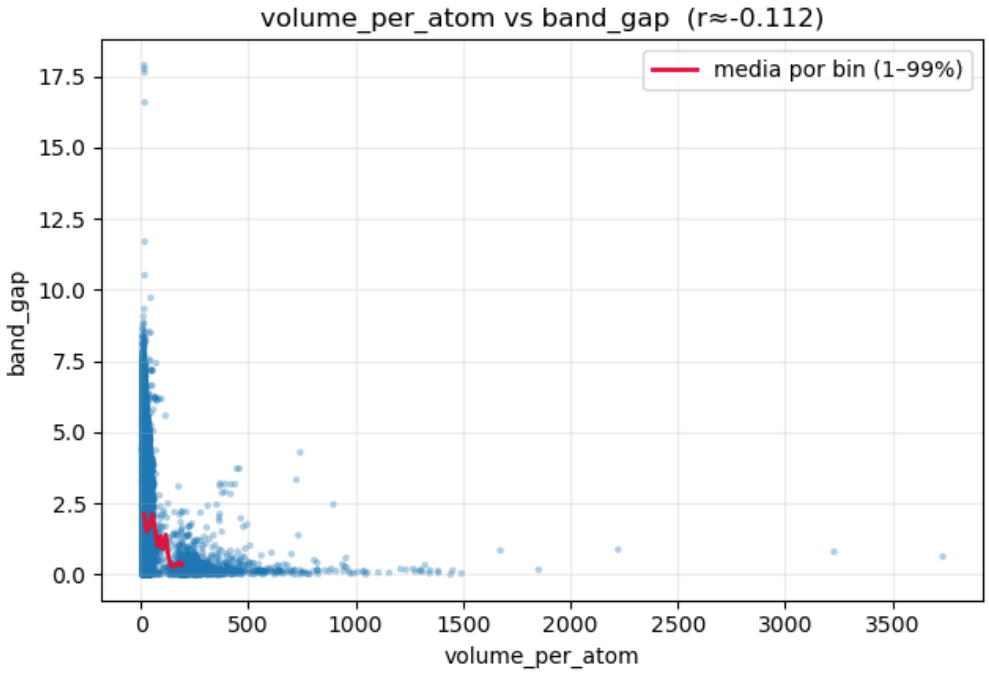
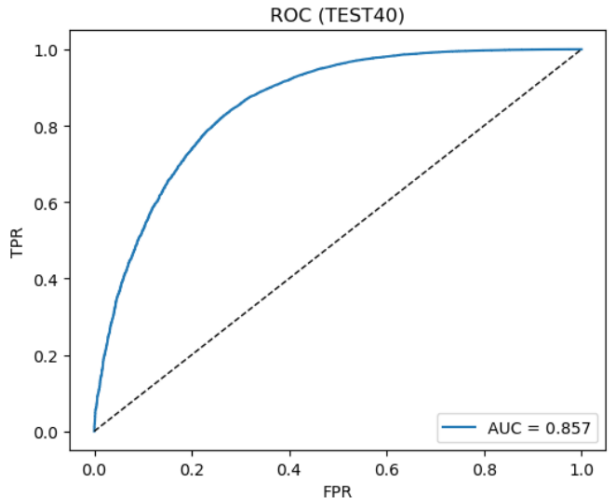


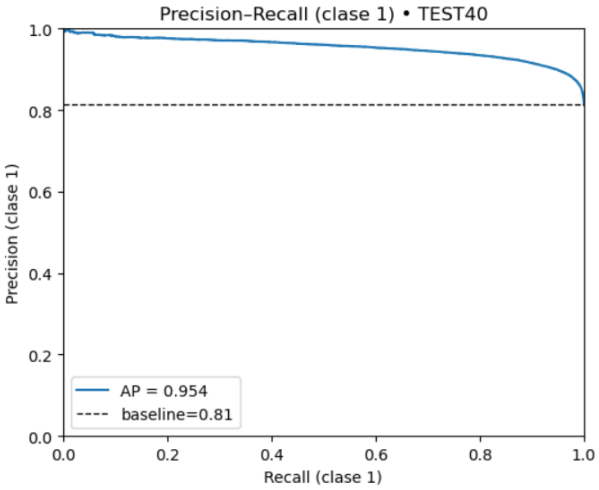
Figure 16: Volume vs band gap

2 Model evaluation

2.1 Semiconductor (Classification)



(a) ROC (AUC=0.857)



(b) Precision–Recall (AP=0.954)

Figure 17: Semiconductor (Classification): curves on TEST.

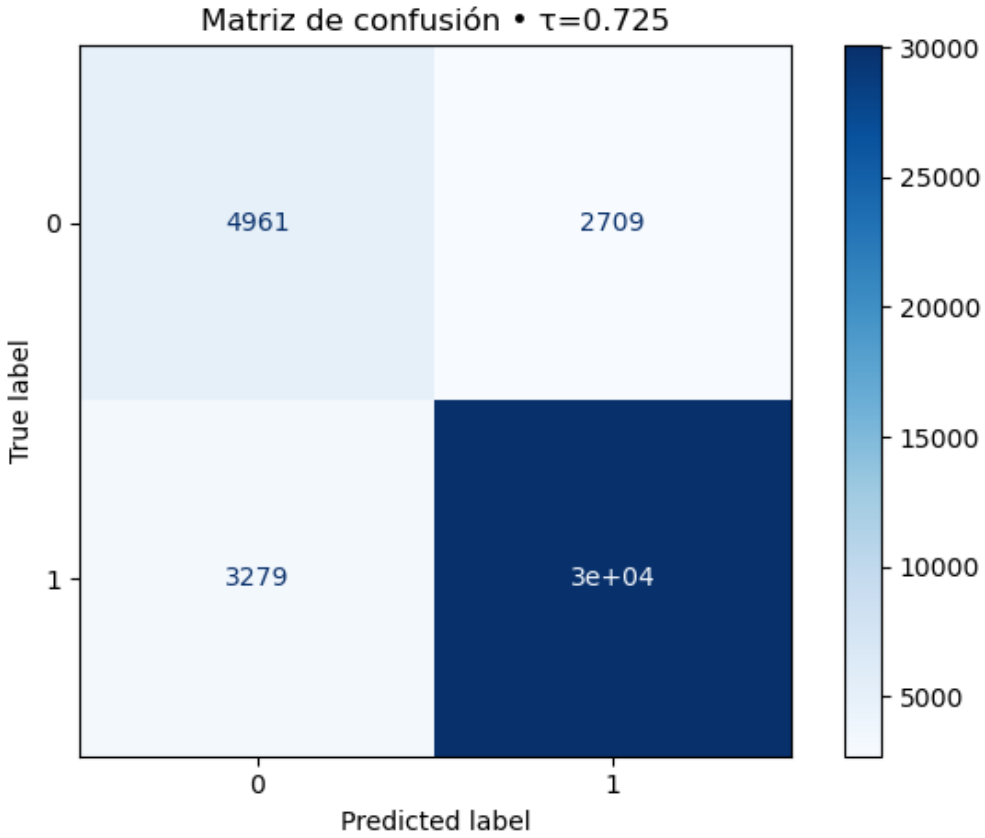


Figure 18: Confusion matrix

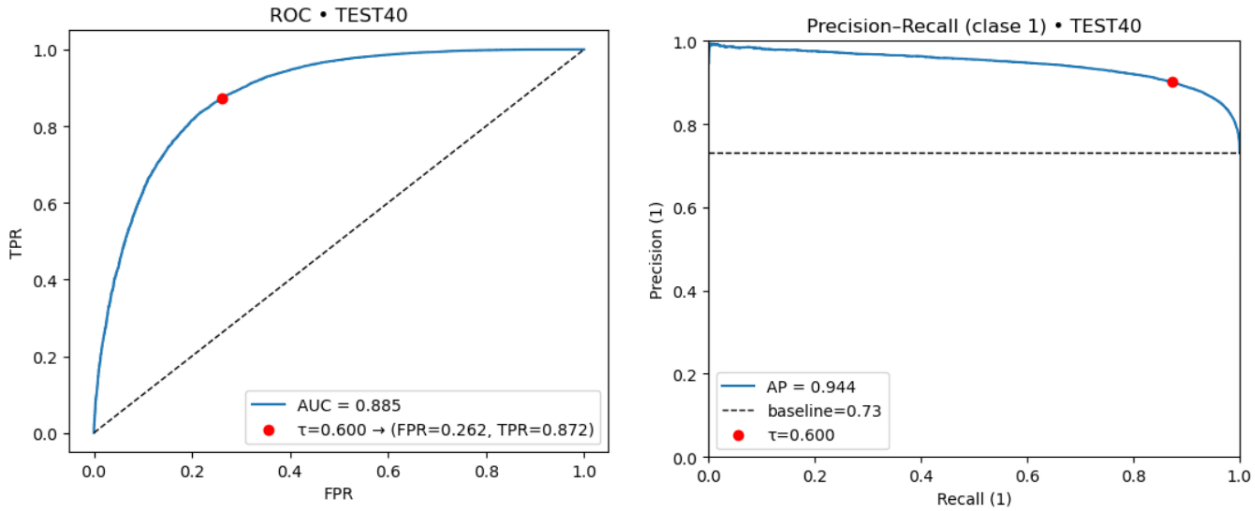
```
Reporte de clasificación:
              precision    recall  f1-score   support

           0      0.602     0.647     0.624      7670
           1      0.917     0.902     0.909     33329

    accuracy                          0.854     40999
   macro avg      0.760     0.774     0.767     40999
weighted avg      0.858     0.854     0.856     40999
```

Figure 19: Metrics (VALIDATION and TEST)

2.2 Stability (Classification)



(a) ROC (AUC=0.872)

(b) Precision–Recall (AP=0.944)

Figure 20: Semiconductor (Classification): curves on TEST.

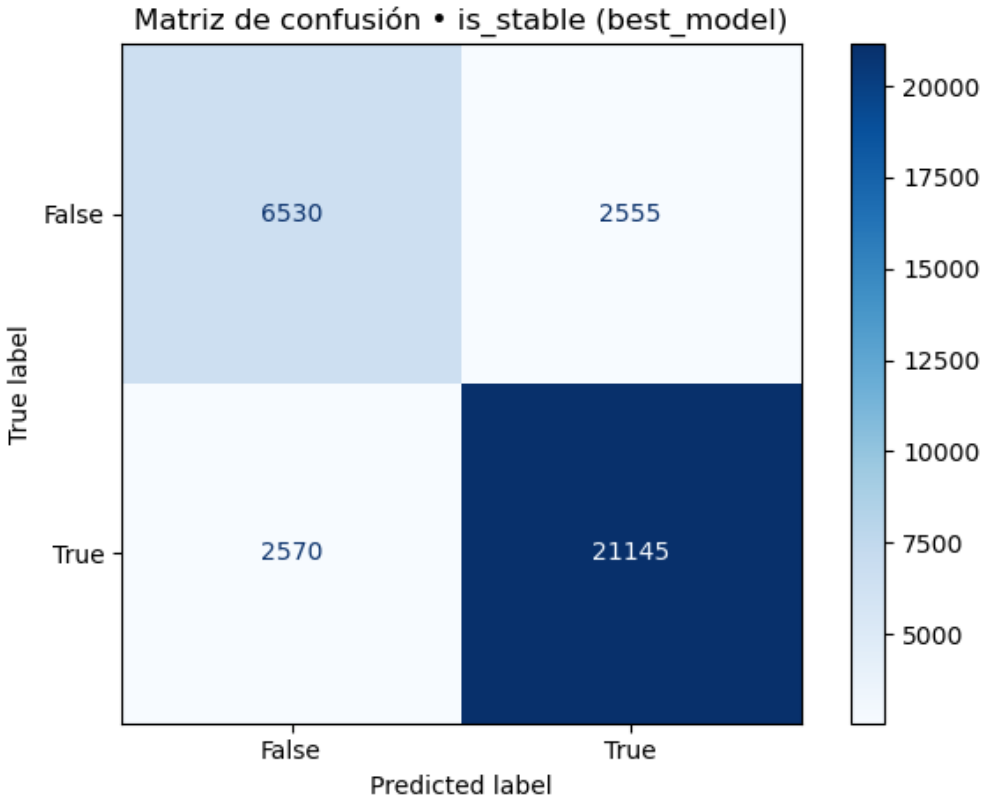
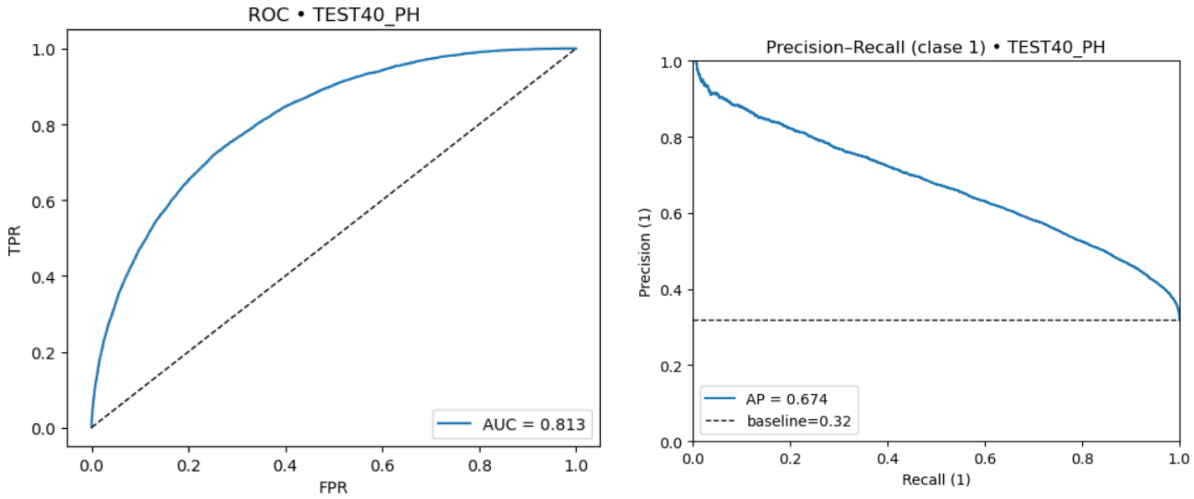


Figure 21: Confusion matrix

	precision	recall	f1-score	support
False	0.718	0.719	0.718	9085
True	0.892	0.892	0.892	23715
accuracy			0.844	32800
macro avg	0.805	0.805	0.805	32800
weighted avg	0.844	0.844	0.844	32800

Figure 22: Metrics (VALIDATION and TEST)

2.3 Photovoltaic (Classification)



(a) ROC (AUC=0.813)

(b) Precision–Recall (AP=0.674)

Figure 23: Semiconductor (Classification): curves on TEST.

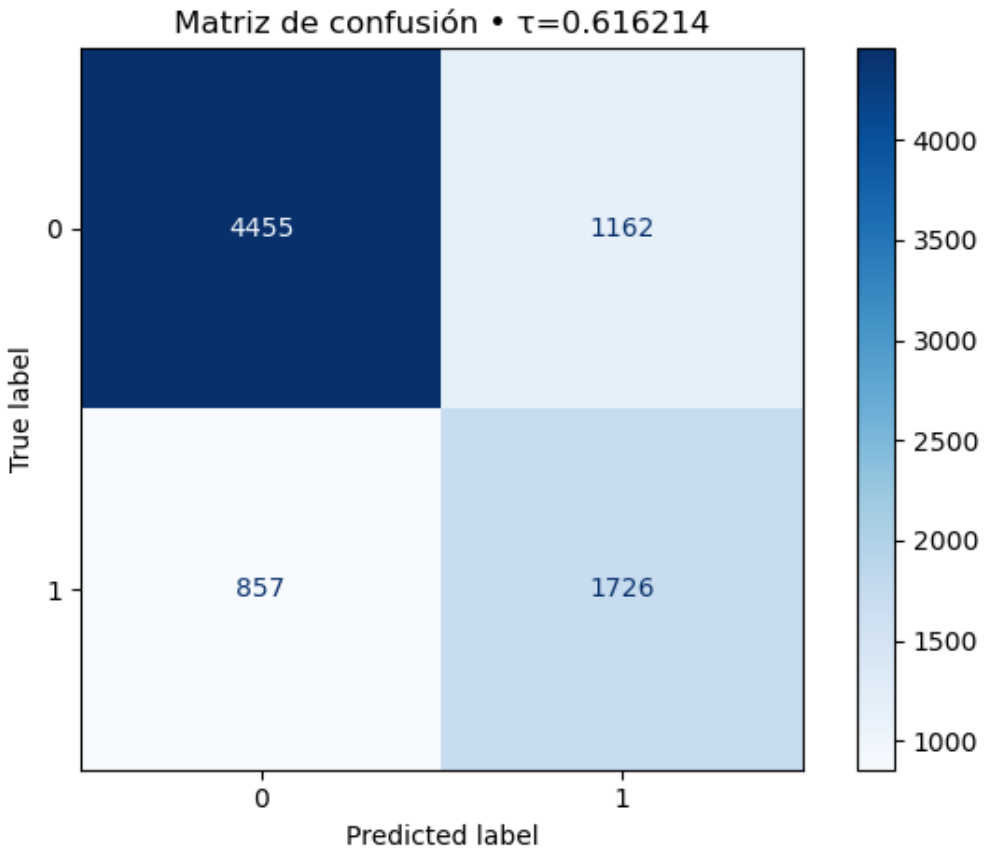
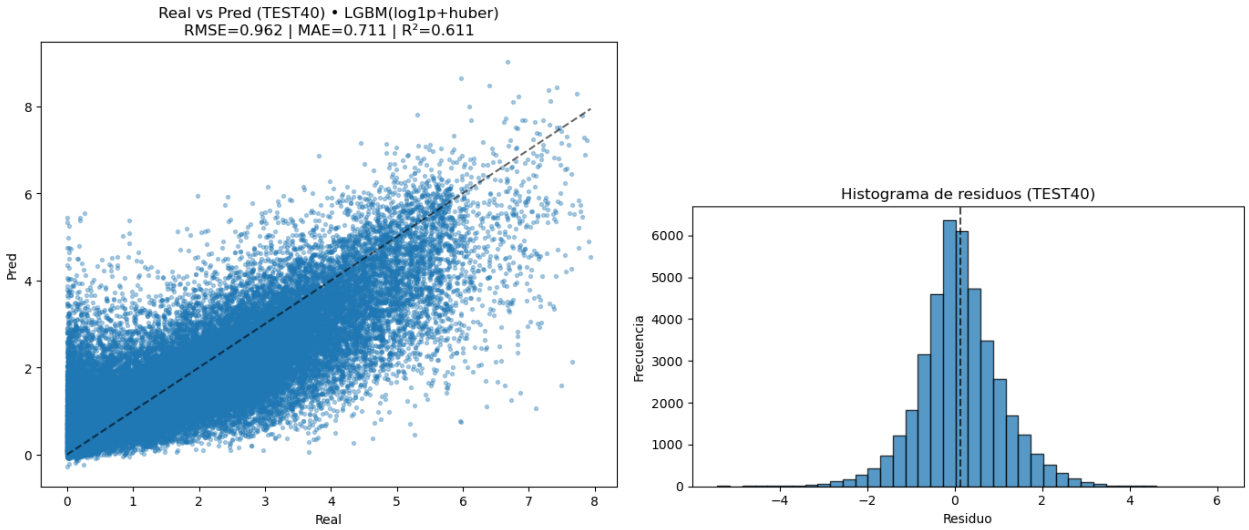


Figure 24: Confusion matrix

	precision	recall	f1-score	support
0	0.829	0.801	0.815	27862
1	0.607	0.650	0.628	13137
accuracy			0.753	40999
macro avg	0.718	0.726	0.721	40999
weighted avg	0.758	0.753	0.755	40999

Figure 25: Metrics (VALIDATION and TEST)

2.4 Band gap (Regression)



(a) Real vs pred (TEST)

(b) Residuals

Figure 26: Band gap (Regression): diagnostics on TEST.

3 Simulation with unseen data

In this section we evaluate our model with 40K previously unseen samples.

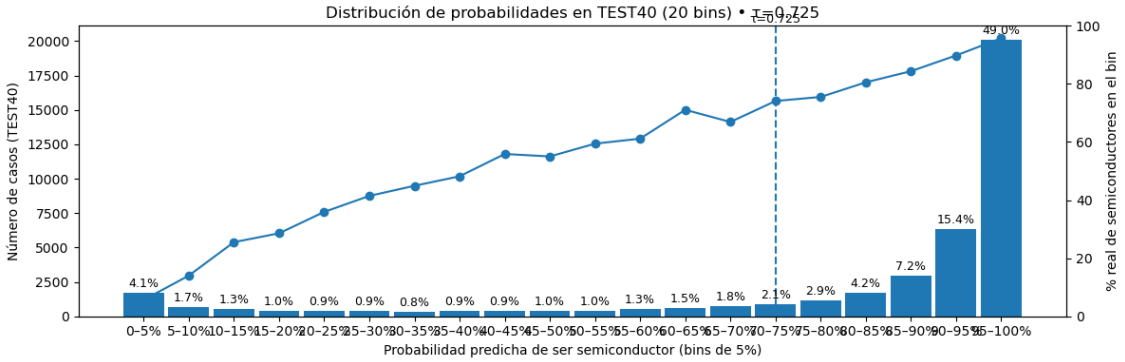


Figure 27: Calibration / Reliability (TEST40, Semiconductor (Classification)).

3.1 Stability (Classification)

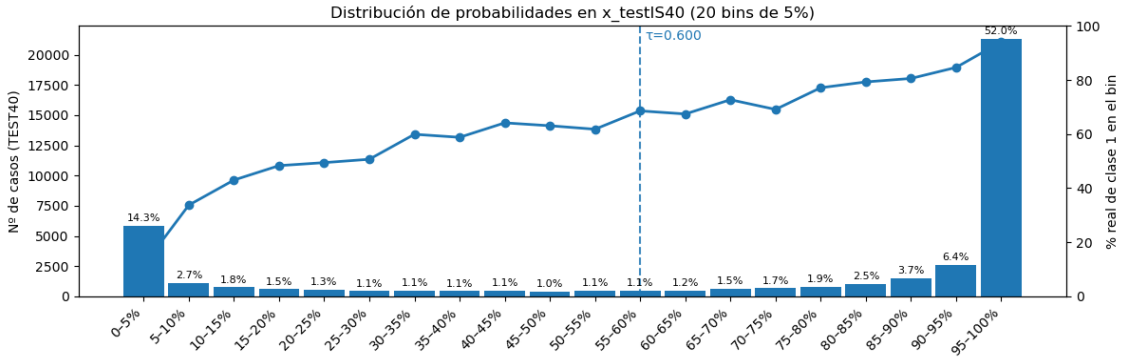


Figure 28: Calibration / Reliability (TEST40, Semiconductor (Classification)).

3.2 Photovoltaic (Classification)

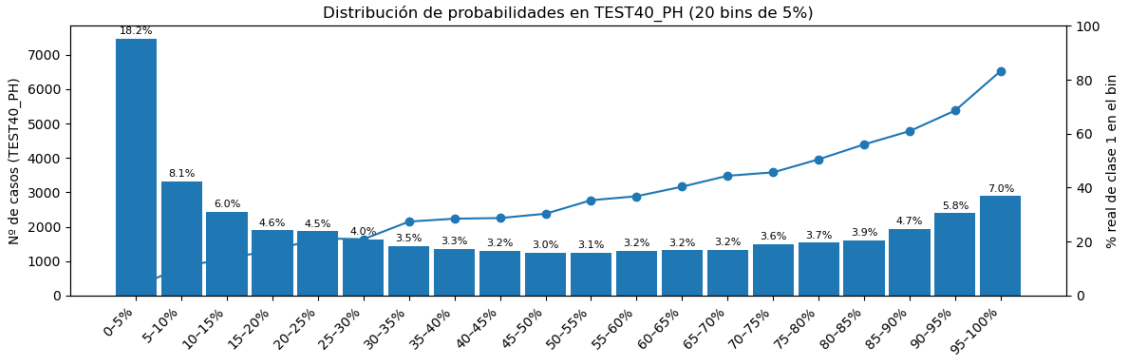
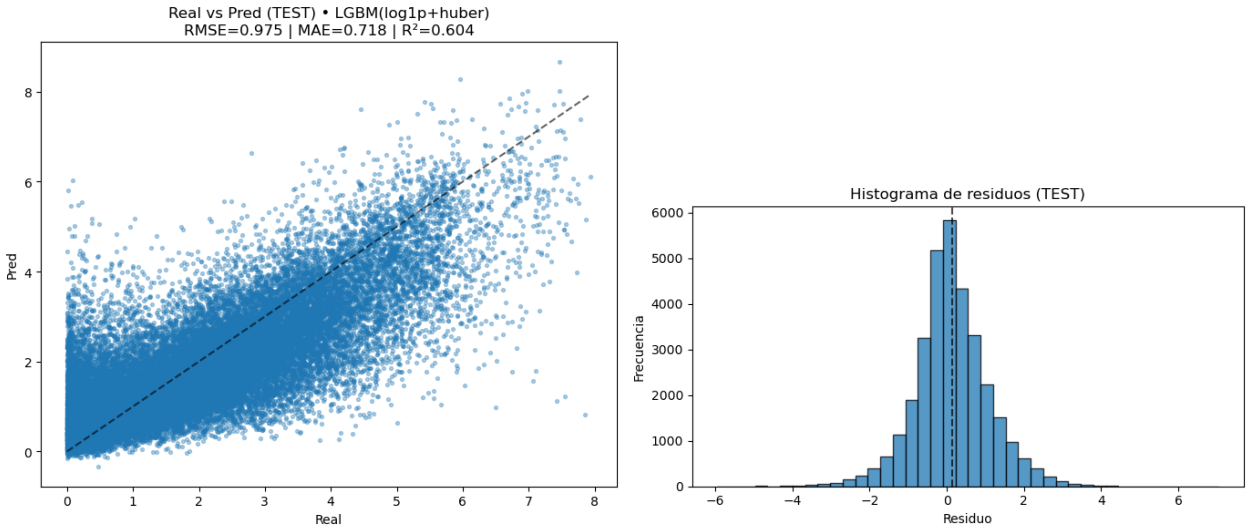


Figure 29: Calibration / Reliability (TEST40, Semiconductor (Classification)).

3.3 Band gap (Regression)



(a) Real vs pred (TEST40)

(b) Residuals

Figure 30: Band gap (Regression) on unseen data.

4 Physical explanation and interpretability

Physical explanation of semiconducting materials. A solid is a semiconductor if $E_g > 0$ and the intrinsic density is small:

$$n_i = \sqrt{N_c N_v}\, e^{-\frac{E_g}{2k_B T}} \quad \Rightarrow \quad \text{if } E_g \gg k_B T \text{ (at 300 K), it is not metallic.}$$

Electronegativity range $\Delta\chi$ (bond type $\rightarrow E_g$). The fraction of ionicity (Pauling/Phillips–Van Vechten) increases with $\Delta\chi$:

$$f_{\mathrm{i}} \;\approx\; 1 - \exp\Big[-\frac{(\Delta\chi)^2}{4}\Big],$$

and the *band gap* increases with ionicity and decreases with covalency (more *hopping*). In a *tight-binding* model:

$$E_g \;\approx\; \underbrace{(\varepsilon_A - \varepsilon_B)}_{\text{polarity}\,\propto\,\Delta\chi} \;- 2|t(d)|, \qquad t(d) \simeq t_0\, e^{-\alpha d}.$$

Larger $\Delta\chi \Rightarrow$ larger polarity and, typically, larger E_g.

Atomic mass / atomic number (SOC and bandwidth). The spin–orbit coupling scales as

$$\Delta_{\mathrm{SO}} \;\propto\; \frac{Z^4}{n^3},$$

and can open or modify the gap near the VBM/CBM:

$$E_g' \simeq E_g \pm \Delta_{\mathrm{SO}}.$$

Moreover, heavy atoms usually give stronger screening and narrower bands, which (in the absence of large *hopping*) favors $E_g > 0$.

Formation energy per atom E_f (stability and defects). Materials with more negative E_f are thermodynamically accessible and tend to exhibit higher defect formation energies:

$$E_f^{\mathrm{def}} = E_{\mathrm{def}} - E_{\mathrm{bulk}} - \sum_i n_i \mu_i + q(E_F + E_{\mathrm{v}}) + \Delta,$$

reducing the density of deep states that would close the effective gap. Thus, E_f does not set E_g, but it filters "good" phases (clean gap, non-metallic).

Conclusion.

$\Delta\chi$ (polarity) $\uparrow \Rightarrow$ $E_g \uparrow$; Z (SOC) $\uparrow \Rightarrow$ E_g' tuned ; $E_f \downarrow \Rightarrow$ stable phase with realizable gap

Physical explanation of stable materials

1) *Formation energy.* This is the *direct* indicator of chemical stability. A more negative formation energy means that the phase, when formed from its references, releases more energy and therefore *"does not benefit"* from decomposing. At equilibrium, the observed phases are those with lower free energy; at moderate pressure and not very high temperature,

$$G \simeq E_{\mathrm{form}} \quad \Rightarrow \quad \text{low } E_{\mathrm{form}} \;\Rightarrow\; \text{stable phase.}$$

2) *Volume.* Stability also depends on pressure through

$$G = U + PV - TS, \qquad \left(\frac{\partial G}{\partial P}\right)_T = V.$$

At higher pressure, phases with *smaller volume per atom* decrease their G more and are favored. This variable clearly distinguishes between *polymorphs* of the same composition.

3) *Density.* For a given composition, high density $\leftrightarrow$ low molar volume. Therefore, the *densest* phases are usually preferred as pressure increases, and they reflect more compact packing (greater cohesion).

Physical explanation of photovoltaic materials

1) *Symmetry.* The crystal symmetry influences the type of *band gap*. In materials with a **direct gap**, optical absorption is more efficient:

$$\alpha_{\text{direct}} \propto \sqrt{h\nu - E_g},$$

whereas in **indirect** ones, phonons are required and absorption is weaker. Therefore, certain crystal systems (such as non-centrosymmetric ones) favor direct transitions and better solar absorption.

2) *Atomic mass.* A wide mass range implies the presence of heavy elements, associated with stronger spin–orbit coupling (SOC):

$$\Delta_{\text{SO}} \propto Z^4.$$

This effect can tune the gap size within the ideal window for photovoltaic conversion (1–2 eV) and improve defect screening, reducing recombination.

3) *Formation energy.* Indicates thermodynamic stability. If the formation energy is low (more negative), the phase is stable and less prone to generating defects:

$$c_{\text{def}} \propto e^{-E_f^{\text{def}}/k_B T}.$$

Fewer deep defects imply lower recombination and higher output voltage.

Physical explanation of materials with band gap

1) Formation energy per atom. Reflects the stability and the type of bonding between atoms. Stronger bonds (more negative formation energy) are usually associated with a greater separation between bonding and antibonding levels, which produces a wider gap:

$$E_g \propto |\varepsilon_{\text{antibonding}} - \varepsilon_{\text{bonding}}|.$$

Therefore, very stable materials tend to have a larger band gap.

2) Energy above hull. Indicates the relative stability with respect to other possible phases. Phases close to the hull ($E_{\text{above hull}} \approx 0$) are those that can actually exist. Although it does not directly determine the value of the gap, it ensures that the computed gap corresponds to a physically realizable structure rather than an unstable phase.

3) Volume per atom. The volume affects interatomic distances and therefore orbital overlap. When the volume decreases (atoms closer together), the overlap increases and the gap tends to shrink:

$$t(d) \propto e^{-\alpha d}, \qquad E_g \approx (\varepsilon_A - \varepsilon_B) - 2|t(d)|.$$

Conversely, more open structures (larger volume) usually present a larger band gap.