JOSÉ LAMAS-VALVERDE

UN EXEMPLE SIMPLE POUR COMPRENDRE LE PRINCIPE DE L'APPRENTISSAGE AUTOMATIQUE

Affiner les données pour mieux développer les capacités des petites entreprises

Les solutions d'intelligence artificielle sont souvent perçues comme des produits finis, ce qui freine la créativité dans leur conception. Il est essentiel de mieux comprendre les phases de collecte de données, de création d'algorithmes et d'apprentissage automatique. Cet article se concentre sur ce dernier aspect par le biais d'un exemple simplifié.

1. INTRODUCTION

Les fondements mathématiques de l'apprentissage automatique *(machine learning)* reposent sur les statistiques et les méthodes d'optimisation mathématique. D'une manière générale, l'IA apprend à partir de données et d'instructions présentées sous la forme d'algorithmes.

Les données jouent un rôle stratégique crucial dans l'espace numérique. Leur analyse, ainsi que l'utilisation d'indicateurs clés et de l'apprentissage automatique, permettent aux PME de mieux valoriser leurs données et de gérer les risques de manière plus efficace.

1.1 Qu'est-ce qu'un algorithme d'apprentissage automatique? Il s'agit d'un ensemble de règles ou de processus utilisés par un système intelligent pour effectuer des tâches, le plus souvent pour découvrir de nouveaux éclairages et de nouveaux modèles, ou prédire des valeurs de sortie à partir d'un ensemble donné de variables d'entrée. Ce sont les algorithmes et les données qui permettent à l'apprentissage automatique de se développer.

L'apprentissage supervisé repose sur un ensemble de données étiquetées, où les variables (p. ex. salaire et âge) sont prédéfinies. Ces données comprennent des entrées et des sorties correctes, permettant au modèle d'apprendre progressivement. L'algorithme mesure sa précision à l'aide d'une fonction de perte, qu'il ajuste jusqu'à ce que l'erreur soit corrigée.

JOSÉ LAMAS-VALVERDE, PHD PHYSIQUE, EXPERT ANALYSE DES RISQUES, MÉTHODES ANALYSE QUANTITATIVE ET ANALYSE DÉCISIONNELLE, FONDATEUR-DIRECTEUR GESTRISK, JOSE.LAMAS@GESTIONDESRISQUES.CH

Contrairement à l'apprentissage supervisé, l'apprentissage non supervisé travaille avec des données non étiquetées. L'algorithme y crée des modèles qui aident à résoudre les problèmes de regroupement ou d'association. Cette méthode est particulièrement utile lorsque les experts ne connaissent pas précisément les propriétés communes d'un ensemble de données.

Un système d'apprentissage automatique exécute l'une des fonctions suivantes:

→ descriptive: le système utilise les données pour expliquer ce qu'il se passe;
→ prédictive: le système utilise les données pour prédire ce qu'il va se passer;
→ prescriptive: le système utilise les données pour suggérer ou recommander des actions à entreprendre.

L'exemple développé ci-dessous aborde la fonction prédictive. Il s'agit d'un exemple de modèle de régression simple utilisé pour prédire les salaires selon l'âge et correspondant à un apprentissage supervisé.

2. EXEMPLE: MODÉLISATION DU SALAIRE SELON L'ÂGE

Comme dans tout projet de modélisation, la démarche commence par une simplification de la réalité, la collecte de données et l'acceptation de quelques conditions de départ:

→ la première condition concerne le contexte: le salaire varie avec l'âge selon un modèle (une tendance générale du marché) à déterminer, dans un cadre professionnel et une région spécifique. Bien que plusieurs facteurs influencent cette relation, avec des pondérations diverses, ils n'entrent pas dans le cadre de cet exemple;
→ la deuxième condition porte sur la quantité de données collectées. Au lieu d'un échantillon important de 20 000 personnes, un échantillon de 20 personnes (n=20) sera utilisé, sans que cela affecte la validité de l'exemple. Le but est de construire un modèle de calcul adapté au contexte et ajusté

Tableau 1: **«TRAINING DATA SET»**

Âge	Salaire
25	47 500
27	72 500
30	62 500
35	105 000
40	151 000
45	131 500
50	133 000
55	140 000
60	118 000
65	134 500

aux données (appelé M1) permettant de prédire pour un âge donné (donnée d'entrée) le salaire le plus proche de la réalité (donnée de sortie).

→ Le modèle M1 est construit à l'aide d'un premier ensemble de données (n=10 personnes) appelé «ensemble de données d'entraînement» *(training data set)*. Le modèle sera ensuite soumis à validation avec un deuxième ensemble de données (n=10) nommé «ensemble de données de validation» *(validation data set)*. Le modèle sera affiné grâce à un troisième ensemble de données hors échantillon (n=10) appelé «ensemble de données de test» *(test data set)*. Il est important de noter que les deux premiers sous-ensembles de données font partie de l'échantillon global n=20; ils doivent être distincts et extraits de manière aléatoire. Ce dernier point est partiellement appliqué dans l'exemple et constitue la troisième condition de simplification.

2.1 Mesure de l'erreur de prévision du modèle. Les écarts de salaire seront calculés en comparant la différence entre les salaires observés dans les données d'entraînement et ceux prédits par le modèle. La même démarche sera appliquée pour les salaires dans l'ensemble de validation et l'ensemble de test. Ces écarts serviront à calculer l'erreur du modèle.

2.2 À la recherche du meilleur modèle. Plusieurs modèles seront construits, leurs erreurs mesurées et comparées. En principe, le modèle présentant des erreurs similaires pour l'ensemble d'entraînement et l'ensemble de validation sera préféré, garantissant ainsi sa stabilité. À l'aide de critères définis tels que le sur-ajustement et le sous-ajustement, le modèle qui répond le mieux aux données sera sélectionné. La méthode se déroule en trois étapes.

3. ÉTAPE 1: ÉLABORATION D'UN MODÈLE À PARTIR DES DONNÉES D'ENTRAÎNEMENT

Le *tableau 1* rassemble les données d'entraînement. La *figure 1* est la représentation graphique de l'ensemble de données d'entraînement du *tableau 1*.

Figure 1: **NUAGE DE POINTS DU «TRAINING DATA SET»**

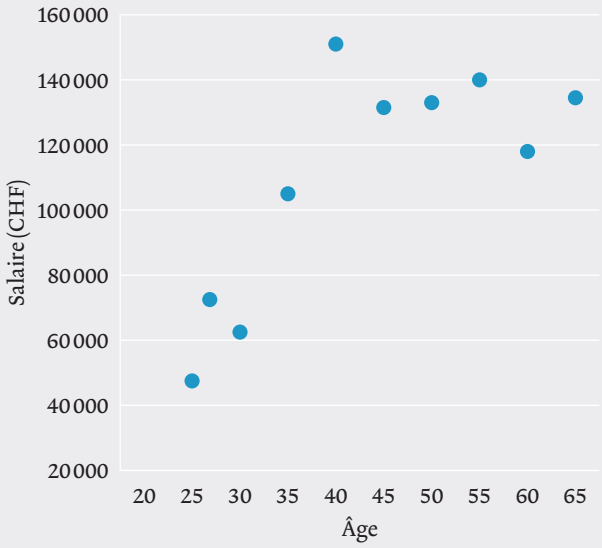
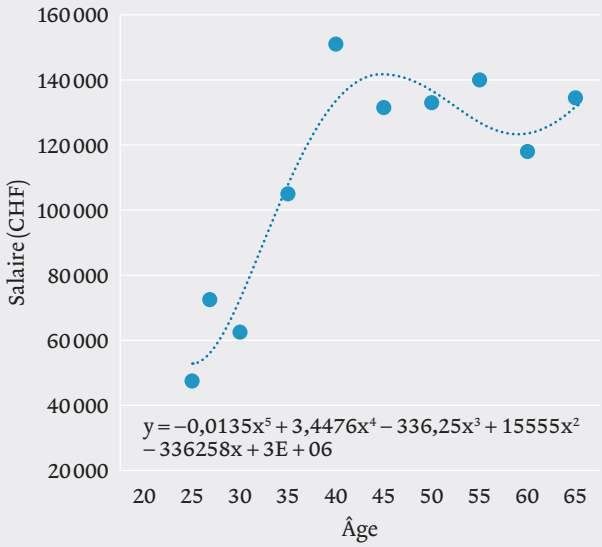


Figure 2: **NUAGE DE POINTS DU «TRAINING DATA SET» ET LA LIGNE D'AJUSTEMENT**



3.1 Conception du modèle. Le nuage de points suggère une courbe avec des collines et une sorte de plateau qui présente des similitudes avec la forme d'un polynôme [1].

Le polynôme de cinquième degré est notre modèle de départ (M1), où la variable *Y* représente le salaire et la variable *X* représente l'âge de la personne, tandis que les coefficients a, b1, b2, b3, b4 et b5 sont des paramètres d'ajustement déterminés dans le modèle. La recherche de ces paramètres d'ajustement de la courbe joue un rôle important dans l'apprentissage automatique des machines.

La *figure 2* montre le graphique de l'ensemble de données d'entraînement du *tableau 1* et la ligne d'ajustement polynomiale de degré 5 superposée. Le polynôme de 5e degré avec ses coefficients d'ajustement peut être obtenu par l'utilisation

Figure 3: **NUAGE DE POINTS DU «VALIDATION DATA SET»**

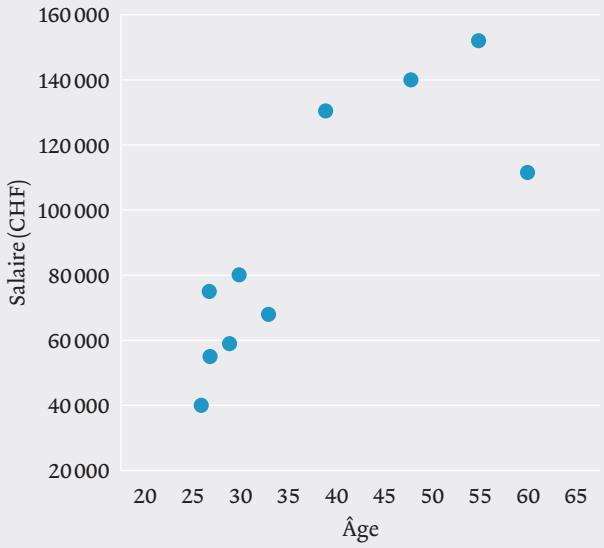
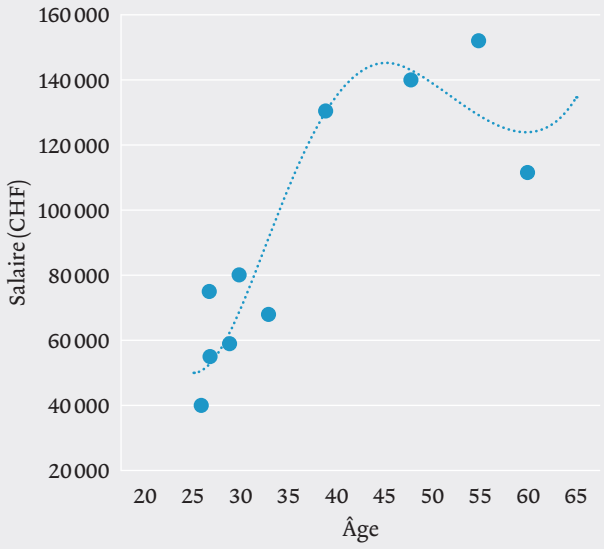


Figure 4: **NUAGE DE POINTS DU «VALIDATION DATA SET» ET LA LIGNE D'AJUSTEMENT**



d'un algorithme de calcul bien connu (méthode des moindres carrés) [2], incorporé dans toutes les calculatrices programmables et aussi dans MS Excel.

Le *tableau 2* montre les données de validation. La *figure 3* montre le graphique de l'ensemble de données de validation du *tableau 2*. Comme le montre la *figure 4*, lorsque le modèle M1, représenté par la ligne d'ajustement polynomiale de 5e degré, est superposé au graphique des données de validation, il ne fonctionne pas bien, notamment pour l'âge au-delà de 50 ans.

Il en ressort que le modèle polynomial de cinquième degré (M1) ne se généralise pas bien.

Tableau 2: **«VALIDATION DATA SET»**

Âge	Salaire
26	40 000
27	75 000
27	55 000
29	59 000
30	80 000
33	68 000
39	130 500
48	140 000
55	152 000
60	111 500

Tableau 3: **ÉCART-TYPE D'ERREUR POUR LE «TRAINING DATA SET»**

Âge	Salaire	Estimation	Erreur
35	105 000	109 003	−4 003
27	72 500	57 541	14 959
30	62 500	73 962	−11 462
25	47 500	53 420	−5 920
40	151 000	134 629	16 371
65	134 500	133 218	1 282
50	133 000	137 833	−4 833
55	140 000	128 200	11 800
60	118 000	124 580	−6 580
45	131 500	143 113	−11 613

3.2 Qualité de l'ajustement. Pour évaluer la qualité de l'ajustement d'un modèle particulier, une mesure objective et pratique est utilisée, fondée sur l'erreur (écart) entre le salaire réel et le salaire estimé par le modèle: le RMSE *(root mean squared error)* également appelé écart-type d'erreur. D'après le *tableau 3*, l'écart-type d'erreur pour l'ensemble des données d'entraînement est de CHF 10628,1 [3]. De même, le *tableau 4* indique que l'écart-type d'erreur pour l'ensemble des données de validation est de CHF 15170,04. Ces résultats ont été obtenus avec la formule classique de l'écart-type.

L'erreur RMSE pour l'ensemble des données de validation apparaît plus élevé que pour l'ensemble des données d'entraînement, contrairement à la stabilité attendue. Par conséquent, on peut conclure que le modèle ne se généralise pas bien aux nouvelles données, révélant un sur-ajustement (trop serré) aux données d'entraînement.

4. ÉTAPE 2: ÉLABORATION DE MODÈLES ALTERNATIFS

À partir des données d'entraînement, plusieurs modèles seront construits en utilisant des algorithmes polynomiaux de

Tableau 4: **ÉCART-TYPE POUR LE «VALIDATION DATA SET»**

Âge	Salaire	Estimation	Erreur
26	40 000	54 547	−14 547
27	75 000	57 541	17 459
27	55 000	57 541	−2 541
29	59 000	67 587	−8 587
30	80 000	73 962	6 038
33	68 000	95 194	−27 194
39	130 500	130 820	−320
48	140 000	141 071	−1 071
55	152 000	128 200	23 800
60	111 500	124 580	−13 080

premier, deuxième, troisième et quatrième degré, afin de les comparer à celui de cinquième degré. La *figure 5* présente le graphique des données ajustées selon les modèles susmentionnés. Les polynômes correspondants ne sont pas affichés ici, mais peuvent être obtenus avec le même logiciel.

4.1 Résultat des modèles. Le *tableau 5* ci-dessous montre le résumé des écart-type d'erreur (RMSE) pour les données d'entraînement et de validation.

Dans le cas du modèle linéaire (degré 1), les erreurs sont similaires mais trop élevées, c'est-à-dire que le modèle sous-ajuste *(under-fits)* les données, tandis que le polynôme du 5e degré sur-ajuste *(over-fits)*, autrement dit l'erreur de validation surpasse l'erreur d'entraînement. Entre les deux, le modèle de 2e degré présente des erreurs similaires et plus faibles. On peut dire qu'il se généralise mieux que le modèle plus

Tableau 5: **ERREURS DES MODÈLES**

Modèle: polynôme	Données d'entraînement	Données de validation
Degré	Erreur du modèle	Erreur du modèle
1	23 553.97	23 438.64
2	13 775.05	14 662.07
3	12 708.12	16 117.50
4	10 849.12	14 629.68
5	10 628.11	15 170.04

complexe de 5e degré. Toutefois, cela ne signifie pas que les modèles plus simples sont systématiquement meilleurs que les modèles plus complexes.

5. ÉTAPE 3: TESTER LE MODÈLE CHOISI

L'ensemble des données de test *(test data set)* présenté dans le *tableau 6* n'a pas été utilisé pour entraîner notre modèle ni pour valider les paramètres d'entrée. Il sert de test final, une fois le modèle définitif choisi, afin d'obtenir la meilleure estimation possible de l'efficacité du modèle sur des données entièrement nouvelles. Cet ensemble de données est également connu sous le nom de données hors échantillon.

À l'issue de l'étape précédente, le modèle de 2e degré avait été choisi. L'application de ce modèle aux données du *tableau 6* permet d'obtenir les estimations de salaire et les erreurs associées à chaque valeur d'âge, comme indiqué dans le *tableau 7*.

À partir du *tableau 7*, l'écart-type d'erreur du modèle de 2e degré est calculé pour l'ensemble du *test data set*; il est égal à CHF 27783.08

Figure 5: **ENSEMBLE DE DONNÉES D'ENTRAÎNEMENT VS SALAIRE ESTIMÉ (5 MODÈLES)**

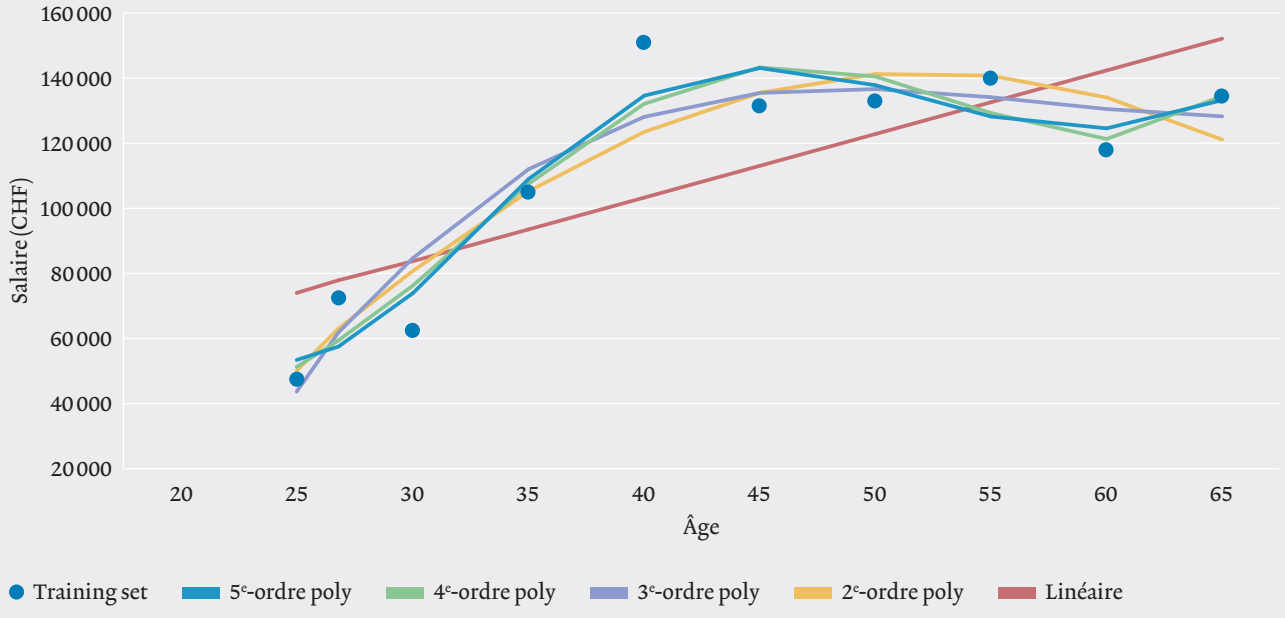


Tableau 6: **«TEST DATA SET»**

Âge	Salaire
26	38 500
52	190 000
38	185 000
60	145 000
64	130 000
41	110 000
34	100 500
46	125 000
57	153 000
55	150 000

Tableau 7: **ÉCART-TYPE D'ERREUR POUR LE «TEST DATA SET»**

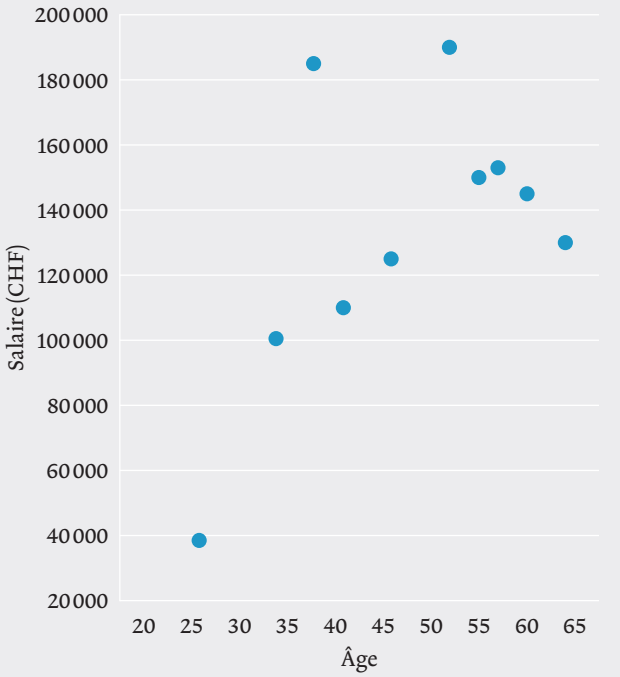
Âge	Salaire	Estimation	Erreur
26	38 500	56 729	–18 229
52	190 000	141 817	48 183
38	185 000	116 956	68 044
60	145 000	134 069	10 931
64	130 000	124 208	5 792
41	110 000	126 399	–16 399
34	100 500	100 872	–372
46	125 000	137 149	–12 149
57	153 000	138 846	14 154
55	150 000	140 782	9 218

En appliquant la même procédure aux cinq modèles, les résultats obtenus sont résumés dans le *tableau 8*. L'erreur de test des modèles de 2e et 3e degré semble similaire.

6. TROUVER L'ÉQUILIBRE ENTRE LE SUR-AJUSTEMENT ET LE SOUS-AJUSTEMENT

Il s'agit d'un problème important en *machine learning*. Certains algorithmes d'apprentissage automatique, comme les réseaux de neurones, peuvent impliquer un très grand nombre de paramètres. Il est alors facile de sur-ajuster, même lorsque l'ensemble des données d'entraînement est volumineux.

Figure 6: **NUAGE DE POINTS DU «TEST DATA SET»**



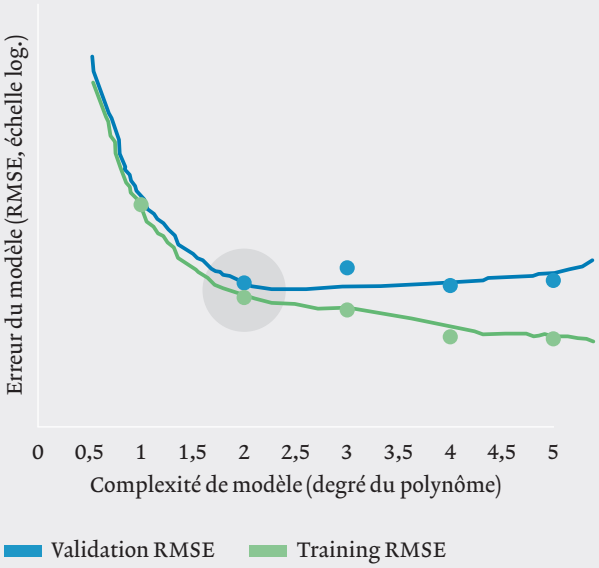
Pour chaque modèle, l'erreur (RMSE) a été mesurée pour les ensembles d'entraînement et de validation. Lorsque la complexité du modèle (soit le degré du polynôme dans notre exemple) est inférieure à X, le modèle se généralise bien, c'est à dire que l'erreur du modèle pour l'ensemble de validation n'est que légèrement supérieure à celle de l'ensemble d'apprentissage. À mesure que la complexité du modèle dépasse X, les erreurs de l'ensemble de validation commencent à augmenter. Le meilleur modèle est donc celui avec une complexité égale à X. Dans notre exemple, X est égal à 2. Une règle empirique pourrait être formulée ainsi:

«la complexité du modèle devrait être augmentée jusqu'à ce que les tests hors échantillon indiquent qu'il ne généralise pas correctement».

Tableau 8: **ERREURS DE TEST**

Modèle: polynôme	Données d'entraînement	Données de validation	Données de test
Degré	Erreur du modèle	Erreur du modèle	Erreur du modèle
1	23 553.97	23 438.64	35 949.60
2	13 775.05	14 662.07	**27 783.08**
3	12 708.12	16 117.50	**28 131.63**
4	10 849.12	14 629.68	29 968.55
5	10 628.11	15 170.04	4 741 969.28

Figure 7: **ÉQUILIBE ENTRE SUR-AJUSTEMENT ET SOUS-AJUSTEMENT**



Cette règle d'équilibre entre l'ajustement excessif et insuffisant est illustrée dans la *figure 7* (les courbes sont tracées pour guider l'œil), où l'on voit que le degré du polynôme idéal se situerait autour de 2, avec des valeurs telles que 2,1 ou 2,2, ce qui resterait à vérifier par des calculs plus poussés.

7. CONCLUSION

La méthode d'apprentissage automatique a été présentée par le biais d'un exemple simplifié adapté à la formation continue. En pratique, l'optimisation des modèles fait appel à des techniques sophistiquées qui dépassent les capacités d'Excel. Des bibliothèques d'algorithmes sont disponibles dans des écosystèmes développés dans des langages de haut niveau, tels que Python [4]. ■

Notes: 1) Un bon ajustement (good fit) pourrait être un polynôme avec la formule suivante: $Y = a + b_1X + b_2X^2 + b_3X^3 + b_4X^4 + b_5X^5$. **2)** La formule est: $Y = 3\times10^6 - (336\,258)X + (15\,555)X^2 - (336{,}25)X^3 + (3{,}4476)X^4 - (0{,}0135)X^5$. **3) P**our l'obtenir, appliquez la fonction STDEV.S à la colonne des erreurs à droite du tableau 3. **4)**La méthodologie de l'apprentissage automatique, en version Excel et Python, est abordée par l'auteur dans ses séminaires dans le cadre de la formation continue d'Expertsuisse.