PROJET DATA · VAL-DE-MARNE · DPE ADEME 2026

Projet DPE : À quel point l’ajout d’informations simples améliore-t-il l’estimation de la lettre DPE d’un logement ?

Mesurer l'apport d'un questionnaire à une estimation indicative de la lettre A–G, avant un premier conseil en rénovation.



Nylsevens Lalanne

Oral · octobre 2026

SOMMAIRE

Cinq étapes, une question

01

Le contexte et les données

Usage visé, source ADEME, population et découpage

02

Les informations testées

Trois paliers : 3, 14 puis 18 informations

03

La modélisation

Deux modèles, six candidats

04

Les résultats

Validation dans le train, puis test final

05

Choix et pistes

Candidat retenu, limites, suites

Fil rouge : **que gagne-t-on, en précision, à ajouter quelques informations simples pour estimer la lettre d'un logement ?**

01 · LE PROBLÈME

Mesurer l'apport de quelques réponses avant le rendez-vous

Un habitant sans DPE récent

Il prépare un rendez-vous avec une maison de la rénovation du Val-de-Marne.

Quelques réponses simples

Le conseiller connaît le logement à grands traits et peut le compléter par un questionnaire.

Une lettre A–G indicative

Pour préparer un premier conseil, toujours avec vérification humaine.

CE QUE LE PROJET NE FAIT PAS

- Pas un DPE officiel
- Pas de décision automatique de travaux
- Un DPE pertinent existant reste prioritaire

Provenance des données et découpage du dataset

SOURCE · OPEN DATA ADEME

DPE du Val-de-Marne établis de janvier à septembre 2026 · une ligne = un diagnostic

41 189
lignes × 218 colonnes

POPULATION RETENUE

39 609 sans les DPE d'immeuble entier

27 403 maisons et appartements · DPE individuel · 9–500 m²

Train · 21 860 logements — F + G : 17,6 %



Test · 5 543 logements — F + G : 16,7 %



DEUX DÉCOUPAGES, DEUX RÔLES

1 Mettre le test de côté · une seule fois



Équilibré par lettre · logements d'une même adresse gardés ensemble · graine 42 : même tirage à chaque relance · filtres appliqués **après**.

2 Dans le train seulement



Validation croisée en **5 plis** : 5 tours, chaque pli valide une fois. Le test ne sert qu'**à la fin**.

CIBLES JAMAIS EN ENTRÉE — **Étiquette** A–G · **Consommation** kWhEP/m²/an · **Émissions** $kgCO_2éq/m^2/an$

02 · LES INFORMATIONS TESTÉES

Trois paliers de renseignements cumulatifs

AVANT L'ENTRETIEN

3

— Période de construction

— Maison ou appartement

— Surface habitable

QUESTIONNAIRE · +11

14

— Énergie et appareil de chauffage

— Installation individuelle ou collective

— Eau chaude : énergie, installation, même équipement que le chauffage ou équipement séparé

— Hauteur, nombre de niveaux

— Présence d'un chauffage secondaire, climatisation, ventilation

ISOLATION · +4

18

— Toiture, murs

— Fenêtres, plancher bas

Exploratoire : difficile à connaître sans document.

3, 14, 18 : des conséquences, pas des objectifs. Identifiants et résultats du DPE ne sont jamais des entrées.

Deux modèles, un même algorithme : la forêt aléatoire

Pourquoi la forêt aléatoire : adaptée aux données tabulaires et aux effets combinés entre caractéristiques.
Elle existe en classification et en régression, et reste robuste avec peu de réglages.
On compare deux façons d'estimer la lettre, pas deux algorithmes.

Modèle 1 · Lettre prédite

OUTIL — **RandomForestClassifier** : 1 forêt choisit une lettre A–G.

ATOUT — Répond directement à la question : une lettre parmi sept.

LIMITE — Toutes les erreurs se valent : A → B compte comme A → G. Peut favoriser les lettres fréquentes.

Modèle 2 · Lettre calculée

OUTIL — **2 × RandomForestRegressor** : kWhEP/m² et $kgCO_2éq/m^2$, puis barème → A–G.

ATOUT — Garde l'information chiffrée derrière la lettre : deux logements D peuvent être très différents.

LIMITE — Une petite erreur près d'un seuil peut faire changer de lettre.

Aucune des deux ne cherche directement à maximiser la détection des passoires thermiques F/G : **on mesure le compromis obtenu**.

Réglages identiques sur les 3 paliers · 200 arbres par forêt · lettre prédite : profondeur ≤ 20, feuilles ≥ 5 · lettre calculée : feuilles ≥ 10

3 paliers × 2 modèles = 6 candidats comparés

INFORMATIONS FOURNIES	MODÈLE 1 · LETTRE PRÉDITE **RandomForestClassifier** → lettre A–G Une seule forêt choisit directement la lettre.	MODÈLE 2 · LETTRE CALCULÉE **2 × RandomForestRegressor** → 2 lettres → **la pire** Règle du DPE : une bonne note en énergie ne masque pas de fortes émissions.
3 **Avant entretien** informations	**Candidat 1** · lettre prédite, 3 infos	**Candidat 2** · lettre calculée, 3 infos
14 **Questionnaire** informations	**Candidat 3** · lettre prédite, 14 infos	**Candidat 4** · lettre calculée, 14 infos
18 **+ Isolation** informations	**Candidat 5** · lettre prédite, 18 infos	**Candidat 6** · lettre calculée, 18 infos

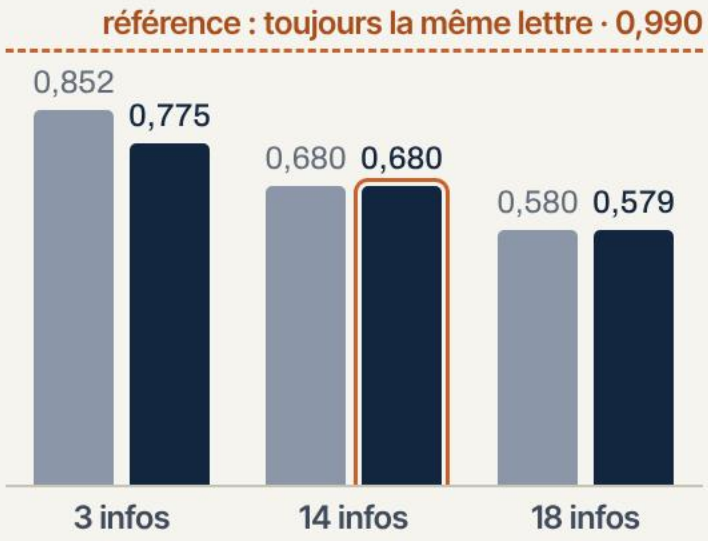
Mêmes logements, mêmes 5 plis, mêmes réglages pour un modèle donné : **seules les informations changent.**

Validation : chaque palier améliore l'estimation

Lettre prédite · Lettre calculée · Candidat retenu, choisi ici

Erreur moyenne (MAE)

en lettres



−31 % d'erreur à 14 infos (0,990 → 0,680) par rapport à une référence qui prédit toujours la lettre la plus fréquente.

Rappel F/G

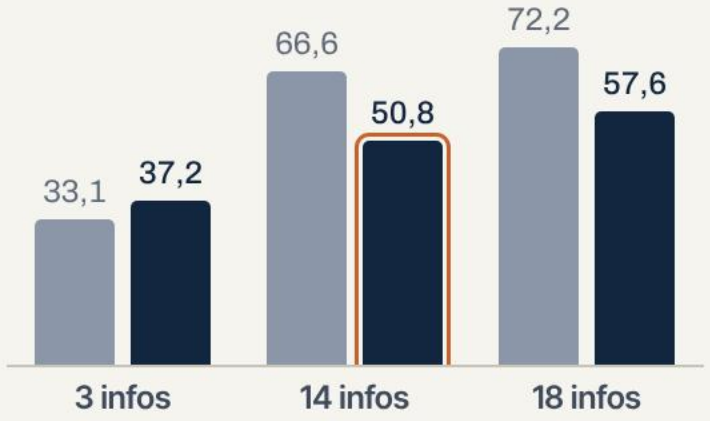
% des vrais F/G retrouvés



Lettre calculée : rappel presque ×3 de 3 à 14 infos.

Précision F/G

% des alertes F/G justes



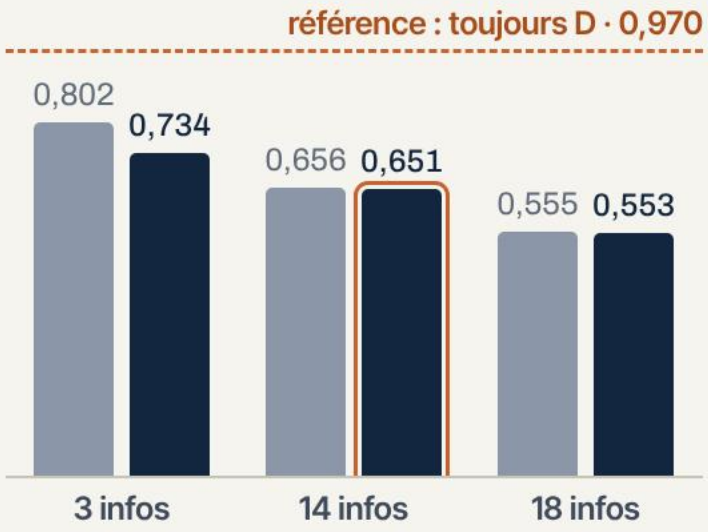
Lettre prédite : précision doublée de 3 à 14 infos.

Test : mêmes tendances, deux compromis

Lettre prédite · Lettre calculée · Candidat retenu en validation

Erreur moyenne (MAE)

en lettres · plus bas = mieux



MAE quasi identiques entre approches à 14 et 18 infos.

Rappel F/G

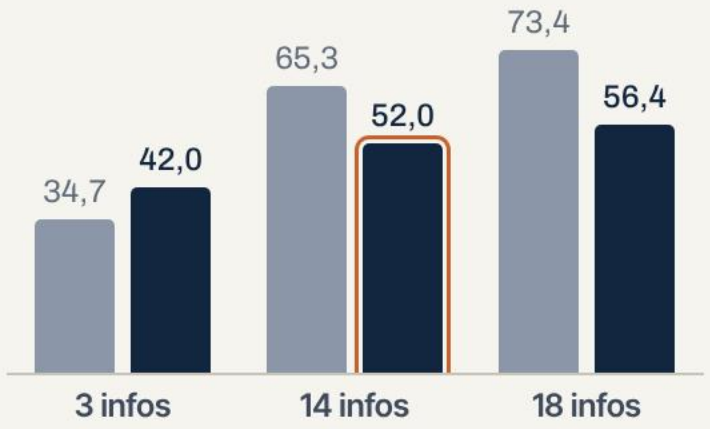
% des vrais F/G retrouvés



Lettre calculée : 36 % des F/G retrouvés au questionnaire, contre 20 %.

Précision F/G

% des alertes F/G justes



Lettre prédite : 65 % d'alertes justes au questionnaire, contre 52 %.

⚠ Au test, sur 100 vrais F, le candidat retenu (lettre calculée, questionnaire) en prédit 58 en E : une seule lettre d'erreur pour la MAE, mais une passoire manquée.

05 · CHOIX PROPOSÉ

Retenir le modèle lettre calculée avec le questionnaire

POURQUOI

Même erreur moyenne que la lettre prédite, **davantage de F/G retrouvés**.

Une fausse alerte se corrige pendant l'entretien. Une passoire manquée, c'est un habitant qui n'est pas orienté vers une rénovation prioritaire ni vers les aides associées.

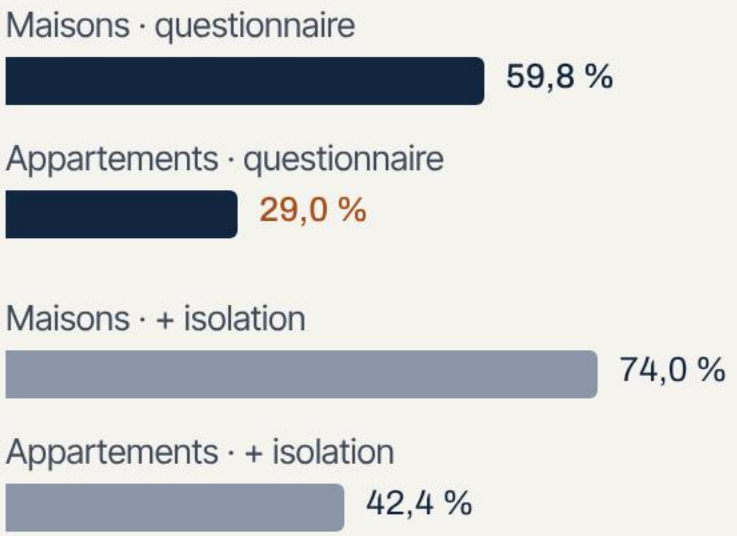
LE PRIX À PAYER

1 sur 2
alertes environ sont incorrectes

2 sur 3
F/G environ restent manqués

47 %
de lettres exactes seulement

F/G RETROUVÉS SELON LE LOGEMENT



L'isolation : en complément, quand elle est connue de manière fiable.
Les causes de l'écart maisons/appartements sont elles non établies.

3 F/G sur 4 sont des appartements, où le repérage est le plus faible (29 %) : c'est pour cela que le rappel global n'est que de 36 %.

Un outil indicatif : trois pistes prioritaires

LIMITES

Entrées simulées
Champs de DPE et pas de vraies réponses d'habitants : les résultats sont peut-être optimistes.

Périmètre restreint
Val-de-Marne, janvier à septembre 2026, logements individuels de 9 à 500 m².

Gain mesuré par blocs
3 → 14 → 18 infos : l'apport de chaque question n'est pas isolé.

Repérage partiel
Environ 2 F/G sur 3 manqués, surtout en appartement ; 1 alerte sur 2 fausse.

PISTES FUTURES

01
Comprendre les écarts
Pourquoi les F/G sont-ils moins bien détectés dans les appartements que dans les maisons ?

02
Simplifier le questionnaire
Tester le retrait d'informations et vérifier si les performances se maintiennent.

03
Tester des réponses réalistes
Mesurer l'effet des réponses inconnues ou approximatives d'un habitant.

Confirmer les résultats sur de nouvelles données et vérifier l'utilité pour le conseiller.

Merci.

Vos questions ?

0,65
lettre d'erreur moyenne avec le questionnaire (test)

36 %
des vrais F/G retrouvés par la lettre calculée

≈ 50 %
avec l'isolation, quand elle est connue de manière fiable

ANNEXE A · RÉGLAGES DES FORÊTS

Les réglages des deux types de forêts

01

Nombre d'arbres

Chaque arbre donne une réponse ; la forêt les combine.

Lettre prédite	Lettre calculée
200	**200**
1 forêt	par forêt · 2 forêts

02

Profondeur maximale

Nombre de questions successives dans un arbre.

Lettre prédite	Lettre calculée
20	**Libre**
questions au plus	bornée par les feuilles

03

Logements par feuille

Taille minimale du groupe final d'un arbre.

Lettre prédite	Lettre calculée
5	**10**
au minimum	au minimum

Bandeau et chiffres orange : réglage différent entre les deux approches.

Identiques pour les 3 paliers · issus d'essais antérieurs : des feuilles plus grandes baissaient l'erreur moyenne, mais aussi le rappel F/G.

ANNEXE B · SCORES COMPLETS

Validation et test, les six candidats

Modèle	Palier	MAE validation	MAE test	Rappel validation	Rappel test	Précision validation	Précision test	Accuracy test
Lettre prédite	3 · avant entretien	0,852	0,802	16,7 %	17,6 %	33,1 %	34,7 %	42,9 %
Lettre prédite	14 · questionnaire	0,680	0,656	18,6 %	20,3 %	66,6 %	65,3 %	47,7 %
Lettre prédite	18 · isolation	0,580	0,555	26,6 %	26,5 %	72,2 %	73,4 %	53,5 %
Lettre calculée	3 · avant entretien	0,775	0,734	11,9 %	13,4 %	37,2 %	42,0 %	43,4 %
Lettre calculée	14 · questionnaire	0,680	0,651	33,6 %	36,3 %	50,8 %	52,0 %	46,7 %
Lettre calculée	18 · isolation	0,579	0,553	48,5 %	49,8 %	57,6 %	56,4 %	52,6 %
Référence	toujours D	0,990*	0,970	—	—	—	—	29,6 %

Accuracy test = part des logements du test dont la lettre estimée est exactement la bonne. La référence prédit la même lettre pour tous : elle ne signale aucun F/G, d'où les « — ». Elle obtient tout de même 80 % des lettres à une lettre près.

Validation : validation croisée dans le train, 5 plis de groupes · test : 5 543 logements, 925 F/G · arrondis à 0,1 point · * en validation, référence = lettre la plus fréquente du train

ANNEXE C · BARÈME

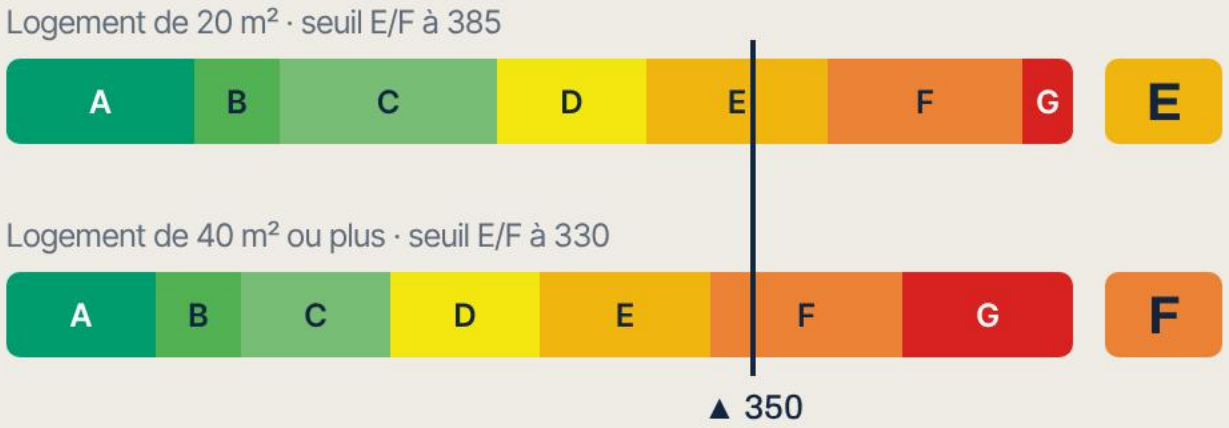
Un barème qui dépend aussi de la surface

01 Deux lettres : une pour la consommation d'énergie et l'autre pour les émissions. On garde **la moins favorable**.

02 Seuils variables jusqu'à 40 m², interpolés entre deux surfaces, fixes au-delà.

99,94 % d'accord avec les lettres enregistrées · 14 écarts sur 21 860

Exemple fictif : énergie 350 kWhEP/m², émissions 20 kgCO_2éq/m²



Émissions : C dans les deux cas. **Même consommation, lettre différente** : E à 20 m², F à 40 m².

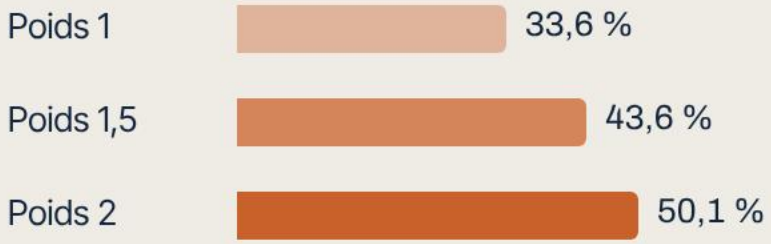
Échelle 0–500 kWhEP/m²/an · table de seuils du projet, de 8 à 40 m²

ANNEXE D · ESSAI EXPLORATOIRE · VALIDATION · LETTRE CALCULÉE

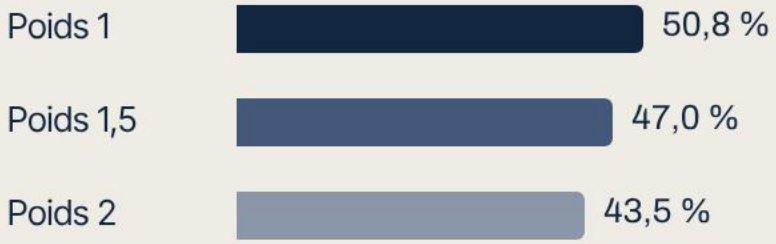
Faire compter davantage les F/G : à quel prix ?

Palier testé : **14 infos · + questionnaire**

RAPPEL · F/G RETROUVÉS (14 INFOS)



PRÉCISION · ALERTES JUSTES (14 INFOS)



14 infos, poids 1 → 1,5 → 2 · MAE 0,680 → 0,698 → 0,721 · erreurs ≥ 3 lettres 1,25 → 1,49 → 1,83 %

Plus d'information ≠ plus de poids. Le palier isolation (18 infos) sans pondération retrouve 48,5 % des F/G avec 57,6 % de précision (MAE 0,579) ; le palier questionnaire (14 infos) au poids 2 : 50,1 % mais 43,5 % (MAE 0,721).

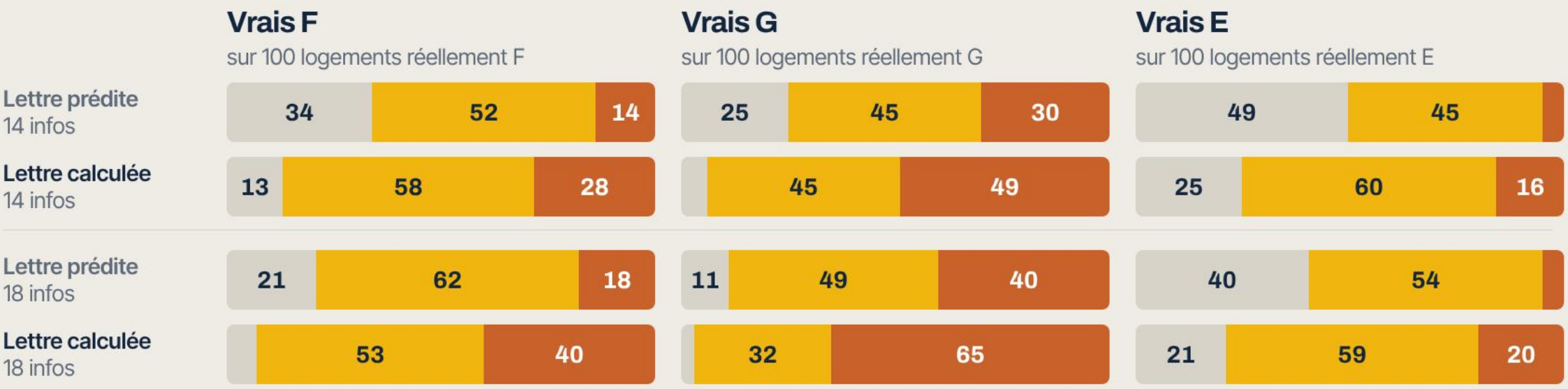
Même compromis aux autres paliers (poids 1 → 2) · **3 infos** : rappel 11,9 → 30,5 %, précision 37,2 → 33,7 % · **18 infos** : rappel 48,5 → 62,2 %, précision 57,6 → 49,6 %

Poids appliqués à l'apprentissage seulement · métriques non pondérées · pas de résultat sur le test

ANNEXE E · MATRICES DE CONFUSION RÉSUMÉES · TEST

Où vont les vrais F, G et E ?

prédits A à D · prédits E · signalés F/G



Le manque vient surtout de E : 52 à 62 % des vrais F sont prédits E, une seule lettre d'écart.

Lettre calculée + isolation : 40 vrais F et 65 vrais G sur 100 signalés.

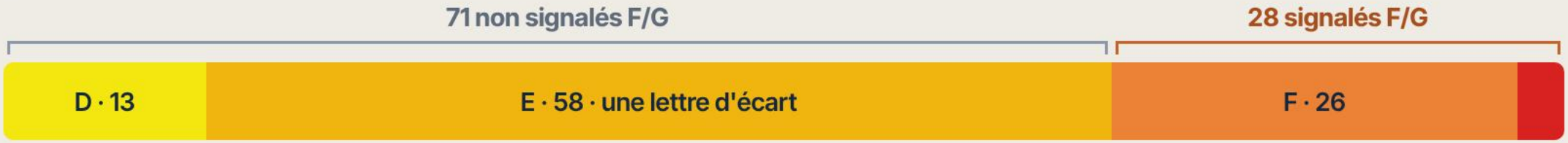
Contrepartie : jusqu'à 20 vrais E sur 100 signalés à tort F/G.

Lignes des matrices de confusion du test, regroupées · pourcentages arrondis (total ≈ 100)

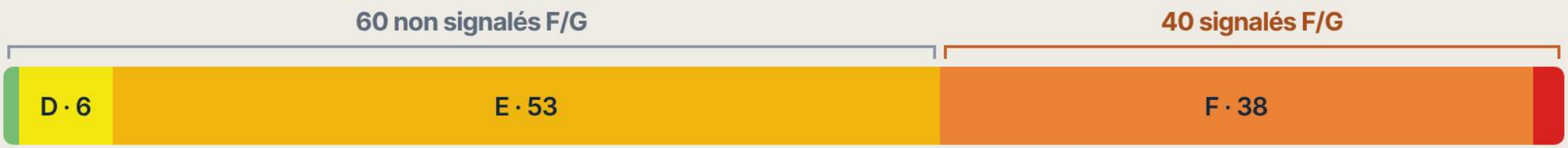
La majorité des vrais F manqués sont prédits E

Sur 100 logements réellement classés F, quelle lettre le modèle prédit-il ?

Questionnaire · 14 informations



+ Isolation · 18 informations



⚠ Dans le cas d'un vrai F prédit E : ce n'est qu'une **seule lettre d'erreur** pour la MAE, mais **une passoire F/G manquée** par le rappel.

Scores issus des lignes « vrai F » des matrices de confusion du test · les pourcentages sont arrondis (total ≈ 100)