Modèle de prédiction des litiges pour les brevets

Rapport de projet

Réalisé par :

Romain Tancrez
Adrien Morlet
Zakaria Ettouhami
Zyad Cherkaoui
Kiane Lachkar

[ETU33]

Encadrants :

M. François Maublanc
M. Naïm Zoghlami

Table des matières

1 Introduction **3**
- 1.1 Contexte du projet . . . 3
- 1.2 Objectifs . . . 3
- 1.3 Données et sources . . . 4
- 1.4 Enjeux . . . 4
- 1.5 Considérations éthiques . . . 4

2 Identification du besoin et état de l'art **5**
- 2.1 Définition des litiges de brevets . . . 5
- 2.2 Travaux existants . . . 5
- 2.3 Régression logistique . . . 6
- 2.4 Elastic Net . . . 6
- 2.5 Random Forest . . . 6
- 2.6 XGBoost . . . 7
- 2.7 Techniques de rééquilibrage : SMOTE . . . 7
- 2.8 Évaluation par AUC et courbe ROC . . . 7

3 Analyse du besoin client **8**
- 3.1 Spécifications fonctionnelles . . . 8
- 3.2 Contraintes . . . 8
- 3.3 Considérations économiques . . . 8

4 Comparaison des modèles **12**
- 4.1 Méthodologie . . . 12
- 4.2 Régression logistique . . . 15
- 4.3 Modèle Elastic Net . . . 16
- 4.4 Random Forest . . . 17
- 4.5 XGBoost . . . 19
- 4.6 Réseau de Neurones . . . 20
- 4.7 Synthèse comparative . . . 23

5 Résultats et interface **24**
- 5.1 Interprétation des performances et du ratio TP/FP . . . 24
- 5.2 Interface utilisateur . . . 24

6 Conclusion et perspectives **27**
- 6.1 Bilan du projet . . . 27
- 6.2 Limites et améliorations possibles . . . 27
- 6.3 Déploiement potentiel . . . 28

Bibliographie **28**

Chapitre 1

Introduction

1.1 Contexte du projet

Dans un contexte d'innovation rapide, notamment dans les secteurs des technologies de l'information et de la communication et des biotechnologies, les inventions sont de plus en plus souvent « cumulatives ». Cela signifie qu'elles s'appuient sur des innovations antérieures, créant ainsi une forte interdépendance entre les brevets. Ce phénomène engendre des risques juridiques accrus pour les entreprises, notamment celui de violer involontairement un brevet existant, ou de se retrouver dans une position de dépendance technologique vis-à-vis d'un brevet fondamental, favorisant des comportements de type *hold-up*.

Les litiges liés aux brevets sont coûteux et complexes, avec un impact financier et stratégique important, en particulier pour les PME qui ne disposent pas toujours des ressources juridiques suffisantes. Aux États-Unis, environ 1% des brevets délivrés entre 1998 et 2017 ont fait l'objet d'un contentieux, avec des dommages-intérêts médians atteignant 6 millions de dollars par affaire. Dans ce contexte, il devient essentiel de pouvoir anticiper les risques juridiques liés aux brevets dès leur dépôt.

En France, une étude relayée par l'IEEPI [3] rappelle que les litiges de brevets, bien que rares en proportion, sont fréquents dans certains secteurs comme la pharmacie ou les télécommunications. Le coût moyen d'un contentieux est estimé entre 80 000 et 150 000 euros, avec des pointes allant jusqu'à plusieurs millions dans les cas complexes. La durée des procédures peut s'étendre sur plusieurs années, bien qu'une réforme récente ait permis d'accélérer le traitement des affaires. Ces données confirment l'intérêt d'un outil d'anticipation précoce du risque contentieux.

1.2 Objectifs

Ce projet vise à développer un modèle de machine learning capable de prédire si un brevet donné a des chances d'être impliqué dans un litige. Plus précisément, il s'agit d'exploiter une base de données de 600 000 brevets délivrés aux États-Unis entre 2002 et 2005, et d'identifier les caractéristiques pertinentes des brevets permettant de prédire le risque de contentieux. Ce modèle devra aider les entreprises à mieux anticiper les litiges potentiels et à ajuster leurs stratégies de propriété intellectuelle.

Les objectifs sont les suivants :
— Concevoir un pipeline de traitement des données adapté à un jeu de données déséquilibré et multivarié ;
— Comparer plusieurs algorithmes d'apprentissage supervisé (régression logistique, XGBoost, réseaux de neurones, etc.) ;
— Identifier les variables les plus influentes dans la prédiction de litiges ;
— Proposer un modèle robuste, interprétable, et avec un bon compromis entre rappel et précision ;
— Fournir une interface ou un outil d'interprétation facilitant l'usage du modèle pour les non-experts.

1.3 Données et sources

La base de données utilisée pour ce projet contient des informations issues de brevets américains accordés entre 2002 et 2005. Chaque ligne représente un brevet, accompagné de nombreuses variables descriptives : informations administratives (date de dépôt, année de délivrance), données techniques (nombre de revendications, secteur technologique, citations antérieures et postérieures), et indicateurs qualitatifs (qualité du brevet, généralité, statut universitaire, etc.).

1.4 Enjeux

Les enjeux de ce projet sont à la fois théoriques, pratiques et économiques. D'un point de vue académique, il s'agit d'appliquer des méthodes de Machine Learning à un problème socio-économique complexe. D'un point de vue pratique, le modèle développé pourrait permettre aux entreprises de mieux évaluer leurs risques juridiques dès la phase de dépôt de brevet, en facilitant les audits de portefeuille et la stratégie d'innovation.

Enfin, sur le plan économique, la prédiction de litiges peut permettre des économies substantielles, éviter des procès coûteux, et soutenir la compétitivité des PME dans un environnement technologique de plus en plus concurrentiel.

1.5 Considérations éthiques

Il est important de noter que ce type de modèle, bien qu'utile pour anticiper les litiges, ne doit pas être interprété comme une preuve juridique. Il sert uniquement d'outil d'aide à la décision, et nécessite une analyse complémentaire par des experts en propriété intellectuelle.

La mise à disposition de ce type d'outil doit aussi être réfléchie en pesant le risque de son utilisation potentielle à des fins agressives, par exemple pour détecter les brevets les plus vulnérables sur lesquels attaquer des entreprises qui n'auraient pas les moyens juridiques de se défendre.

Chapitre 2

Identification du besoin et état de l'art

2.1 Définition des litiges de brevets

Un litige de brevet désigne un conflit juridique concernant les droits exclusifs accordés par un brevet, généralement en cas d'accusation de contrefaçon ou de demande d'invalidation. Ces litiges relèvent du droit de la propriété industrielle, mais ont des impacts bien au-delà du juridique : ils peuvent influencer des décisions commerciales, stratégiques et technologiques majeures.

Bien que relativement rares (moins de 2 % des brevets déposés), les brevets impliqués dans des litiges sont souvent associés à des technologies à fort potentiel économique ou stratégique. Anticiper de tels litiges est donc un enjeu clé pour les entreprises innovantes, notamment dans des secteurs comme la pharmacie, les télécoms ou les biotechnologies, où l'innovation est cumulative.

Cependant, cette tâche est rendue complexe par la diversité des facteurs impliqués (techniques, juridiques, économiques) et la faible occurrence de litiges. D'où l'intérêt d'une approche algorithmique capable de prédire, à partir des métadonnées des brevets, la probabilité d'un futur litige.

À l'échelle européenne, la France se distingue par un nombre particulièrement élevé de décisions de justice en matière de brevets. D'après une étude comparative citée dans le rapport de l'IEEPI [3], la France a enregistré 380 décisions entre 2000 et 2019, pour un total de plus de 113 millions d'euros d'indemnisation accordée — bien au-dessus de l'Allemagne, l'Espagne ou le Royaume-Uni. Ces chiffres soulignent l'importance stratégique du contentieux brevet en France et la nécessité d'une approche préventive par la donnée.

2.2 Travaux existants

Plusieurs travaux récents ont cherché à anticiper les litiges de brevets en exploitant les métadonnées associées aux documents brevets. Les approches varient selon les hypothèses, les sources d'information utilisées et les modèles algorithmiques mobilisés.

Juranek et Otneim (2024) proposent un modèle d'apprentissage supervisé basé sur des variables numériques et catégorielles disponibles dès le dépôt du brevet : nombre de revendications, citations antérieures, taille de la famille, etc. Leur objectif est d'identifier les caractéristiques intrinsèques aux brevets qui corrèlent fortement avec la probabilité de litige. Ils insistent sur l'importance de l'équilibre entre précision et rappel, et appliquent plusieurs méthodes classiques comme la régression logistique, les arbres de décision et

XGBoost. Leurs résultats montrent que le nombre de citations antérieures, la qualité du brevet ou encore son originalité sont de bons prédicteurs de contentieux [1].

Wongchaisuwat et al. (2024), quant à eux, adoptent une approche plus hybride. Leur modèle combine des variables tabulaires (issues des métadonnées juridiques et techniques) avec des représentations vectorielles extraites du texte des brevets (résumés, revendications, etc.). Ils appliquent des modèles non linéaires pour estimer à la fois la probabilité de litige et le temps jusqu'au litige. Ils soulignent notamment le rôle crucial des réseaux de citations dans la diffusion d'une valeur juridique, ainsi que l'importance d'acteurs stratégiques comme les universités ou les non-practicing entities (NPE) [2].

Ces deux approches se rejoignent sur plusieurs points :

— Elles s'appuient sur de grands jeux de données structurées issus des bases de l'USPTO ;
— Elles considèrent le litige comme un événement rare, donc nécessitant une attention particulière au déséquilibre des classes ;
— Elles privilégient des modèles interprétables ou explicables, dans un objectif d'aide à la décision stratégique.

2.3 Régression logistique

La régression logistique est souvent utilisée comme modèle de base pour la classification binaire. Elle permet de modéliser la probabilité qu'un brevet soit litigieux à partir d'une fonction logistique appliquée à une combinaison linéaire des variables explicatives [5]. Cependant, ce modèle suppose un équilibre des classes, ce qui est rarement le cas ici, puisque la proportion de brevets impliqués dans un litige est très faible (environ 1,1 %) [1]. Cela peut conduire à une mauvaise calibration des probabilités et à un grand nombre de faux négatifs si l'on n'adapte pas le seuil de décision ou le jeu de données.

2.4 Elastic Net

Elastic Net est un modèle régularisé combinant les approches Ridge (L2) et Lasso (L1), ce qui permet à la fois de limiter la complexité du modèle et de sélectionner automatiquement les variables les plus pertinentes [6]. Il est particulièrement utile lorsque les variables sont nombreuses et potentiellement corrélées. Ce modèle est testé dans l'étude de Juranek et Otneim [1], avec des performances équivalentes à la régression logistique, mais un meilleur contrôle du surapprentissage grâce à la pénalisation.

2.5 Random Forest

La méthode Random Forest, abordée dans le cadre du cours de data mining [5], repose sur l'agrégation de plusieurs arbres de décision entraînés sur des sous-échantillons bootstrap du jeu de données. Chaque arbre partitionne l'espace des variables pour minimiser une mesure d'impureté (comme l'indice de Gini), et la prédiction finale est obtenue par vote majoritaire. Dans l'étude de Juranek et Otneim [1], Random Forest obtient un score AUC de 0.815, ce qui le place comme un modèle performant, robuste au déséquilibre et peu sensible au bruit.

2.6 XGBoost

XGBoost (Extreme Gradient Boosting) est une amélioration des méthodes de boosting classiques, où les arbres sont construits séquentiellement et chaque nouvel arbre corrige les erreurs du précédent. Une régularisation est également intégrée à la fonction de coût afin d'éviter le surapprentissage [1]. Ce modèle est le plus performant dans les expérimentations menées par les auteurs, avec un AUC de 0.818, ce qui le rend particulièrement adapté à la prédiction de litiges sur des données déséquilibrées.

2.7 Techniques de rééquilibrage : SMOTE

Le problème du déséquilibre des classes (brevets litigieux très minoritaires) est central dans les travaux sur la prédiction de litiges. Pour y remédier, Juranek et Otneim [1] utilisent la méthode SMOTE (Synthetic Minority Over-sampling Technique), qui génère artificiellement de nouveaux exemples de la classe minoritaire en interpolant des observations existantes. Cette technique améliore la sensibilité des modèles sans altérer excessivement la spécificité, en particulier pour les modèles de type arbre (Random Forest, XGBoost).

2.8 Évaluation par AUC et courbe ROC

L'évaluation des performances est effectuée via la courbe ROC, qui trace le taux de vrais positifs (TPR) contre le taux de faux positifs (FPR) pour différents seuils. L'aire sous cette courbe (AUC) est une mesure synthétique de la capacité du modèle à discriminer les deux classes [1]. Un AUC de 0.5 correspond à un modèle aléatoire, tandis qu'un AUC de 1 désigne un classifier parfait. Les modèles testés par les auteurs obtiennent des scores AUC entre 0.793 (régression logistique) et 0.818 (XGBoost), confirmant l'avantage des méthodes arborescentes.

Conclusion

Les travaux existants montrent que les modèles arborescents, en particulier XGBoost et Random Forest, offrent les meilleures performances dans un contexte de forte asymétrie des classes. Cependant, ils restent complexes à interpréter, ce qui pose un défi dans les contextes juridiques. La régularisation (Elastic Net) et le rééquilibrage (SMOTE) apparaissent également comme des leviers clés pour améliorer la robustesse et la pertinence des prédictions.

Chapitre 3

Analyse du besoin client

3.1 Spécifications fonctionnelles

L'objectif principal de ce projet est de concevoir un modèle de machine learning capable d'anticiper les litiges potentiels associés aux brevets. Il s'agit donc de classifier automatiquement un brevet comme étant *susceptible de litige* ou non, à partir de données historiques disponibles au moment du dépôt.
Les spécifications fonctionnelles attendues sont les suivantes :

- Le modèle doit accepter en entrée les données descriptives d'un brevet (*informations temporelles, citations, indices de qualité, origine géographique, etc.*).
- Il doit produire en sortie une probabilité d'avoir un litige ou non.
- L'approche doit permettre une interprétabilité suffisante pour identifier les facteurs influents.
- Le modèle doit minimiser le coût opérationnel des erreurs de classification, notamment les faux positifs.

3.2 Contraintes

La modélisation est soumise à plusieurs contraintes :

- Déséquilibre fort des classes : moins de 2% des brevets sont associés à un litige, rendant difficile l'apprentissage sans techniques spécifiques (sur-échantillonnage, pondération).
- Coût asymétrique des erreurs : une erreur de type faux positif (FP) est plus pénalisante qu'un faux négatif (FN), car elle peut mener à l'abandon prématuré d'un brevet.
- Complexité du contexte juridique : le modèle ne doit pas reposer sur des informations indisponibles à la date de dépôt (pas de fuite de données).
- Exigence de performance opérationnelle : les prédictions doivent être suffisamment fiables pour servir d'outil d'aide à la décision.

3.3 Considérations économiques

Dans un cadre industriel, les performances d'un modèle ne se mesurent pas uniquement en termes statistiques. L'enjeu est clair : générer un gain économique net tangible. Cela implique de lire les résultats à travers une grille d'analyse orientée métier.

Impact économique des prédictions

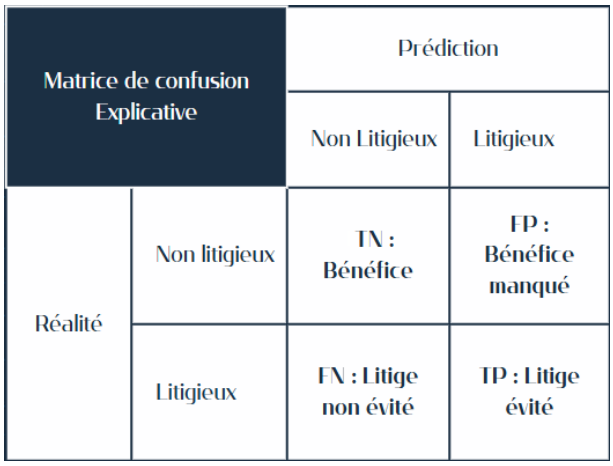

FIGURE 3.1 – Matrice de Confusion Explicative

Chaque décision du modèle influence directement la rentabilité de l'entreprise :

— **True Positive (TP)** : un brevet litigieux est correctement identifié. L'entreprise anticipe ainsi un contentieux potentiel et évite des frais juridiques significatifs (*gain* $= TP \times C_J$). Avec

— **False Positive (FP)** : un brevet non litigieux est signalé à tort. Il risque d'être mis de côté, entraînant une perte d'opportunité commerciale. (*perte* $= FP \times C_B$)

Avec :

— C_J : coût moyen d'un litige (6 000 000 $ dans notre cas)

— C_B : gain moyen associé à un brevet non litigieux conservé (1 000 000 $)

Formule de calcul du bénéfice net

Sans utilisation du modèle, le gain lié à l'exploitation des brevets est :

$$(FP + TN) \times C_B - (TP + FN) \times C_J \tag{3.1}$$

Ceci correspond aux gains des brevets non litigieux moins les pertes des brevets litigieux.

L'utilisation du modèle va conduire à renoncer aux brevets prédits litigieux et à garder les brevets prédits non-litigieux. Le gain devient alors :

$$TN \times C_B - FN \times C_J \tag{3.2}$$

En soustrayant 3.1 de 3.2, nous obtenons l'économie E réalisée en utilisant le modèle :

$$E = (TP \times C_J) - (FP \times C_B) \tag{3.3}$$

L'économie relative ainsi réalisée correspond au ratio :

$$\frac{E}{(FN + TP) \times C_J} = r \cdot \left(1 - \frac{1-p}{p} \cdot \gamma\right) \tag{3.4}$$

Qui s'exprime directement en fonction des résultats de la matrice de confusion : la précision $p = \frac{TP}{TP+FP}$, le rappel $r = \frac{TP}{TP+FN}$ et d'une constante $\gamma = \frac{C_B}{C_J}$.

Exemple d'impact financier concret

Considérons une application dans le domaine des dispositifs médicaux, où chaque brevet protège un composant critique.

— **60 brevets litigieux** sont correctement identifiés : **360 millions de $** de litiges évités
— **40 brevets sains** sont écartés à tort : **40 millions de $** d'opportunités commerciales perdues
— **Bénéfice net total** : **320 millions de $** économisés par l'utilisation du modèle.

Cet exemple montre l'impact économique direct d'un bon calibrage du modèle.

Pourquoi le ratio `TP/FP` est stratégique

Le ratio TP/FP [1] ne mesure pas seulement une performance statistique : il mesure le seuil à partir duquel le modèle devient rentable :

— si $TP/FP > \gamma$ l'économie réalisée en utilisant le modèle est positive. Cela signifie que les alertes émises sont en grande majorité pertinentes — chaque alerte est susceptible d'économiser plus qu'elle ne coûte.
— si $TP/FP < \gamma$ l'économie réalisée en utilisant le modèle est négative. Cela signifie que le modèle prédit trop de faux positifs qui diluent la valeur des vraies alertes. Dans ce cas le manque à gagner du aux brevets abandonnés à tort est plus important que l'économie réalisée en évitant les litiges.

Il est important de noter que la présence de faux positifs n'est pas problématique en soi. Elle devient acceptable tant que le ratio $\eta = \frac{\gamma}{TP/FP} = \gamma \cdot \frac{1-p}{p}$ reste économiquement favorable. Ce ratio apparaît dans l'Eq. 3.4 que l'on peut réécrire comme :

$$\%E = r \cdot \left(1 - \gamma \cdot \frac{1-p}{p}\right) = r \cdot (1 - \eta) \tag{3.5}$$

L'objectif n'est pas de tendre vers zéro erreur, mais vers un optimum économique en maximisant $r \cdot (1 - \eta)$. On observe aussi que dans l'Eq. 3.5 η joue un rôle capital : non seulement sa valeur va influer sur le % d'économie réalisée en utilisant le modèle mais elle aussi conditionner le signe donc le gain ou la perte liés à l'utilisation du modèle :

— si $\eta < 1$ (ce qui correspond à $TP/FP > \gamma$ une économie est réalisée en utilisant le modèle.
— si $\eta > 1$ (ce qui correspond à $TP/FP < \gamma$) l'utilisation du modèle va conduire à une perte pour l'utilisateur.
— η représente aussi la marge de robustesse par rapport aux hypothèses considérées pour estimer les valeurs C_J et C_B. Pour une même économie relative $\%E$, un modèle avec une valeur de η plus faible sera à privilégier car il sera plus robuste à des évolutions ou imprécisions dans l'estimation de C_J et C_B.

1. TP/FP est directement lié à la précision du modèle avec $TP/FP = \frac{p}{1-p}$

Modèle vs analyse manuelle

En l'absence de modèle, l'entreprise serait contrainte d'évaluer tous les brevets manuellement ou d'appliquer une approche uniforme, sans priorisation du risque. Le modèle agit ici comme un filtre intelligent : il hiérarchise les brevets en fonction de leur potentiel de litige, permettant une allocation plus efficace des ressources juridiques et commerciales.

Le véritable enjeu du projet n'est pas seulement de classifier correctement, mais de transformer chaque bonne prédiction en un gain financier. En ce sens, les ratios `TP/FP` (ou p, ou η) & $\%E$ deviennent des indicateurs stratégiques, alignant les performances algorithmiques sur les objectifs économiques réels de l'entreprise.

Chapitre 4

Comparaison des modèles

4.1 Méthodologie

Cette section décrit de manière rigoureuse la démarche suivie pour la construction, l'entraînement et la comparaison de plusieurs modèles de prédiction des litiges liés aux brevets. L'ensemble des étapes s'inscrit dans un pipeline homogène, allant du prétraitement à l'évaluation finale, avec un accent particulier sur les métriques pertinentes dans un contexte industriel où les erreurs ont un coût asymétrique.

Structure du pipeline

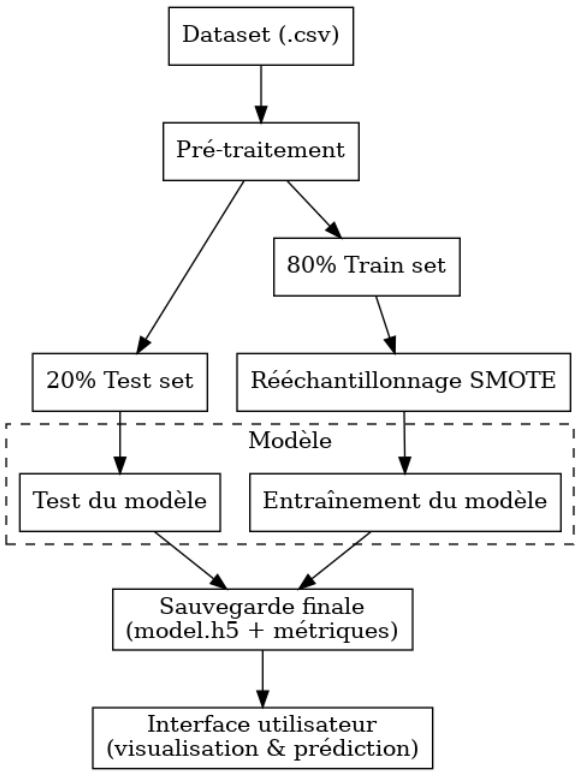

FIGURE 4.1 – Schéma du Pipeline du Projet

Tous les modèles ont été entraînés selon un même schéma de traitement, assurant ainsi la cohérence de la comparaison. Le pipeline est structuré en huit étapes principales :

1. **Préparation des données** : nettoyage des colonnes, imputation des valeurs manquantes numériques par la médiane, et encodage des variables catégorielles.
2. **Séparation des jeux** : découpage stratifié en ensemble d'entraînement (80 %) et de test (20 %), en respectant la proportion de brevets litigieux.
3. **Rééquilibrage des classes (SMOTE)** : application exclusive sur le jeu d'entraînement. La méthode SMOTE (Synthetic Minority Over-sampling Technique) génère des exemples synthétiques de la classe minoritaire à partir de ses voisins les plus proches. Un ratio cible de 30 % a été fixé pour limiter le déséquilibre tout en évitant la surreprésentation artificielle.
4. **Standardisation** : les variables numériques ont été centrées et réduites (z-score). Cette étape est indispensable pour les modèles sensibles à l'échelle des variables (régression, Elastic Net, MLP [1]). Elle a néanmoins été appliquée systématiquement à l'ensemble des modèles pour garantir l'uniformité du traitement.
5. **Entraînement des modèles** : chaque modèle a été entraîné sur les données rééquilibrées et normalisées, avec des ajustements propres à sa nature (régularisation, pondération des classes, architecture).
6. **Sélection de variables** : une réduction supervisée de la dimensionnalité a été réalisée, selon une stratégie adaptée à chaque famille de modèle.
7. **Optimisation du seuil de décision** : le seuil de classification a été ajusté via une validation croisée interne, en maximisant le F1-score et la précision p.
8. **Évaluation finale** : tous les modèles sont évalués sur le même jeu de test, non modifié, selon des critères identiques.

Modèles évalués

Cinq modèles, issus de familles algorithmiques complémentaires, ont été sélectionnés :
— **Régression logistique** : modèle de base, rapide à entraîner et interprétable.
— **Elastic Net** : version régularisée combinant pénalités L1 et L2, adaptée aux contextes de forte colinéarité.
— **Random Forest** : modèle d'ensemble basé sur des arbres décisionnels, robuste aux données bruitées et aux non-linéarités.
— **XGBoost** : algorithme de boosting par gradient, performant sur données tabulaires déséquilibrées.
— **Perceptron multicouche (MLP)** : réseau de neurones dense, capable de modéliser des interactions complexes.

Encodage des variables catégorielles

Les variables qualitatives ont été encodées par fréquence d'apparition, attribuant à chaque modalité sa proportion d'occurrence dans le jeu d'entraînement. Ce type d'encodage préserve l'information statistique tout en limitant la dimensionnalité, contrairement à un encodage one-hot.

1. MLP : "Multi Layer Perceptron"

Rééquilibrage des classes

La très faible proportion de litiges (environ 1,2 %) rend indispensable une correction du déséquilibre. Celle-ci a été appliquée uniquement sur le jeu d'entraînement via la méthode SMOTE, avec un ratio cible fixé à 30 % de classe minoritaire. Cela permet aux modèles de mieux apprendre les motifs associés aux cas litigieux sans altérer la représentativité du jeu de test.

Standardisation des variables

Une standardisation des variables numériques (z-score) a été réalisée systématiquement. Elle est indispensable pour les modèles utilisant des gradients ou des pénalités (régression logistique, Elastic Net, MLP). Bien qu'elle soit inutile pour les modèles arborescents, elle a été conservée pour garantir un pipeline unifié.

Sélection de variables

La sélection de variables a été adaptée selon le type de modèle :

- **Modèles linéaires (logistique, Elastic Net)** : utilisation de la méthode `SelectFromModel`, en retenant :
  - les coefficients au-dessus de la moyenne (régression logistique),
  - les coefficients dépassant 1.5 fois la moyenne absolue (Elastic Net).

  Cette approche vise à améliorer l'interprétabilité et à limiter le sur-apprentissage.
- **XGBoost** : aucune sélection préalable. Le modèle intègre naturellement une pondération interne des variables. Une analyse post hoc de l'importance des variables a néanmoins été conduite.
- **Réseau de neurones (MLP)** : le MLP a été entraîné sur le même jeu de données que celui utilisé pour le modèle XGBoost, incluant toutes les variables disponibles après nettoyage. Aucune réduction stricte de dimension n'a été appliquée en amont, mais la régularisation du modèle a été assurée par des couches `Dropout` et un mécanisme d'arrêt anticipé (`EarlyStopping`).

Optimisation du seuil de classification

Le seuil de classification n'a pas été fixé arbitrairement à 0.5. Il a été optimisé pour chaque modèle à l'aide d'une validation croisée interne, selon deux critères p & r

Critères d'évaluation

Tous les modèles ont été évalués sur un même jeu de test réel, à partir des métriques suivantes :

- **Accuracy** : $a = \frac{TP+TN}{TP+TN+FP+FN}$ proportion globale de bonnes classifications,
- **Précision** : $p = \frac{TP}{TP+FP}$ part des litiges prédits qui sont réellement litigieux,
- **Rappel** : $r = \frac{TP}{TP+FN}$ part des litiges réels effectivement détectés,
- **F1-score** : $F1 = \frac{2 \cdot p \cdot r}{p+r}$ moyenne harmonique entre précision et rappel,
- **R-Eco** : $\%E = r \cdot (1 - \eta)$ indicateur économique clé du projet.
- **Robustness** : $\eta = \frac{1-p}{p} \cdot \gamma$indicateur de la marge de robustesse.

Cette approche méthodologique garantit une évaluation cohérente, rigoureuse et alignée avec les objectifs stratégiques du projet.

4.2 Régression logistique

La régression logistique constitue un modèle linéaire de classification particulièrement adapté à la prédiction d'événements binaires. Bien que peu adaptée à des cas aussi déséquilibrés et non linéaires que les litiges brevets, elle a été choisie ici comme modèle de référence pour amorcer le projet. Son intérêt réside dans sa robustesse, sa rapidité d'entraînement, et surtout sa forte interprétabilité. Elle a permis de valider la chaîne de traitement (prétraitement, équilibrage, standardisation, sélection, évaluation) et d'identifier les premières variables discriminantes.

Adaptation du pipeline

Le pipeline appliqué à la régression logistique suit rigoureusement les étapes décrites dans la section *Méthodologie*, notamment :

— l'imputation des valeurs manquantes par la médiane,
— l'encodage fréquentiel des variables catégorielles,
— l'application de SMOTE (30 %) pour corriger le déséquilibre,
— la standardisation z-score des variables numériques.

Nous détaillons ci-dessous les spécificités propres à ce modèle.

Apprentissage et pondération

Le modèle a été entraîné à l'aide du solveur `saga`, particulièrement adapté aux jeux volumineux et compatible avec les régularisations ℓ_1 et ℓ_2. Afin de renforcer la prise en compte de la classe minoritaire, une pondération asymétrique a été introduite : les exemples positifs ont été affectés d'un poids cinq fois supérieur à ceux de la classe majoritaire.

Sélection de variables

Une sélection automatique a été appliquée via la méthode `SelectFromModel`. Seules les variables dont l'importance excède la moyenne des coefficients absolus ont été conservées, ramenant le modèle à 12 variables pertinentes. Cette réduction permet de limiter le surapprentissage tout en facilitant l'interprétation.

Les variables les plus contributives incluent :

— `year_filing`, `year_grant`, `grant_lag` : informations temporelles,
— `renewal`, `quality_index_4`, `DIV` : indices de qualité du brevet,
— `country_JP_PAD`, `UNIVERSITY`, `gov_int` : informations institutionnelles.

Optimisation du seuil de classification

Le seuil de décision a été optimisé à partir de la courbe précision/rappel, en recherchant le point maximisant le F1-score. Le seuil retenu est de 0.8971, ce qui reflète une volonté de limiter les faux positifs tout en capturant un maximum de cas litigieux.

Résultats obtenus

Les performances du modèle sur le jeu de test sont les suivantes :

— **Exactitude globale** : 96,1 %
— **Précision (classe litige)** : 5,4 %

— **Rappel (classe litige)** : 13,9 %
— **Score F1 (classe litige)** : 7,8 %
— **%E** : -27 % (<0, utilisation du modèle non rentable)
— η : 292 %

Matrice de confusion

	Prédit : Non-litige	**Prédit : Litige**
Réel : Non-litige	123 572 (TN)	3 710 (FP)
Réel : Litige	1 311 (FN)	212 (TP)

Le ratio TP/FP est de 0,057, soit environ 1 litige correctement détecté pour 17 fausses alertes.

Discussion

Sans surprise, les résultats révèlent les limites structurelles de la régression logistique dans un contexte aussi déséquilibré et non linéaire :
— faible capacité à capturer des interactions complexes,
— sensibilité persistante au déséquilibre des classes,
— ratio TP/FP encore trop faible pour un usage industriel.

La régression logistique a joué un rôle fondamental dans le projet : établir une première ligne de base robuste, évaluer la qualité du pipeline, et comprendre les variables clés du problème. Si ses performances sont insuffisantes pour un déploiement opérationnel, elle constitue une excellente référence pour évaluer les apports des modèles plus avancés.

4.3 Modèle Elastic Net

Après avoir établi une première ligne de base avec la régression logistique, nous avons exploré le modèle **Elastic Net**. Il conserve l'interprétabilité des modèles linéaires tout en y ajoutant une régularisation mixte (L1/L2), utile dans des contextes de multicolinéarité et de sélection de variables. Il constitue une extension naturelle et potentiellement plus robuste face au sur-apprentissage.

Adaptation du pipeline

Le pipeline appliqué suit rigoureusement les étapes décrites dans la section *Méthodologie*, à savoir :
— imputation des valeurs manquantes par la médiane,
— encodage fréquentiel des variables catégorielles,
— application de SMOTE (30 %) sur l'ensemble d'entraînement uniquement,
— standardisation des variables numériques (z-score).
Le modèle a été entraîné avec les paramètres suivants :
— `alpha = 0.1`,
— `l1_ratio = 0.7`,
— `max_iter = 5000`,
— seuil de classification : `0.5`.

Formulation mathématique Elastic Net minimise la fonction de coût suivante :

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \beta)^2 + \alpha \left[(1-\rho)\frac{1}{2}||\beta||_2^2 + \rho||\beta||_1 \right] \right\} \tag{4.1}$$

où α régule l'intensité de la pénalisation globale, et ρ ajuste le compromis entre L1 (sparse) et L2 (stabilité).

Résultats obtenus (avant sélection stricte)

Les performances sur le jeu de test sont les suivantes :
— **Exactitude (Accuracy)** : 54,84 %
— **Précision (classe litige)** : 1,99 %
— **Rappel (classe litige)** : 76,95 %
— **Score F1 (classe litige)** : 3,87 %
— **%E** : -551%(<0, utilisation du modèle non rentable)
— η : 817 %

Matrice de confusion

	Prédit : Non-litige	**Prédit : Litige**
Réel : Non-litige	69 463 (TN)	57 819 (FP)
Réel : Litige	351 (FN)	1 172 (TP)

Ratio TP / FP : 0,0203 (soit environ 1 litige correctement détecté pour 49 fausses alertes).

Réduction de dimension stricte

Une sélection stricte a été appliquée à l'aide de la méthode `SelectFromModel`, avec un seuil défini à 1.5× la moyenne des coefficients absolus. Seules 4 variables ont été conservées :
— `bwd_cits`,
— `quality_index_4`,
— `DIV`,
— `foreign_priority`.

Le modèle a ensuite été réentraîné sur ce sous-ensemble, donnant les mêmes résultats que précédemment. L'interprétabilité est ainsi renforcée, sans perte de performance notable.

Le modèle Elastic Net améliore le rappel au détriment d'une forte dégradation de la précision. Ce modèle n'est pas économiquement rentable avec une perte encore plus importante que celle de la regression logistique.

4.4 Random Forest

La *forêt aléatoire* est une méthode d'ensemble reposant sur une agrégation d'arbres de décision entraînés de manière indépendante. Elle offre une bonne robustesse aux données bruitées et permet de modéliser des relations non linéaires entre les variables. Son efficacité et sa simplicité d'implémentation en font un choix classique pour des tâches de classification.

Adaptation du pipeline

Le modèle Random Forest a été intégré au pipeline général tel que défini dans la section *Méthodologie*, incluant :

— Imputation des valeurs manquantes par la médiane,
— Encodage fréquentiel des variables catégorielles,
— Application de SMOTE sur le jeu d'entraînement (ratio 30 %),
— Standardisation par z-score (appliquée ici pour uniformité, bien que non indispensable aux arbres).

Paramétrage

Le modèle a été entraîné avec les paramètres suivants :

— `n_estimators = 200` : nombre d'arbres,
— `max_depth = 4` : profondeur maximale, pour limiter le sur-apprentissage,
— `class_weight = "balanced"` : pondération automatique des classes selon leur fréquence,
— `random_state = 42` : pour la reproductibilité.

Résultats sur l'ensemble de test

Le seuil de classification a été optimisé à l'aide de la courbe précision/rappel (F1-max). Les performances sur le jeu de test sont les suivantes :

— **Exactitude (Accuracy)** : 96,96 %
— **Précision (litiges)** : 6,89 %
— **Rappel (litiges)** : 12,54 %
— **F1-score (litiges)** : 8,90 %
— **%E** : -15%(<0, utilisation du modèle non rentable)
— η : 225 %

Matrice de confusion

	Prédit : Non-litige	**Prédit : Litige**
Réel : Non-litige	124 702 (TN)	2 580 (FP)
Réel : Litige	1 332 (FN)	191 (TP)

Analyse et interprétation

Malgré le rééquilibrage SMOTE, la forêt aléatoire peine à bien détecter la classe minoritaire. Le rappel reste modéré et la précision relativement faible, ce qui traduit un grand nombre de faux positifs. Cela s'explique notamment par :

— la structure rigide des arbres peu profonds,
— le choix conservateur de l'hyperparamètre `max_depth`,
— la forte dominance de la classe négative, difficile à compenser entièrement.

Conclusion

Le modèle Random Forest montre une bonne capacité de généralisation globale (accuracy élevée), mais ses performances sur la classe litigieuse restent insuffisantes dans un contexte d'usage opérationnel. Son utilisation n'est pas rentable économiquement. Il constitue toutefois un bon benchmark pour des approches plus ciblées comme XGBoost, mieux adaptées au déséquilibre des classes.

4.5 XGBoost

Le modèle *Extreme Gradient Boosting* XGBoost s'impose comme une référence incontournable pour des tâches de classification supervisée, notamment dans les contextes de forte non-linéarité et de déséquilibre de classes. Il repose sur un enchaînement d'arbres de décision faibles corrigés successivement, avec des mécanismes internes de régularisation et de contrôle du sur-apprentissage.

Motivations

Contrairement aux modèles linéaires, XGBoost est capable de capturer des interactions complexes entre variables, sans nécessiter de transformation explicite. Il supporte à la fois la régularisation ℓ_1 et ℓ_2, et offre des options fines de pondération pour traiter le déséquilibre de classes.

Prétraitement et pipeline

Le modèle a été intégré dans le pipeline décrit en début de section, incluant :
— Imputation par la médiane,
— Encodage fréquentiel des variables catégorielles,
— Rééquilibrage par SMOTE (ratio 30 %),
— Standardisation (non obligatoire mais conservée pour homogénéité),
— Suppression des variables à risque de fuite (`Year_Litigation`, `foreign_priority`).

Paramétrage spécifique

Les hyperparamètres ont été ajustés pour maximiser le rappel tout en limitant les faux positifs. Les choix notables incluent :
— `learning_rate = 0.01` : convergence lente mais stable,
— `max_depth = 3`, `gamma = 40` : contrôle de la complexité,
— `scale_pos_weight = 10` : accent mis sur la classe minoritaire,
— `subsample = 0.8`, `colsample_bytree = 0.7` : bagging interne,
— `reg_lambda = 40`, `reg_alpha = 25` : forte régularisation.

Résultats sur l'ensemble de test

— **Exactitude (accuracy)** : 84,97 %
— **Précision (litiges)** : 3,32 %
— **Rappel (litiges)** : 41,63 %
— **F1-score** : 6,15 %
— **%E** : -161%(<0, utilisation du modèle non rentable)

— η : 485 %
— **TP / FP** : 0,0343 (soit un vrai litige pour 29 faux positifs)

Matrice de confusion

	Prédit : Non-litige	**Prédit : Litige**
Réel : Non-litige	108 811 (TN)	18 471 (FP)
Réel : Litige	889 (FN)	634 (TP)

Sélection et importance des variables

Aucune réduction de dimension n'a été appliquée en amont pour le modèle XGBoost, celui-ci étant capable d'attribuer automatiquement des poids faibles aux variables peu informatives. Cela nous a permis d'inclure toutes les variables disponibles dans l'apprentissage initial, tout en bénéficiant des mécanismes internes de régularisation du modèle.

Cependant, certaines variables ont été retirées manuellement a posteriori, en particulier `Year_Litigation`, dont la corrélation directe avec la cible provoquait une fuite d'information, faussant les performances observées en créant du sur-apprentissage. De même, `foreign_priority` a été exclue en raison de sa distribution fortement déséquilibrée et de son pouvoir prédictif anormalement élevé.

Après entraînement, l'importance des variables a été mesurée à l'aide du critère de *gain moyen*, qui reflète l'apport moyen d'une variable dans la réduction de la fonction de perte lors des splits.

Les dix variables les plus importantes sont :
— `country_JP_PAD`, `fwd_cits5`, `quality_index_4`,
— `bwd_cits`, `renewal`, `generality`,
— `DIV`, `tech_field`, `claims`, `UNIVERSITY`.

Ces variables recouvrent à la fois des aspects techniques (citations, diversité), administratifs (renouvellement), et institutionnels (type de déposant). Leur interprétation a permis de valider les signaux identifiés par d'autres modèles (logistique, Elastic Net), tout en confirmant la robustesse de XGBoost à capturer des structures complexes dans les données.

Discussion

XGBoost surpasse nettement les modèles linéaires en rappel, mais souffre encore d'un taux élevé de faux positifs. Son potentiel reste élevé, à condition d'un réglage plus fin du seuil ou d'une combinaison avec d'autres modèles. Pour autant, ce modèle n'est pas rentable lui non plus. Son utilisation entraînerait des pertes conséquentes pour l'utilisateur. En l'état il n'est pas adapté à notre problématique.

4.6 Réseau de Neurones

Aucun des modèles testés précédemment n'est rentable et ceci de manière robuste. Nous avons donc décidé d'implémenter un réseau de neurones multi couches tout en restant de petite taille, afin de tester si un modèle plus profond était plus adapté à notre problématique.

Dès les premiers tests, une structure très simple de 2 couches de neurones a conduit à un modèle rentable économiquement et présentant un bon niveau de robustesse. Nous

avons testé d'autres structures en augmentant le nombre de couches et en affinant les hyper-paramètres pour optimiser notre prédiction. Une fois notre meilleure configuration identifiée, nous l'avons implémentée dans une interface utilisateur.

Architecture du modèle

L'architecture choisie repose sur une structure séquentielle classique :
— Une couche d'entrée de taille égale au nombre de variables explicatives ;
— 2 à 4 couches cachées comportant respectivement 64, 32 (puis 16 et 8) neurones, activés par la fonction ReLU ;
— Une couche de sortie à un neurone, avec activation sigmoïde, produisant une probabilité d'appartenance à la classe « litige ».

Des couches de *dropout* (30 %) sont insérées entre chaque couche cachée afin de limiter le sur-apprentissage. La fonction de perte utilisée est l'entropie croisée binaire (`binary_crossentropy`), optimisée avec l'algorithme Adam.

Évaluation par validation croisée

Pour évaluer la robustesse du modèle et éviter toute dépendance au découpage initial des données, nous avons opté pour une validation croisée 5-Folds. Cette méthode garantit une évaluation plus fidèle des performances, en préservant la proportion de litiges dans chaque sous-échantillon.

À chaque itération (Fold), un pipeline complet est réappliqué :

1. Découpage des données d'entraînement et de test selon les indices du `StratifiedKFold` ;
2. Rééquilibrage par SMOTE (ratio 30 %) sur l'ensemble d'entraînement ;
3. Standardisation des variables numériques (z-score) ;
4. Entraînement avec `EarlyStopping` pour éviter le sur-apprentissage ;
5. Évaluation sur le jeu de test associé au pli courant.

Les poids des modèles obtenus sur chacun des 5 Folds sont sauvegardés et leurs matrices de confusion sont comparées. Nous avons comparé en particulier la précision p, le rappel r, l'économie relative estimée $\%E$ et l'indicateur η de la robustesse du gain économique.

Résultats comparés des MLPs

Nous avons réalisé 5 apprentissages pour chacune des 3 structures avec 2, 3 ou 4 couches de neurones cachés (hors "dropout"). A l'issu de chaque apprentissage, le modèle est testé sur le dataset de test. Les résultats sont synthétisés dans le tableau 4.1.

TABLE 4.1 – Comparatif de différents modèles MLPs

Modèle	**# Fold**	**p**	**r**	**Eta**	**%E**
2 couches					
2layers-64-32	1	61%	12%	11%	11%
2layers-64-32	2	57%	10%	13%	9%
2layers-64-32	3	57%	11%	13%	10%
2layers-64-32	4	60%	10%	11%	9%
2layers-64-32	5	44%	13%	21%	10%
3 couches					
3layers-64-32-16	1	57%	13%	13%	11%
3layers-64-32-16	2	56%	14%	13%	12%
3layers-64-32-16	3	46%	12%	20%	10%
3layers-64-32-16	4	53%	13%	15%	11%
3layers-64-32-16	5	52%	11%	15%	9%
4 couches					
4layers-64-32-16-8	1	58%	13%	12%	11%
4layers-64-32-16-8	2	52%	12%	15%	10%
4layers-64-32-16-8	3	52%	12%	15%	10%
4layers-64-32-16-8	4	56%	12%	13%	10%
4layers-64-32-16-8	5	53%	10%	15%	9%

Nous pouvons observer que les performances varient légèrement selon les Folds, mais restent relativement stables. Pour l'ensemble des structures, les modèles sont tous économiquement rentables avec une économie positive située entre 9% et 12%. Les valeurs de η comprises entre 11% et 21% montrent une bonne robustesse aux hypothèses de valeurs des brevets et des coûts des litiges. Même s'ils étaient corrigés d'un facteur 5, l'utilisation du modèle resterait rentable. Les performannces évoluent faiblement avec l'ajout ou le retrait d'une couche cachée. En moyenne, on constate 1 point d'écart sur $\%E$ et η en passant d'un modèle à 3 couches à un modèle à 2 ou 4 couches.

Le rappel global est modeste (entre 10 % et 14 %), indiquant que de nombreux litiges ne sont pas détectés. Toutefois, dans notre contexte applicatif, ce compromis reste acceptable : l'enjeu est moins de tout détecter que de proposer des alertes suffisamment fiables pour être analysées sans saturer les ressources humaines. En ce sens, un ratio TP/FP très supérieur à 0,17 (équivalent à $\eta < 1$ ou $p > 14\%$) constitue une avancée significative et confère un intéret économique à notre modèle. Des améliorations futures pourraient viser à renforcer le rappel sans trop sacrifier la précision, par exemple en ajustant dynamiquement le seuil de décision, ou en explorant des architectures plus profondes.

Nous avons retenu en référence le modèle "3layers-64-32-16" du Fold-2 qui présente la meilleure performance en $\%E$ et un η parmi les plus faibles. Ce sont ses poids qui on été implémentés dans l'interface utilisateur. Sa structure avec 3 couches cachées reste très légère et permet une exécution rapide même sur un PC portable dépourvu de GPU.

Matrice de confusion du modèle sélectionné

	Prédit : Non-litige	**Prédit : Litige**
Réel : Non-litige	127 115 (TN)	166 (FP)
Réel : Litige	1 311 (FN)	213 (TP)

— **Précision (classe litige)** : 56%
— **Rappel (classe litige)** : 14%
— **F1-score (classe litige)** : 22%
— **Exactitude globale (accuracy)** : 98.9%
— **%E** : 12%(>0, utilisation du modèle rentable)
— η : 13% («1, bonne robustesse aux hypothèses)
— **TP / FP** : 1,28

4.7 Synthèse comparative

La présente section propose une vue d'ensemble des performances obtenues par les cinq modèles évalués, afin de mettre en évidence les compromis entre précision, rappel, robustesse et interprétabilité. Chaque modèle est replacé dans le contexte du problème posé, avec une attention particulière portée aux ratios $\%E$ et η, les indicateurs clés pour notre problématique.

Comparaison des métriques principales

Modèle	**Accuracy**	**Précision**	**Rappel**	**F1-score**	$\%E$	η
Régression logistique	96.1 %	5.4 %	13.9 %	7.8 %	-27 %	292 %
Elastic Net	54.8 %	2.0 %	76.9 %	3.9 %	-552 %	817 %
Random Forest	96.96 %	6.9 %	12.5 %	8.9 %	-16 %	225 %
XGBoost	84.97 %	3.3 %	41.6 %	6.2 %	-160 %	485 %
Réseau de Neurones (MLP)	98.9 %	56 %	14 %	22 %	12 %	13 %

TABLE 4.2 – Comparaison des performances sur la classe minoritaire (litiges)

Analyse transversale

Les modèles Elastic Net, XGBoost présentent des niveaux de rappel supérieurs à tous les autres. Toutefois cette performance en terme de rappel est obtenue au détriment de la précision qui est inférieure à 4%. Dans notre contexte, un niveau aussi faible de précision est rédhibitoire. En effet, une précision $p < 14\%$ conduit mécaniquement à des valeurs de $\eta > 1$ et donc à une utilisation du modèle entraînant des coûts supérieurs que sans l'utiliser. Le critère principale pour la viabilité économique est donc d'avoir un modèle qui propose une valeur de $\eta < 1$ (voire $\eta << 1$ pour s'assurer d'une robustesse suffisante par rapport à nos estimations) donc un niveau de précision $p > 14\%$. Seuls les modèles MLP offrent un niveau de précision qui dépasse ce seuil. Par conséquent, seuls les modèles MLP offrent une rentabilité positive pour notre étude.

Chapitre 5

Résultats et interface

5.1 Interprétation des performances et du ratio TP/FP

La performance du modèle MLP ne peut être résumée à un simple score d'accuracy. En effet, dans un contexte fortement déséquilibré comme celui des litiges de brevets, les métriques classiques peuvent être trompeuses. L'équation 3.5 a fait apparaître que que l'économie ou la perte financière résultant de l'utilisation du modèle sont conditionnées par le signe de $1 - \eta$. Pour que l'utilisation du modèle soit rentable il est nécessaire que $\eta < 1$ ce qui est équivalent à $TP/FP > \gamma$ soit $0,17$ avec nos hypothèses ou encore à $p > \frac{\gamma}{1+\gamma}$ soit 14%.

C'est pourquoi notre attention s'est portée en premier lieu sur la valeur du coefficient η (ou sur `TP/FP`, ou sur p comparés à leur seuils respectifs), qui sont les indicateurs de rentabilité de l'utilisation du modèle.

Il convient de souligner que le modèle ne vise pas l'exhaustivité donc un haut rappel, mais plutôt la qualité des alertes donc précision et rentabilité. Ce positionnement stratégique s'aligne avec les enjeux économiques définis en amont du projet.

Avec des précisions très supérieures à 14% et malgré un niveau de rappel très faible, les modèles MLP sont les seuls à offrir une rentabilité positive, donc à présenter un intérêt économique.

5.2 Interface utilisateur

Afin de rendre le modèle accessible aux utilisateurs non techniques, une interface graphique a été développée (cf. figure 5.1). Celle-ci permet de renseigner manuellement les caractéristiques d'un brevet au moment de son dépôt. Le bouton `Submit` déclenche une prédiction en temps réel, affichant la probabilité estimée de litige, tandis que le bouton `Save to Logs` permet d'enregistrer la saisie pour un traitement ultérieur.

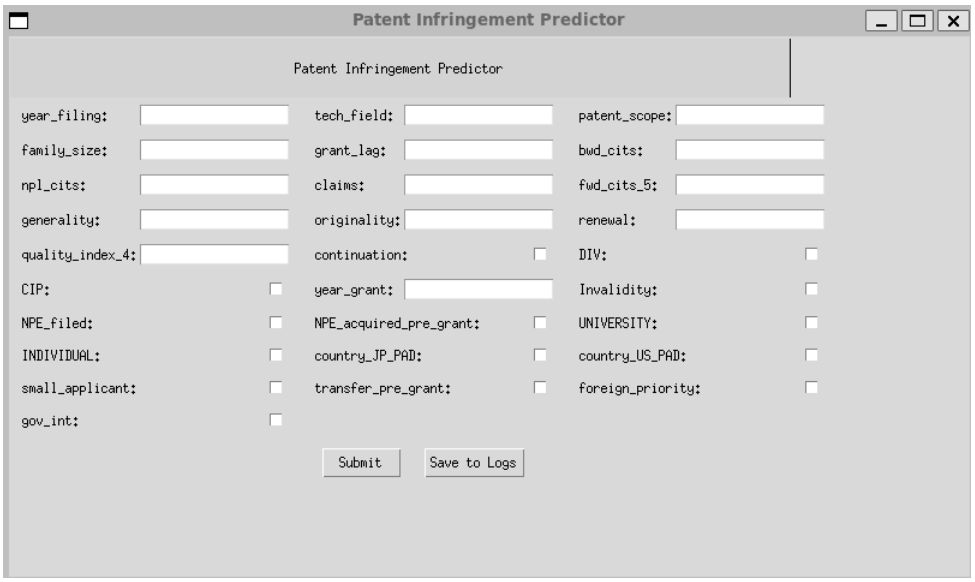

FIGURE 5.1 – Interface graphique de prédiction du risque de litige

Les champs avec un carré simple représentent des variables binaires du dataset. Ce prototype offre un socle fonctionnel pour un outil d'aide à la décision. Il pourrait, à terme, être intégré à des systèmes d'information brevets, ou couplé à des interfaces web à destination des services juridiques et R&D.

Nous illustrons ci-dessous deux cas concrets d'utilisation de l'interface, basés sur des exemples réels du dataset. Ils montrent la capacité du modèle à distinguer efficacement les brevets à risque de ceux qui ne le sont pas.

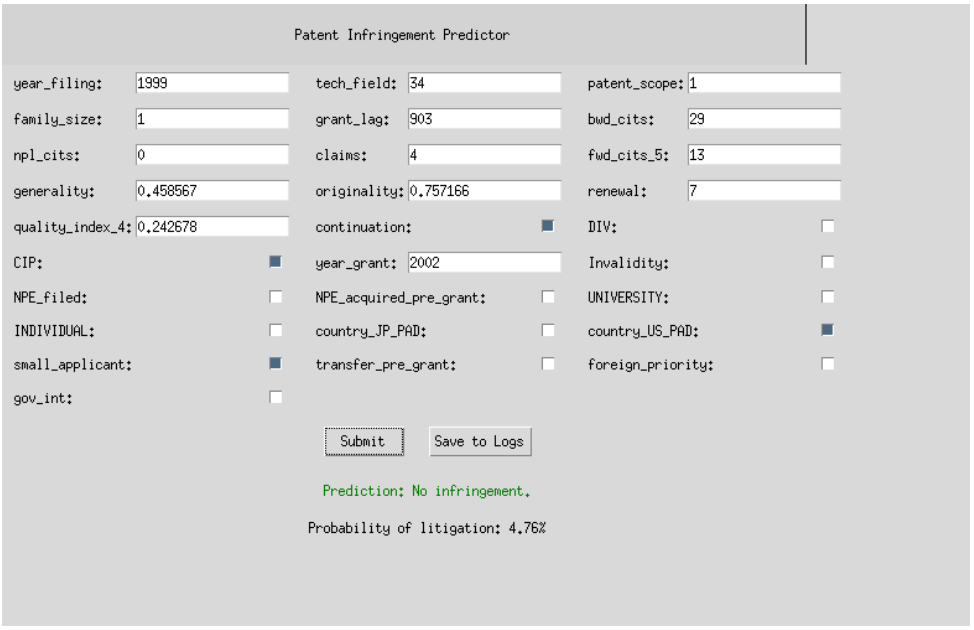

FIGURE 5.2 – Exemple de prédiction correcte : brevet non litigieux détecté comme sûr (probabilité de litige : 4.76 %)

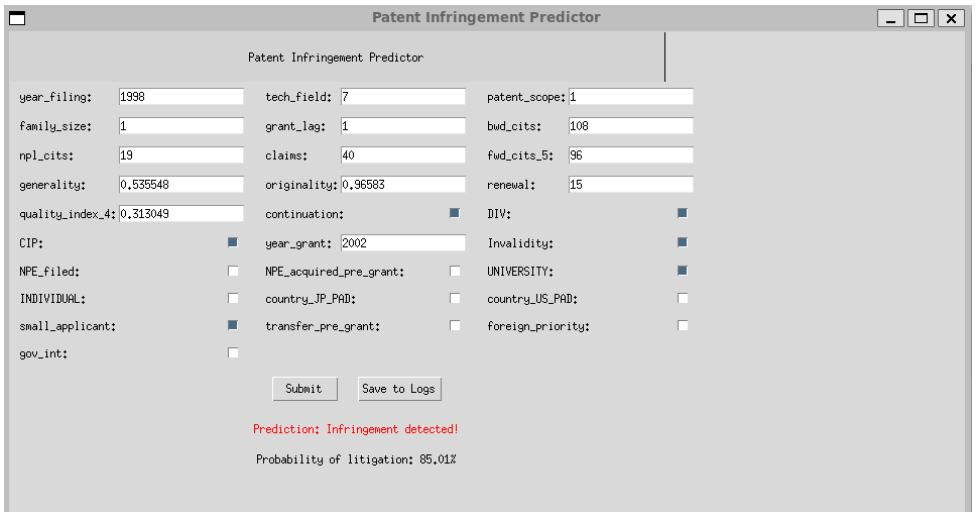

FIGURE 5.3 – Exemple de prédiction correcte : brevet litigieux détecté comme risqué (probabilité de litige : 85.01 %)

La première prédiction (figure 5.2) correspond à un brevet historiquement non litigieux, pour lequel le modèle renvoie une faible probabilité de litige. La deuxième (figure 5.3) concerne un brevet effectivement impliqué dans un contentieux, détecté comme tel avec une forte probabilité. Ces cas illustrent la cohérence entre les résultats produits par le modèle et les données historiques.

Chapitre 6

Conclusion et perspectives

6.1 Bilan du projet

Ce projet a permis de construire un pipeline complet de prédiction du risque de litige, depuis la préparation des données jusqu'à l'intégration dans une interface fonctionnelle. Plusieurs familles de modèles ont été comparées — linéaires, arborescents, et neuronaux.

Ce projet a été l'occasion de mettre en œuvre de manière plus approfondie ces différents types de modèles. Il nous a aussi permis de prendre du recul sur les métriques classiquement utilisées et de constater que même si elles restaient au cœur des critères pertinents pour notre étude, elles étaient ici combinées d'une manière très spécifique et nous ont conduit à rechercher avant tout à dépasser un seuil de précision même au détriment du rappel si on voulait garantir l'intérêt économique pour l'utilisateur.

Le réseau de neurones s'est imposé comme le modèle offrant le meilleur compromis entre détection et pertinence des alertes.

6.2 Limites et améliorations possibles

Malgré ces résultats encourageants, plusieurs limites doivent être soulignées :

- Le rappel reste modeste ($\sim$14 %), ce qui signifie que 86% des litiges ne sont pas détectés. Même si l'utilisation du modèle reste économiquement rentable, d'un point de vue marketing annoncer que le modèle ne détectera pas 86% des brevets litigieux est assez pénalisant pour la crédibilité de notre outil.
- Le modèle reste tributaire de la qualité des variables disponibles, certaines dimensions juridiques n'étant pas représentées.

Plusieurs pistes peuvent être envisagées pour améliorer les performances et l'utilisabilité du modèle comme le recours à des architectures hybrides (XGBoost + MLP) pour améliorer le rappel tout en conservant le niveau de précision. Le recours à des modèles NLP pour extraire de l'information complémentaire basée sur le contenu effectif du brevet est aussi une piste à explorer.

6.3 Déploiement potentiel

L'interface développée constitue une preuve de concept prometteuse. Dans une perspective de déploiement :

— une interface web ou une API pourrait remplacer la version locale actuelle ;
— le modèle pourrait être mis à jour régulièrement avec de nouvelles données pour suivre l'évolution du contentieux ;
— l'outil pourrait être utilisé en phase d'analyse de portefeuille brevets, pour hiérarchiser les risques et prioriser les audits juridiques.

En définitive, bien que le modèle actuel n'offre pas une solution parfaite ni entièrement automatisable, il constitue un levier opérationnel utile pour mieux anticiper les litiges, avec une logique orientée vers le retour sur investissement. Il pose les bases d'une approche plus fine, plus stratégique, et plus économique de la gestion de la propriété industrielle.

Bibliographie

[1] Juranek, S., & Otneim, H. (2024). *Predicting patent lawsuits with machine learning*. International Review of Law & Economics, 80, 106228. https://dynresmanagement.com/uploads/3/5/2/7/35274584/patent_predictions.pdf

[2] Wongchaisuwat, P., Klabjan, D., & McGinnis, J. O. (2024). *Predicting litigation likelihood and time to litigation for patents*. Patent Predictions. https://doi.org/10.1016/j.irle.2024.106228

[3] Millon de La Verteville, O., & Desrousseaux, G. (2021). *Le contentieux des brevets en France – Paroles d'experts*. IEEPI. https://www.ieepi.org/paroles-dexperts-contentieux-brevets-france/

[4] OMPI (2022). *Faits et chiffres en matière de PI 2022*. WIPO-PUB-943-2022-FR, Organisation Mondiale de la Propriété Intellectuelle. https://www.wipo.int/edocs/pubdocs/fr/wipo-pub-943-2022-fr-wipo-ip-facts-and-figu pdf

[5] (2024). *Cours de Data Mining*. CYTECH.

[6] Powell, K. (2024). *Elastic Net Regression : Theory and Application*. Vidéo YouTube. https://www.youtube.com/watch?v=M0UVUHy7l1M&ab_channel=KodyPowell