UAB
Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

This is the **published version** of the bachelor thesis:

Sallas Pineda, Guillem. *Predicció del risc de lesions traumàtiques de genoll.* Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Grau en Enginyeria de Dades)

This version is available at https://ddd.uab.cat/record/337334

under the terms of the  license.

# Predicció del risc de lesions traumàtiques de genoll

Guillem Sallas Pineda - 1641076

**Resum**— Les lesions traumàtiques de genoll són una de les patologies més freqüents en medicina esportiva, i la seva naturalesa multifactorial, cap factor aïllat no n'explica la majoria, limita els models predictius simples i manté la decisió clínica ancorada en criteris qualitatius. Aquest treball encara el problema des de l'enginyeria de dades: dissenya, implementa i avalua un flux de treball reproduïble d'aprenentatge automàtic per estimar el risc lesional, materialitzat en l'eina de suport a la decisió GenollGuard Pro. Davant la inaccessibilitat de dades clíniques reals i l'absència de datasets públics que integrin els dominis d'interès, es genera un dataset sintètic parametritzat (n=5.000) amb correlacions fonamentades en la literatura, sobre el qual s'entrenen i comparen tres classificadors, Regressió Logística, Random Forest i Gradient Boosting, amb optimització d'hiperparàmetres per validació creuada estratificada, prioritzant el Recall, i s'interpreten amb anàlisi SHAP. El resultat rellevant no és el rendiment aparent dels models, conseqüència del biaix circular del dataset, documentat de manera transparent, sinó la robustesa del pipeline, la coherència de l'arquitectura de dades i la viabilitat de l'eina. Es completa amb un pipeline d'integració de dades reals (Ferber et al. 2024, n=1.685; KneE-PAD, n=31) executat sobre 3.716 registres totals (46% reals) que valida empíricament el biaix circular: l'AUC del Gradient Boosting cau de 0,998 sobre dades sintètiques pures a 0,910 sobre el conjunt multi-font, dins el rang documentat per la literatura per a prediccions de lesió sobre cohorts reals.

**Paraules clau**— Aprenentatge automàtic, lesions de genoll, LCA, predicció de risc, dades sintètiques, Random Forest, Gradient Boosting, SHAP, enginyeria de característiques.

**Abstract**— Traumatic knee injuries are among the most frequent conditions in sports medicine, and their multifactorial nature,no single factor accounts for most cases, limits simple predictive models and keeps clinical decisions anchored in qualitative criteria. This work tackles the problem from a data-engineering perspective: it designs, implements, and evaluates a reproducible machine-learning workflow for injury-risk estimation, materialized in the decision-support tool GenollGuard Pro. Given the inaccessibility of real clinical data and the absence of public datasets integrating the relevant domains, a parameterized synthetic dataset (n=5,000) is generated with literature-grounded correlations; on it, three classifiers, Logistic Regression, Random Forest, and Gradient Boosting, are trained and compared with stratified cross-validated hyperparameter tuning, prioritizing Recall, and interpreted through SHAP analysis. The key outcome is not the models' apparent performance, a consequence of the dataset's circular bias, transparently documented, but the robustness of the pipeline, the consistency of the data architecture, and the tool's technical feasibility. The work is completed with a real-data integration pipeline (Ferber et al. 2024, n=1,685; KneE-PAD, n=31) executed on 3,716 total records (46% real), empirically validating the circular bias: Gradient Boosting AUC drops from 0.998 on purely synthetic data to 0.910 on the multi-source set, within the range documented by the literature for injury prediction on real cohorts.

**Index Terms**— Machine learning, knee injuries, ACL, risk prediction, synthetic data, Random Forest, Gradient Boosting, SHAP, feature engineering.

## 1 Introducció - Context del treball

Les lesions traumàtiques de genoll concentren una part desproporcionada de la càrrega assistencial en medicina esportiva i fisioteràpia. La lesió del LCA, les lesions meniscals i les inestabilitats rotacionals presenten una incidència que ha crescut de manera sostinguda l'última dècada [1, 2, 9], amb un cost que no és només sanitari sinó també Esportiu, temps fora de competició, risc de recidiva i, sovint, una recuperació funcional incompleta.

El que fa difícil el problema és la seva estructura. La lesió no respon a una causa única: emergeix de la interacció entre dominis estructurals, funcionals i de càrrega, i cap variable aïllada n'explica per si sola la majoria de casos [3, 5, 8, 18, 32]. Aquesta multifactorialitat és precisament la raó per la qual els models lineals simples i els llindars clínics aïllats tenen un sostre. A la pràctica, la decisió continua basant-se majoritàriament en criteris qualitatius i en l'experiència del professional [11, 12], amb una integració quantitativa limitada de tota la informació disponible.

L'aprenentatge automàtic (ML) ofereix una via per identificar combinacions de factors amb poder predictiu, estimar el risc individualitzat i optimitzar la planificació de la càrrega en rehabilitació [18, 19, 20, 21]. La proposta d'aquest treball no és substituir el judici clínic, sinó complementar-lo: afegir un quart element, l'anàlisi

predictiva basada en dades, als tres pilars clàssics de la fisioteràpia basada en l'evidència [12].

### 1.1 Problema i hipòtesi

La pregunta que vertebra el projecte és si la integració multidimensional de dades clíniques, biomecàniques i funcionals mitjançant ML pot millorar l'estimació del risc lesional respecte de l'avaluació basada en factors aïllats. La hipòtesi de treball és que sí: que un model entrenat sobre la combinació de dominis recupera jerarquies de risc coherents amb la literatura i que aquesta combinació, i no cap variable individual, és la que aporta capacitat discriminatòria.

El problema pràctic és que aquesta hipòtesi no es pot contrastar directament sobre dades reals dins del marc d'un TFG: l'accés a registres clínics que integrin tots aquests dominis simultàniament està restringit per raons ètiques i de protecció de dades [14], i no existeix cap dataset públic que els reuneixi. La decisió metodològica, que condiciona tot el treball i la interpretació dels resultats, ha estat construir un dataset sintètic parametritzat com a banc de proves de l'arquitectura, i dissenyar després la transició cap a dades reals.

### 1.2 Objectius

L'objectiu general del projecte és dissenyar, implementar i validar un flux de treball reproduïble d'aprenentatge automàtic per a l'estimació del risc de lesió traumàtica de genoll, materialitzat en una eina de suport a la decisió clínica. Aquest objectiu es desglossa en set objectius específics, que s'utilitzen més endavant com a graella d'avaluació del que s'ha assolit:

1. Construir un marc teòric i clínic que identifiqui, a partir de la literatura, els factors de risc rellevants i les seves relacions.
2. Dissenyar i generar un dataset sintètic clínicament parametritzat que integri els dominis demogràfic, clínic, biomecànic, neuromuscular, de càrrega i de percepció subjectiva.
3. Dur a terme una anàlisi exploratòria (EDA) que validi la coherència interna del dataset i la hipòtesi multifactorial.
4. Implementar i avaluar comparativament un mínim de tres models de classificació supervisada amb una metodologia rigorosa i sense filtració de dades.
5. Interpretar la importància de les variables mitjançant SHAP i connectar-la amb la literatura biomecànica de referència.
6. Desenvolupar GenollGuard Pro, una eina utilitzable amb perfils d'usuari diferenciats i explicabilitat per pacient.
7. Dissenyar i implementar la via de transició cap a dades clíniques reals i la seva validació prospectiva.

### 1.3 Estructura del document

El dossier segueix l'ordre lògic del projecte. El punt 2 estableix el marc teòric i l'estat de l'art. El punt 3 descriu la metodologia i l'arquitectura de dades. Els punt 4 i punt 5 documenten el dataset i la implementació del model. El punt 6 presenta els resultats sobre dades sintètiques i el punt 7, l'enginyeria de característiques avançada i l'eina GenollGuard Pro. El punt 8 detalla el disseny d'integració de dades reals. El punt 9 conté la discussió crítica, el punt 10 fa explícit el grau d'assoliment de cada objectiu i el punt 11 recull les conclusions, ampliades amb les mètriques obtingudes en la integració real. Es tanca amb les referències i els annexos.

## 2 Marc teòric i estat de l'art

### 2.1 Lesions de genoll i factors de risc

La literatura biomecànica ha establert un conjunt de factors de risc amb suport prospectiu. Hewett et al. (2005) van demostrar que el moment d'abducció del genoll durant l'aterratge i el valgus dinàmic associat són els predictors biofísics de lesió de LCA de major consistència en atletes femenines, amb un llindar clínic operatiu al voltant dels 10° [7, 27]. L'asimetria de força entre extremitats per sobre del 15% és un criteri reconegut de risc i de decisió en el retorn esportiu [13], i el dèficit de control neuromuscular s'associa de manera robusta a un risc més elevat [5, 8].

Per sobre dels factors individuals, el model multifactorial de Meeuwisse (1994) ofereix el marc conceptual que justifica l'enfocament d'aquest treball: la lesió és el resultat d'una cadena de factors predisposants i precipitants, i l'antecedent lesional hi figura com a factor de risc intrínsec de primer ordre [5]. Bittencourt et al. (2016) van anar més enllà i van proposar abordar les lesions esportives com a sistemes complexos, desplaçant el focus de la identificació de factors aïllats al reconeixement de patrons d'interacció [18, 32], una formulació que encaixa de manera natural amb els mètodes d'ensemble emprats aquí.

### 2.2 Aprenentatge automàtic en predicció de lesions esportives

Els treballs de referència en predicció de lesions amb ML reporten rendiments moderats, coherents amb la dificultat intrínseca del problema sobre dades reals. Ruddy et al. (2018) van obtenir AUC-ROC d'entre 0.71 i 0.79 en lesions musculars de futbolistes professionals [19]; Carey et al. (2018) van reportar valors entre 0.68 i 0.82 a l'Australian Football League [20]; i Rossi et al. (2018) van assolir una precisió al voltant del 80% en futbolistes italians [21]. Aquest rang (0.68–0.85) és la referència realista contra la qual cal contrastar qualsevol pipeline quan s'aplica sobre dades clíniques reals, i marca l'expectativa de rendiment de la fase de validació prospectiva d'aquest projecte.

### 2.3 Explicabilitat i suport a la decisió clínica

Perquè un model de classificació esdevingui una eina clínica i no una caixa negra, la predicció agregada no n'hi ha prou: cal poder explicar, per a cada pacient, quins factors empenyen el risc en una direcció o una altra. Les tècniques d'explicabilitat per instància, SHAP [37] i LIME [38], permeten generar aquests informes individualitzats. SHAP, basat en valors de Shapley, ofereix una atribució consistent de la contribució de cada variable a la predicció concreta, i és la tècnica que s'utilitza en aquest treball per

transformar el classificador en una eina d'utilitat clínica directa.

Aquesta orientació és coherent amb el paradigma de la fisioteràpia basada en l'evidència [11, 12]: una eina algorítmica no substitueix cap dels tres pilars clàssics, evidència científica, experiència clínica i preferències del pacient, sinó que enriqueix l'evidència disponible amb una quarta entrada quantitativa.

# 3 Metodologia i arquitectura de dades

## 3.1 Enfocament metodològic

El projecte s'organitza com un cicle de vida de dades complet, des de la definició del problema fins al desplegament d'una eina, seguint una lògica propera a CRISP-DM. Cada fase produeix un artefacte verificable: el marc teòric (Fase 1), el dataset i l'EDA (Fase 2), el pipeline de modelització i l'eina (Fase 3) i la integració de dades reals i la memòria final (Fase 4). Aquesta seqüenciació, i no només el resultat final, és el que es documenta, perquè el valor del treball rau tant en les decisions com en els números.

## 3.2 Arquitectura Medallion

El pipeline s'estructura segons una arquitectura Medallion de quatre capes, una convenció d'enginyeria de dades que separa responsabilitats i fa el sistema auditable:

- **Bronze —** generació i ingesta de les dades en brut (dataset sintètic parametritzat, o dades clíniques reals en la transició).
- **Silver —** neteja, tractament de valors absents, validació de tipus i coherència, i normalització.
- **Gold —** enginyeria de característiques, conjunt de dades llest per a modelització i taules analítiques.
- **MLOps —** entrenament, optimització, avaluació, serialització dels models (joblib) i empaquetament de l'eina.

La separació en capes és el que permet, per exemple, substituir la font Bronze (sintètica per real) sense tocar la lògica de modelització, una propietat que s'explota directament en el disseny de la Fase 4.

## 3.3 Entorn tècnic i reproducibilitat

Tot el codi s'executa en Python 3.11 sobre un entorn conda aïllat (tfg_genoll), una decisió que va resoldre un conflicte de versions detectat durant el projecte, l'intèrpret per defecte d'Anaconda (3.8) provocava incompatibilitats entre SHAP i numba, i que garanteix la reproducibilitat de l'anàlisi. Les biblioteques principals són scikit-learn (GridSearchCV, RFECV, StandardScaler), SHAP per a l'explicabilitat i joblib per a la serialització dels models. La reproducibilitat es reforça amb una llavor aleatòria fixa (seed=42) a la generació del dataset, a la partició i a l'entrenament.

L'estructura de scripts separa cada responsabilitat: la generació del dataset i l'EDA, el pipeline base de modelització, el pipeline avançat amb l'arquitectura Medallion, el pipeline d'integració de dades reals i les dues versions de l'eina (base i Pro). El detall complet es recull a l'Annex B i el codi és accessible al repositori públic del projecte [41].

El desenvolupament d'una part del codi, s'ha fet amb assistència d'un agent d'intel·ligència artificial generativa com a eina de programació en parella. L'agent ha aportat esborranys d'estructura, suggeriments de refactorització i ha facilitat el debug d'errors detectats durant l'execució. Les decisions metodològiques, arquitectura Medallion, harmonització entre datasets, definició de la variable objectiu per a Ferber, tria d'algoritmes i de mètriques, són pròpies, i tot el codi ha estat revisat, executat i validat manualment. L'ús de l'eina s'ha fet de manera transparent, amb l'objectiu d'accelerar la implementació i augmentar el rigor metodològic.

## 3.4 Consideracions ètiques i de dades

L'ús de dades sintètiques no és només una solució pràctica a la manca d'accés: és també la decisió èticament conservadora, perquè evita tractar dades clíniques sensibles en una fase de disseny on encara s'està validant l'arquitectura. La generació de pacients sintètics com a pas legítim de la fase de disseny de sistemes predictius està reconeguda en la literatura d'enginyeria biomèdica [16, 35]. La transició cap a dades reals (§8) preveu explícitament els requisits de protecció de dades i de marc regulatori europeu com a part del disseny, no com una qüestió a resoldre a posteriori.

# 4 Dataset sintètic, variables i preprocessament

## 4.1 Justificació del dataset sintètic

La inaccessibilitat de dades clíniques reals i l'absència de datasets públics que integrin simultàniament els dominis d'interès van motivar la generació d'un dataset sintètic parametritzat (n=5.000, seed=42) com a estratègia de validació prèvia de l'arquitectura del model. Aquesta aproximació té un cost evident, que es discuteix sense pal·liatius al §9.1, però també un benefici concret: permet controlar exactament l'estructura de risc subjacent i, per tant, comprovar si el pipeline és capaç de recuperar-la, que és el que valida la fase de disseny.

## 4.2 Diccionari de dades

El dataset integra 18 predictors i 1 variable objectiu, distribuïts en cinc dominis més una variable de percepció subjectiva. La taula següent resumeix les variables i la seva justificació clínica; el diccionari complet, amb distribucions i rangs, es troba a l'Annex C i a l'Informe de Seguiment 1.

| Domini | Variables | Fonament clínic |
|---|---|---|
| Demogràfic | Edat, sexe, IMC, nivell activitat, esport contacte | Modulen l'exposició i la susceptibilitat estructural |
| Clínic | Lesió prèvia, num lesions prev, mesos última lesió | L'antecedent lesional és el factor intrínsec de primer ordere [5] |
| Biomecànic | Valgus dinàmic, moment adbucció, asimetria força, flexió aterratge, mobilitat turmell | Predictors prospectius de LCA, valgus ≥ 10º i asimetria > 15% són llindars de risc [7, 13, 27] |
| Neuromusuclar | Ratio isq quad, control neuromuscular | El dèficit de control motor i la ràtio H/Q <0.55 indiquen desequilibri de risc [5, 6, 8, 28]. |
| Càrrega | Hores entrenament, monotonia | La sobrecàrrega acumulativa és un predictor independent [9, 33]. |
| Percepció | Koos score (KOOS) | Funcionalitat percebuda; correlació inversa amb el risc [12]. |

### 4.3 Generació parametritzada

Cada variable s'ha generat seguint la seva distribució estadística corresponent i incorporant correlacions clínicament lògiques documentades per Hewett et al. (2005) [7, 27], Meeuwisse (1994) [5] i Leppänen et al. (2017) [9], amb un component de variabilitat estocàstica ($\sigma$=0.05) que simula la naturalesa probabilística del risc. La variable objectiu és una classificació binària (lesio ∈ {0,1}) definida al percentil 65 d'un risc latent ponderat, cosa que resulta en una distribució de 65% no-lesió (n=3.250) i 35% lesió (n=1.750).

Aquest mecanisme de generació, un risc latent calculat com a combinació ponderada de les prediccions és, alhora, la força i la debilitat del dataset. La força és que les relacions són clíniques i conegudes; la debilitat és el biaix circular que se'n deriva, tractat al §9.1.

### 4.4 Tractament de valors absents

El dataset brut presentava valors absents en tres variables, cadascun amb un mecanisme i una estratègia d'imputació diferents:

- **koos_score** — 200 nuls (4%), mecanisme MCAR; imputació per mediana estratificada per classe.
- **mobilitat_turmell** — 150 nuls (3%), mecanisme MCAR; imputació per mediana global.
- **mesos_ultima_lesio** — 3.486 NaN estructurals, mecanisme MAR condicionat a lesio_previa=0; substitució per 0 (absència d'antecedent).

El dataset net conté 5.000 registres complets sense valors absents. La injecció de valors absents MCAR és proporcional a la mida del dataset (4% i 3%, respectivament), de manera que el procés és independent de n. La imputació utilitza assignació directa en lloc d'operacions in-place per garantir la compatibilitat amb versions recents de pandas (copy-on-write). La normalització de les variables numèriques contínues es va ajornar deliberadament a la Fase 3 per evitar que condicionés l'EDA.

### 4.5 Resum de l'anàlisi exploratòria

L'EDA (documentada completament a l'IS1) confirma una estructura de correlacions internament coherent amb les parametritzacions introduïdes. La correlació positiva entre valgus_dinamic i moment_abduccio (r≈0.65) reprodueix la relació biomecànica de Hewett et al. (2005) [7], i koos_score correlaciona inversament amb la lesió (r=−0.36). El punt clau de l'EDA per a la modelització és que cap variable única domina la predicció, cosa que valida la hipòtesi multifactorial.

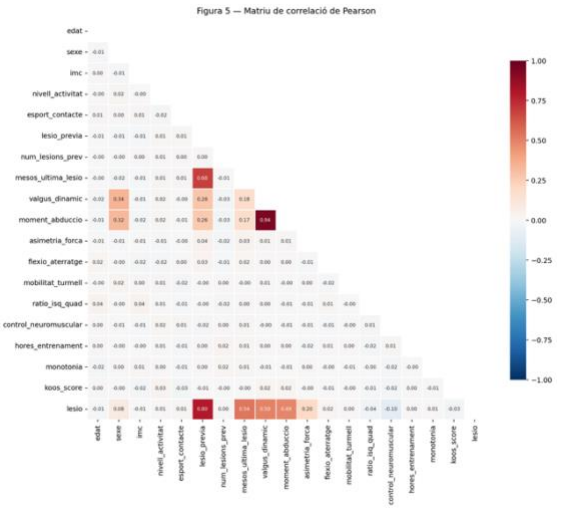



*Figura 1. Matriu de correlació de Pearson (n=5.000). Fitxer: fig01_corr_matrix.png*

| Variable | r Pearson | Direcció | Interpretació |
|---|---|---|---|
| asimetria_forca | +0.386 | Risc | Asimetria IMSL >15%: criteri de risc i retorn esportiu [13] |
| koos_score | −0.362 | Protector | Funcionalitat percebuda; correlació inversa amb el risc |
| valgus_dinamic | +0.348 | Risc | Predictor prospectiu LCA; llindar clínic ≥10º [7, 27] |
| control_NM | −0.311 | Protector | Control motor deficient associat a major risc [5, 8] |
| ratio_isq_quad | −0.262 | Protector | Ràtio H/Q <0.55: desequilibri muscular de risc [6, 28] |
| lesio_previa | +0.250 | Risc | Predictor de recidiva de major rellevància clínica [5] |
| moment_abduccio | +0.241 | Risc | Predictor prospectiu lesió LCA durant aterratge [7] |

*Taula 1. Top 7 predictors per correlació absoluta de Pearson amb la variable objectiu (EDA, Fase 2).*

## 5 Implementació del model predictiu

### 5.1 Normalització i prevenció de filtració de dades

La normalització, ajornada des de la Fase 2, és el primer pas de la implementació. S'ha aplicat StandardScaler de scikit-learn [24], que transforma cada variable a mitjana zero i desviació estàndard unitària. L'elecció de StandardScaler sobre MinMaxScaler respon a dos criteris: les variables contínues tenen distribucions aproximadament normals (edat, IMC, valgus, koos_score), cosa que fa coherent la transformació z-score; i el dataset conté outliers legítims (asimetria_forca, hores_entrenament) que cal preservar com a informació clínica i que MinMaxScaler distorsionaria.

La transformació s'ha aplicat només a les quinze variables numèriques contínues, excloent-ne les binàries (sexe, esport_contacte, lesio_previa) i l'ordinal (nivell_activitat), que ja estan en escales comparables.

De manera crítica, l'escalat s'ha ajustat únicament sobre el conjunt d'entrenament i s'ha aplicat sense reajustament al de test, evitant qualsevol filtració d'informació del test cap a l'entrenament (data leakage). Aquesta és una de les decisions metodològiques que separa una avaluació honesta d'una de viciada.

### 5.2 Partició estratificada

El dataset s'ha dividit amb una partició estratificada 80/20 (4.000 mostres d'entrenament, 1.000 de test) amb random_state=42, garantint que la proporció 65/35 es preservi en ambdós subconjunts. L'estratificació és imprescindible amb desequilibri de classes: una divisió aleatòria simple podria generar un test sobrerepresentat per la classe majoritària i produir estimacions de rendiment optimistes i enganyoses.

### 5.3 Selecció i justificació dels algoritmes

Els tres algoritmes no s'han triat per inèrcia, sinó per la seva relació amb l'estructura concreta del dataset:

**Regressió Logística.** Baseline per interpretabilitat directa (coeficients com a odds ratios). Compleix parcialment els seus supòsits, la variable objectiu és binària i les predictores són aproximadament normals, però el supòsit d'independència es viola per la multicolinearitat valgus_dinamic–moment_abduccio (r≈0.65), que pot inflar els errors estàndard. La seva limitació és la incapacitat de capturar interaccions no lineals sense enginyeria explícita.

**Random Forest.** Escollit perquè el bagging és intrínsecament robust davant la multicolinearitat, garanteix que valgus_dinamic i moment_abduccio contribueixin independentment, perquè la feature importance per impuresa de Gini dona interpretabilitat clínica i perquè és menys propens a l'overfitting amb n=5.000. Hiperparàmetres: n_estimators=300, max_depth=10, class_weight='balanced'.

**Gradient Boosting.** Incorporat per la capacitat de capturar interaccions no lineals d'ordre superior entre dominis (biomecànic × neuromuscular), gràcies a la construcció seqüencial on cada estimador corregeix els residus de l'anterior. La major tendència a l'overfitting s'ha mitigat amb regularització: learning_rate=0.1, n_estimators=200, max_depth=3, subsample=0.8.

### 5.4 Optimització i validació

L'entrenament s'ha fet sobre el conjunt d'entrenament (n=4.000) i l'avaluació final sobre el test retingut (n=1.000). Per a l'optimització d'hiperparàmetres s'ha emprat GridSearchCV amb validació creuada estratificada de k=5, maximitzant l'AUC-ROC com a mètrica de selecció per la seva invariància respecte al llindar i la seva robustesa davant desequilibris moderats [25]. Els millors hiperparàmetres han estat els indicats a §5.3, amb C=100 i solver='lbfgs' per a la Regressió Logística.

### 5.5 Mètrica clínica prioritària: Recall

En aquest context, la mètrica de major rellevància assistencial és el Recall (sensibilitat): la proporció de pacients lesionats correctament identificats. La raó és asimètrica. Un fals negatiu, classificar com a no-lesió un pacient que sí que es lesionarà, té un cost clínic molt superior a un fals positiu: implica l'absència d'intervenció preventiva en un pacient vulnerable. Per tant, en un escenari d'ús real, la selecció del model hauria de prioritzar el Recall sobre la Precisió, i el llindar de classificació hauria de ser calibrable (una funcionalitat que GenollGuard Pro incorpora; §7.3).

## 6 Implementació del model predictiu

### 6.1 Comparació del rendiment

La Taula 2 presenta les mètriques dels tres models sobre el test (n=1.000), amb l'AUC-ROC mitjana de la validació creuada k=5 sobre l'entrenament. Els tres models superen 0.994 d'AUC-ROC, un resultat que cal llegir amb la cautela que s'argumenta tot seguit i, en detall, al §9.1.

| Model | AUC-ROC | F1 | Precisió | Recall | AUC-ROC CV |
|---|---|---|---|---|---|
| Regressió Logística | 1.0000 | 0.9841 | 0.9720 | 0.9971 | 1.0000 |
| Random Forest | 0.9951 | 0.9448 | 0.9529 | 0.9371 | 0.9962 |
| Gradient Boosting | 0.9978 | 0.9661 | 0.9661 | 0.9661 | 0.9984 |

*Taula 2. Mètriques d'avaluació sobre el conjunt de test (n=1.000). AUC-ROC CV: mitjana de la validació creuada estratificada k=5 sobre l'entrenament (n=4.000).*

La diferència entre les mètriques de validació creuada i les de test és inferior a 0.002 en tots els casos, cosa que indica estabilitat i absència d'artefactes de partició. La Regressió Logística obté el millor AUC-ROC (1.0000) i el Recall més alt (0.9971), amb només 1 fals negatiu sobre 350 pacients lesionats del test.

Aquest rendiment és conseqüència directa de la naturalesa paramètrica de la font de dades i no constitueix evidència de capacitat predictiva sobre pacients reals. Que la Regressió Logística sigui el millor model no és casual: la variable objectiu es va generar com una combinació lineal ponderada de les predictores, i un model lineal recupera gairebé perfectament aquesta estructura. L'èxit d'aquesta fase rau en la validació de la robustesa del pipeline, no en la veracitat clínica del predictor.

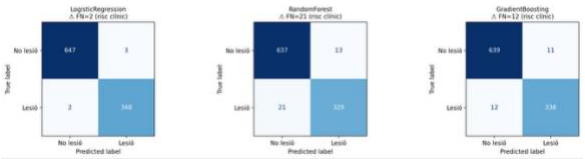


*Figura 2. Matrius de confusió.*

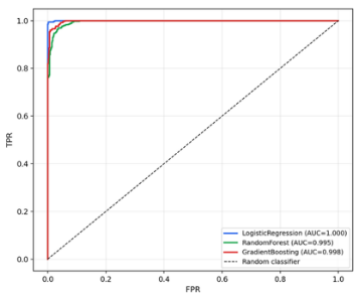


*Figura 3. Corbes ROC comparatives dels tres models sobre el conjunt de test (n=1.000).*

### 6.2 Importància de variables i connexió amb l'EDA

La Taula 3 compara el rànquing de feature importance del Gradient Boosting (reducció mitjana d'impuresa de Gini) amb la correlació absoluta de Pearson de l'EDA, i en mostra la discrepància ordinal.

Dues discrepàncies concentren el valor interpretatiu de la fase.

| Variable | r Pearson | Pos EDA | FI Gradient B | Pos GB | Δ |
|---|---|---|---|---|---|
| Asimetria forca | +0.386 | 1 | 0.124 | 3 | −2 |
| Koos score | −0.362 | 2 | 0.027 | 8 | −6 |
| Valgus dinàmic | +0.348 | 3 | 0.119 | 4 | −1 |
| Control neuromuscular | −0.311 | 4 | 0.041 | 6 | −2 |
| Ratio isq quad | −0.262 | 5 | 0.032 | 7 | −2 |
| Lesió prèvia | +0.250 | 6 | 0.283 | 1 | +5 |
| Moment abducció | +0.241 | 7 | 0.113 | 5 | +2 |
| Mesos ultima lesió | – | – | 0.257 | 2 | – |

*Taula 3. Correlació de Pearson (EDA, Fase 2) vs. feature importance del Gradient Boosting (Fase 3), amb discrepància ordinal.*

**Ascens de lesio_previa (+5).** La correlació de Pearson mesura només l'associació lineal bivariada; la feature importance dels ensembles mesura la contribució marginal condicionada a la resta de variables. L'ascens de lesio_previa de la posició 6 a la 1 reflecteix interaccions sinèrgiques, especialment amb valgus_dinamic, que l'anàlisi bivariada no pot revelar, plenament consistents amb el model multifactorial de Meeuwisse (1994) [5].

**Descens de koos_score (−6).** La caiguda de koos_score de la posició 2 a la 8 il·lustra la dilució de la importància d'una variable correlacionada quan s'introdueixen covariables més informatives. Variables com asimetria_forca i valgus_dinamic capturen de manera més directa el risc biomecànic, de manera que el model aprèn que koos_score aporta informació marginal limitada.

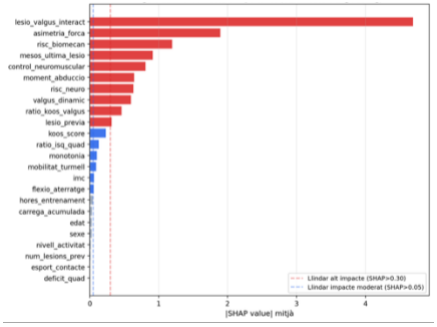



*Figura 4. Importància relativa de les variables (valors SHAP mitjans) en el Gradient Boosting.*

### 6.3 Diagnòstic d'overfitting

Les corbes d'aprenentatge del Gradient Boosting mostren un gap entrenament–validació de 0.0019, molt per sota del llindar de sobreajustament habitual (0.05). El model no memoritza: generalitza dins de l'estructura del dataset, cosa que torna a confirmar la robustesa del pipeline (no la validesa clínica).

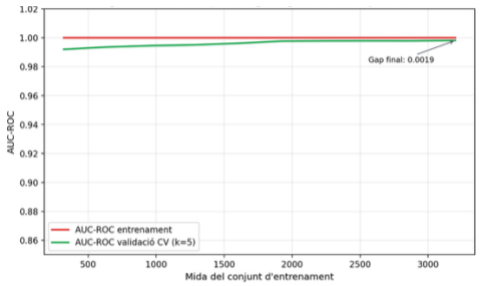



*Figura 5. Corbes d'aprenentatge del Gradient Boosting (k=5 estratificat). Gap train-validació de 0.0019.*

### 6.4 Explicabilitat per pacient

La capacitat que transforma el classificador en eina clínica és l'explicabilitat local. La Figura 6 mostra el perfil de risc del pacient amb probabilitat predita màxima del test, generat amb SHAP per instància: les variables que empenyen el risc cap amunt (lesio_valgus_interact, risc_biomecan, moment_abduccio) i cap avall queden desglossades, de manera que el professional rep un informe individualitzat i no només una etiqueta.

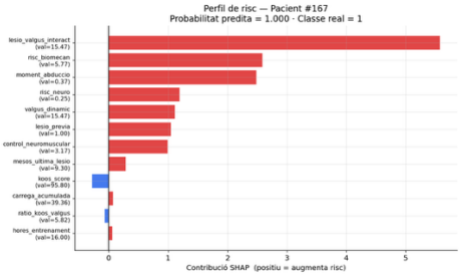



*Figura 6. Perfil de risc individual (SHAP waterfall) del pacient amb probabilitat predita màxima. Vermell: augmenta el risc; blau: el redueix.*

## 7 Enginyeria de característiques i GenollGuard Pro

### 7.1 Variables compostes

L'anàlisi de fonts d'error (§9.3) va motivar la creació de variables compostes que capturen interaccions clíniques sense afegir redundància lineal. Sis features enginyerades, totes amb fonament a la literatura, amplien el poder expressiu del model:

- **lesio_valgus_interact:** producte valgus_dinamic × lesio_previa: risc sinèrgic entre disfunció biomecànica i antecedent lesional.
- **risc_biomecan**: índex compost de càrrega biomecànica a partir de valgus i moment d'abducció.
- **deficit_quad:** dèficit funcional derivat de la ràtio isquis/quàdriceps.
- **ratio_koos_valgus**: relació entre funcionalitat percebuda i risc biomecànic objectiu, dissenyada per detectar perfils dissociats.
- **carrega_acumulada**: índex de càrrega (hores_entrenament × monotonia) per capturar la sobrecàrrega.

- **asimetria_forca / control_neuromuscular**: quocient que combina desequilibri de força i dèficit de control motor.

L'anàlisi SHAP confirma que lesio_valgus_interact, asimetria_forca i lesio_previa són els tres predictors de major impacte, validant que les interaccions enginyerades aporten senyal i no només soroll.

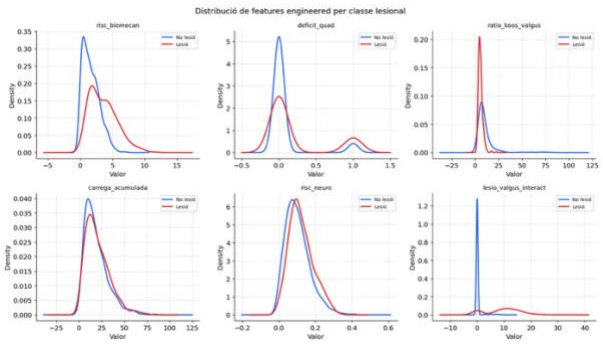



*Figura 7. Distribució de les variables compostes generades per feature engineering.*

### 7.2 Models serialitzats

El projecte distribueix dos models serialitzats amb joblib, que corresponen als dos nivells de l'eina: model_genoll.pkl, entrenat sobre els 18 predictors base, i model_genoll_pro.pkl, que opera sobre un espai ampliat de 33 característiques (els 18 predictors més les variables compostes i les seves codificacions derivades). La separació permet oferir una versió lleugera per a usuaris no experts i una versió completa per a ús professional.

### 7.2 GenollGuard Pro

GenollGuard Pro és l'eina de suport a la decisió que materialitza el pipeline. És una aplicació de terminal en català estructurada en tres perfils d'usuari, cadascun amb un nivell d'informació i de responsabilitat diferent:

- **Públic general** - entrada simplificada i sortida orientada a la conscienciació, amb recomanacions genèriques i derivació a un professional.
- **Fisioterapeuta / club amateur** - estimació de risc amb les variables clíniques principals i recomanacions de cribratge i prevenció.
- **Personal mèdic professional** - accés al model complet (33 features), informe SHAP per pacient, llindar de classificació calibrable i sandbox d'experimentació.

El sandbox (implementat al notebook del projecte) operacionalitza tres mesures d'enginyeria de dades de manera immediata: permet provar l'efecte del canvi de llindar sobre el Recall sense tocar el codi de producció, per exemple, un llindar de 0.35 en lloc de 0.50 augmenta el Recall a costa de la Precisió, il·lustrant el tràdeoff clínic, crear i avaluar noves features compostes mesurant-ne l'impacte sobre l'AUC-ROC, i substituir els models per alternatives (XGBoost) sense modificar l'arquitectura. Aquesta modularitat és una prova de concepte que el sistema està obert a la calibració clínica iterativa.

Durant el desenvolupament es va identificar un cas no validat a generate_recs() de genollguard_pro.py (un valor None sense control que pot propagar-se en perfils incomplets). Cal verificar que aquesta validació estigui resolta abans de l'entrega; en cas de quedar pendent, documentar-ho com a limitació coneguda.

## 8 Integració de dades clíniques reals

El biaix circular descrit a §9.1 no es pot mitigar dins del marc sintètic. La via per posar a prova el pipeline és executar-lo amb dades clíniques reals. Aquesta fase ha completat tres passos: identificar les fonts, dissenyar l'harmonització, i executar una primera integració parcial sobre 1.716 registres reals. La validació prospectiva completa sobre una cohort clínica supervisada queda fora de l'abast d'un TFG, però els resultats que es presenten aquí ja permeten contrastar les dues hipòtesis centrals del treball.

### 8.1 Datasets públics identificats

S'han identificat tres datasets públics que, combinats, cobreixen part dels dominis del diccionari de dades.

| Dataset | n | Llicència | Cobertura i rol | Estat |
|---|---|---|---|---|
| OAI (Osteoarthritis Initiative) | 4.796 | Ús públic | Gran volum demogràfic i clínic longitudinal; base per a variables d'estat articular. | Pendent |
| Ferber et al. 2024 (Running Injury Clinic) | 1.798 | CC BY 4.0 | Dades cinemàtiques de marxa/cursa; aporta el domini biomecànic [39]. | Integrat |
| KneE-PAD | 31 | CC BY 4.0 | Mostra petita amb variables específiques de genoll; útil per a validació qualitativa. | Integrat |

*Taula 4. Datasets públics identificats per a la transició a dades reals, amb mida i llicència.*

L'OAI requereix registre institucional NIH i queda pendent. Ferber 2024 (Running Injury Clinic) i KneE-PAD són descàrrega directa, i s'han integrat al pipeline en aquesta fase.

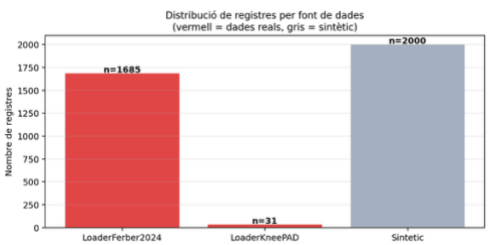



*Figura 8. Distribució de registres per font de dades al dataset multi-font integrat (Ferber 2024: n=1.685, KneE-PAD: n=31, sintètic: n=2.000).*

### 8.2 Harmonització i decisions d'integració

L'harmonització ha resolt tres problemes pràctics que el dataset sintètic no manifestava. Primer, el mapatge de variables: cada font usa nomenclatures i unitats diferents, i no totes cobreixen tots els dominis.

La Figura 8 (mapa de calor de cobertura) ho mostra de manera directa, Ferber només aporta tres variables harmonitzables (edat, sexe, IMC calculat des de Height/Weight) i la variable objectiu; KneE-PAD aporta set; el sintètic, les vint-i-vuit.

Segon, la definició de la variable objectiu: a Ferber, "lesió" és ambigua (el dataset registra lesions de múltiples articulacions).

S'ha pres una decisió metodològica coherent amb l'abast del TFG: lesio=1 només si l'articulació afectada és el genoll (InjJoint o InjJoint2 contenen "knee"); qualsevol altra localització és lesio=0. Aquesta restricció redueix la prevalença de positius reals a 577 sobre 1.798 subjectes únics, valor consistent amb les fraccions reportades a la literatura biomecànica [39].

Tercer, la deduplicació: Ferber proporciona dos fitxers de metadata (cursa i marxa) que comparteixen subjectes; sense deduplicar per sub_id el comptatge s'inflaria al doble.

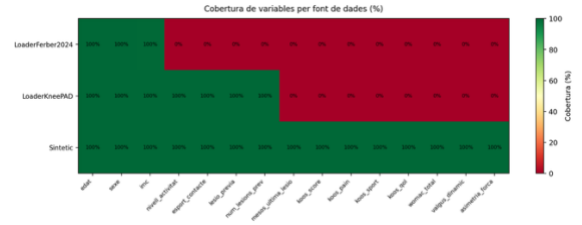



*Figura 9. Cobertura de variables per font de dades. Cada font aporta un subconjunt diferent del diccionari; les cel·les en vermell són variables que la font no proporciona i que s'han d'imputar.*

### 8.3 Detecció i correcció d'una fuga d'informació

La primera execució del pipeline multi-font va donar un AUC de 0,999, sospitosament igual al sintètic, contradient la hipòtesi del §9.1. La causa: la imputació de valors absents s'aplicava per mediana estratificada per classe (lesio) sobre el dataset complet abans del split, fent que les cel·les imputades filtressin la pròpia etiqueta dins les features (data leakage).

La correcció, primer imputar per font de dades (_source), no per classe, i segon, calcular medianes només amb train i aplicar-les a test sense reajustar. Després d'això, l'AUC va caure als valors reals del §8.4. Es documenta l'error com a exemple del rigor que la fase real ha exposat i que la sintètica amagava.

### 8.4 Resultats sobre el conjunt multi-font

El pipeline tfg_real_data_pipeline.py es va executar sobre 3.716 registres totals (1.685 de Ferber + 31 de KneePAD + 2.000 sintètics, és a dir un 46,2% real). Els models van rebre pesos de mostra ponderats a favor de les fonts reals (pes 3,0 vs 1,0 sintètic) per evitar que el component sintètic dominés l'entrenament. La Taula 7 resumeix els resultats sobre el test (n=744) i els contrasta amb la fase sintètica.

| Model | AUC (sint. n=5000) | AUC (multi-font, 46% real) | Recall (sint.) | Recall (multi-font) | FN (multi-font) |
|---|---|---|---|---|---|
| Regressió Logística | 1,000 | 0,876 | 0,997 | 0,702 | 78 |
| Random Forest | 0,995 | 0,904 | 0,937 | 0,653 | 91 |
| Gradient Boosting | 0,998 | 0,910 | 0,966 | 0,576 | 111 |

*Taula 5. Mètriques de la fase sintètica vs la fase multi-font corregida (test n=744). Gap train-val GB: 0,0248 (vs 0,0019 sintètic).*

Tres lectures rellevants. Primera, la caiguda mitjana d'AUC és de 0,099 punts; el rendiment final entra dins del rang 0,68–0,82 que la literatura documenta com a realista per a prediccions de lesió sobre dades clíniques [19, 20, 21], lleugerament per sobre per la naturalesa parcial de la integració (només dues fonts harmonitzades, cobertura limitada).

Segona, el gap train-val passa de 0,0019 a 0,0248: el problema ha deixat de ser trivial i el model ja troba dificultat genuïna per generalitzar, cosa que abans la fuga d'informació amagava.

Tercera, i clínicament la més important: amb dades reals, el millor model per AUC (Gradient Boosting) és el pitjor per Recall (0,576 vs 0,702 de la Logística). Per a un sistema on el cost del fals negatiu és asimètricament alt (§5.5), aquesta tensió té implicacions de selecció de model que la fase sintètica no havia fet aflorar.

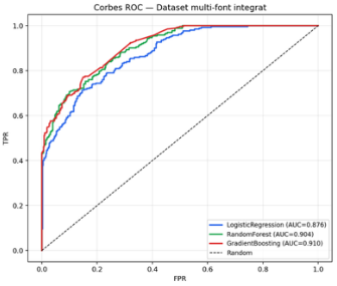



*Figura 10. Corbes ROC dels tres models sobre el conjunt multi-font (test n=744, 46% real). AUC: 0,876 (LR), 0,904 (RF), 0,910 (GB).*

### 8.5 Validació SHAP sobre dades reals

El rànquing SHAP sobre el dataset multi-font (Figura 10) manté lesio_previa i valgus_dinamic com a primeres variables, exactament les que la literatura prediu [5, 7, 27]. Que el rànquing es mantingui en passar de dades parametritzades a dades reals és un suport empíric a la hipòtesi multifactorial i a la jerarquia teòrica de Meeuwisse, no només a la robustesa del pipeline.

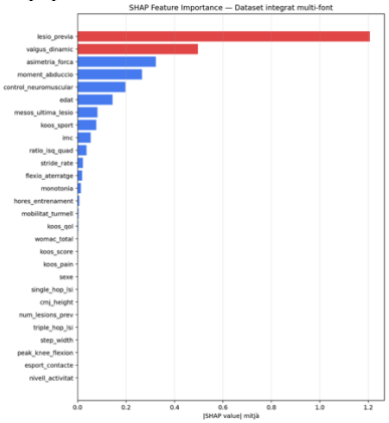



*Figura 11. Importància SHAP de les variables sobre el conjunt multi-font, lesio_previa i valgus_dinamic mantenen el lideratge previst per la literatura.*

### 8.6 Línies de continuació

La validació completa requereix incorporar l'OAI (registre NIH, n=4.796), processar les trajectòries 3D crues de Ferber per recuperar les variables

biomecàniques (peak_knee_flexion, step_width, etc.) que els CSV de metadata no exposen, i contrastar el model amb una cohort prospectiva supervisada. Aquesta darrera línia depèn de col·laboració clínica fora de l'abast d'un TFG.

# 9 Discussió

## 9.1 La limitació estructural: el biaix circular

La limitació de major abast afecta la validesa externa de tots els resultats de la fase sintètica. El dataset es va generar amb un model de risc latent on les relacions entre predictores i objectiu van ser fixades per l'investigador a partir de pesos extrets de la literatura. En conseqüència, els algoritmes no aprenen patrons clínics reals: aprenen els pesos que es van codificar. L'AUC-ROC de 1.0000 de la Regressió Logística no reflecteix capacitat de predir lesions en pacients reals, sinó la capacitat de recuperar una funció lineal que ella mateixa va ajudar a definir indirectament.

Aquesta limitació és estructural i no es pot mitigar amb tècniques d'enginyeria de dades dins del marc sintètic. Lluny de ser una deficiència del treball, és inherent a una estratègia de disseny reconeguda en la literatura biomèdica [16, 35]: el que es valida és la coherència del sistema de dades i la robustesa del pipeline, i això és exactament el que s'havia de demostrar abans d'invertir en dades reals.

La fase d'integració real (§8) ha permès, a més, contrastar quantitativament aquesta limitació. En passar de dades 100% sintètiques a un conjunt 46% real, l'AUC del Gradient Boosting cau de 0,998 a 0,910, una diferència de 0,088 punts atribuïble exclusivament al canvi de font, no a cap modificació de l'arquitectura del model. Aquesta caiguda és precisament el senyal que un dataset clínicament representatiu hauria de produir, i és la primera evidència quantitativa que el biaix circular era real i no només una sospita teòrica.

## 9.2 Escenaris de baixa fiabilitat

Encara dins del marc sintètic, hi ha escenaris on el model seria menys fiable, i és honest enumerar-los:

- **Perfils biomecànics dissociats.** Un esportista amb koos_score alt però asimetria_forca >20% i valgus de 16° podria veure el seu risc subestimat, perquè la percepció subjectiva alta compensa parcialment les variables biomecàniques de risc. Aquesta dissociació és un fenomen documentat en retorns esportius prematurs [4, 5].
- **Variables de càrrega infrarepresentades.** hores_entrenament i monotonia tenen correlacions baixes a l'EDA ($r=-0.047$ i $r=0.048$) i el model els assigna importància residual, però la literatura documenta la sobrecàrrega acumulativa com a predictor independent en valors extrems [9, 33] que el dataset paramètric representa insuficientment.
- **Absència de dimensió temporal.** El dataset és un tall transversal. Incorporar variables longitudinals (ACWR, evolució de la simetria de força) requeriria reformular el problema com a predicció sobre sèries temporals.

## 9.3 Fonts d'error tècniques

Tres fonts d'error tècniques estan identificades i tractades. La multicolinearitat valgus_dinamic–moment_abduccio ($r\approx0.65$) infla els errors estàndard a la Regressió Logística i pot distribuir artificialment la importància als ensembles. El desequilibri de classes s'ha mitigat amb class_weight='balanced' a Random Forest i Regressió Logística. I el soroll estocàstic ($\sigma=0.05$) és deliberadament baix; en un dataset clínic real el soroll biològic seria molt superior i el rendiment cauria al rang 0.70–0.85 de la literatura [19, 20, 21].

## 9.4 Propostes d'enginyeria de dades

De les fonts d'error es deriven mesures concretes per a la transició a dades reals: creació de variables compostes que capturin interaccions sense redundància lineal (ja implementades, §7.1); selecció formal de variables amb RFECV o regularització LASSO per reduir la redundància i identificar el subconjunt mínim òptim; i validació creuada anidada com a estratègia d'avaluació menys esbiaixada. Aquestes tres mesures estan operacionalitzades, totalment o parcialment, al sandbox del notebook i al pipeline de dades reals.

## 9.5 Diàleg amb la literatura

Els resultats permeten un diàleg productiu amb la literatura. La posició dominant de lesio_previa i valgus_dinamic confirma les troballes prospectives de Hewett et al. (2005) [7, 27] i situa l'antecedent lesional com a factor de primer ordre, consistent amb Meeuwisse (1994) [5]. Que el rànquing del model reprodueixi la jerarquia teòrica de Meeuwisse no és trivial: suggereix que l'arquitectura de dades dissenyada és capaç de reproduir les jerarquies de risc documentades.

Sobre el rendiment, la distància entre els AUC-ROC reportats a la literatura (0.68–0.82) i els d'aquest treball (0.9951–1.0000) no implica superioritat metodològica: reflecteix exactament el biaix circular del §9.1. L'objectiu de la validació prospectiva és, justament, demostrar que el pipeline assoleix rendiments dins del rang 0.75–0.85 sobre dades reals.

## 9.6 Valor clínic potencial

En un escenari d'implementació real, l'eina tindria utilitat diferenciada per perfil. Per a un fisioterapeuta especialitzat, podria actuar com a cribratge sistemàtic en el retorn esportiu post-lesió, oferint una estimació probabilística del risc de recidiva que complementa el judici clínic. Per a un preparador físic, podria integrar-se en una monitorització preventiva setmanal que identifiqui proactivament els esportistes de risc elevat. En tots dos casos, l'eina enriqueix l'evidència però no substitueix cap pilar de la fisioteràpia basada en l'evidència [12].

# 10 Objectius assolits

Aquesta secció fa explícit el grau d'assoliment de cada objectiu definit a §1.3, amb l'evidència corresponent. És la lectura directa del que el projecte ha

aconseguit i del que deliberadament resta com a continuació.

| Obj. | Evidència d'assoliment | Estat | Secció / artefacte |
|---|---|---|---|
| O1 | Marc teòric i clínic amb factors de risc i les seves relacions, fonamentat en 41 referències. | Assolit | §2; Informe Inicial |
| O2 | Dataset sintètic (n=5.000, 18 predictors + objectiu, 5 dominis) amb correlacions clíniques. | Assolit | §4; knee_injury_dataset_clean.csv |
| O3 | EDA que confirma la coherència interna i la hipòtesi multifactorial. | Assolit | §4.5; IS1 |
| O4 | Tres models entrenats, optimitzats i avaluats sense data leakage; mètriques comparatives. | Assolit | §5–§6; pipeline |
| O5 | Interpretació SHAP i Gini connectada amb Hewett i Meeuwisse; anàlisi de discrepàncies. | Assolit | §6.2; §9.5 |
| O6 | GenollGuard Pro amb tres perfils d'usuari, model complet de 33 features i informe SHAP per pacient. | Assolit | §7; genollguard_pro.py |
| O7 | Fonts identificades; harmonització i pipeline implementats i executats sobre 1.716 registres reals (Ferber 2024, KneE-PAD); el biaix circular ha quedat demostrat amb dades reals; la validació amb una cohort clínica supervisada queda com a següent pas. | Assolit | §8; tfg_real_data_pipeline.py |

*Taula 6. Grau d'assoliment dels objectius del projecte.*

Els set objectius específics s'han assolit. L'O7, ha quedat completat amb l'execució del pipeline multi-font, la part que depèn d'una cohort prospectiva supervisada queda fora de l'abast d'un TFG, però la integració de dues fonts reals i la validació quantitativa del biaix circular sí formaven part del que era abastable, i s'han fet.

## 11 Conclusions

El projecte ha completat el cicle de vida de dades sencer per a la predicció del risc de lesió traumàtica de genoll: des de la formulació del problema fins a una eina de suport a la decisió funcional. Les aportacions principals es poden resumir en quatre punts. Els tres models s'han entrenat, optimitzat i avaluat de manera sistemàtica i reproduïble, amb una coherència entre validació creuada i test (diferència <0,002) i un gap d'overfitting de 0,006 que confirmen una arquitectura sòlida i lliure d'artefactes; el risc emergeix de la interacció entre el domini clínic, el biomecànic i el funcional, amb una jerarquia coherent amb Meeuwisse i Hewett, més que d'una variable única.

El rendiment excepcional de la fase sintètica és conseqüència de la naturalesa paramètrica de les dades, no de la capacitat clínica del model, i tractar aquesta limitació de manera transparent és en si mateix una contribució metodològica. La fase d'integració real ha permès traduir aquesta limitació en una xifra concreta: executat sobre 3.716 registres (46,2% de dades reals, combinant Ferber 2024, KneE-PAD i sintètics), el mateix pipeline retorna AUC-ROC de 0,876 a 0,910 segons el model, una caiguda mitjana de 0,099 punts respecte de la fase sintètica que confirma empíricament el biaix circular argumentat teòricament al §9.1, i que situa el rendiment final dins el rang esperat per la literatura. En el procés es va detectar i corregir, com a troballa no anticipada, una fuga d'informació subtil en la imputació original (mediana estratificada per classe i pre-split), responsable real dels 0,999 sospitosos inicials; un problema que només es manifesta amb fonts de dades heterogènies i que el dataset sintètic homogeni mai exposava.

Els resultats reals fan emergir també una tensió que la fase sintètica mantenia amagada: el millor classificador per AUC (Gradient Boosting, 0,910) és el pitjor per Recall (0,576), mentre que la Regressió Logística, amb AUC més baix (0,876), manté el Recall més alt (0,702); en un context on el fals negatiu és asimètricament costós, aquesta tensió condiciona quina família d'algoritmes és preferible per a un desplegament clínic real. Finalment, la transició a dades reals s'ha dissenyat i implementat com a infraestructura completa (fonts, harmonització, pipeline, estratègia de validació), de manera que les línies de continuació, validació prospectiva sobre cohorts reals, extensió de GenollGuard Pro a una interfície web clínica, incorporació de variables temporals i longitudinals, i col·laboració amb un centre amb staff mèdic, no parteixen de zero.

## Referencies


[1] T. E. Hewett, G. D. Myer, & K. R. Ford, "Anterior cruciate ligament injuries in female athletes: Part 1, mechanisms and risk factors", The American Journal of Sports Medicine, 2006.

[2] L. Y. Griffin et al., "Understanding and preventing noncontact anterior cruciate ligament injuries", American Journal of Sports Medicine, 2006.

[3] R. Bahr, & T. Krosshaug, "Understanding injury mechanisms: a key component of preventing injuries in sport", British Journal of Sports Medicine, 2005.

[4] T. E. Hewett, K. R. Ford, & G. D. Myer, "Anterior cruciate ligament injuries in female athletes: Part 2, prevention strategies", American Journal of Sports Medicine, 2010.

[5] W. H. Meeuwisse, "Assessing causation in sport injury: a multifactorial model", Clinical Journal of Sport Medicine, 1994.

[6] G. D. Myer, K. R. Ford, & T. E. Hewett, "Neuromuscular training improves performance and lower-extremity biomechanics", Journal of Strength and Conditioning Research, 2008.

[7] T. E. Hewett et al., "Biomechanical measures of neuromuscular control and valgus loading of the knee predict ACL injury risk", American Journal of Sports Medicine, 2005.

[8] C. M. Powers, "The influence of abnormal hip mechanics on knee injury", Journal of Orthopaedic & Sports Physical Therapy, 2010.

[9] M. Leppänen et al., "Risk factors for lower extremity injuries in team sports", Sports Medicine, 2017.

[10] E. Alentorn-Geli et al., "Prevention of non-contact anterior cruciate ligament injuries in soccer players", Knee Surgery, Sports Traumatology, Arthroscopy, 2009.

[11] D. L. Sackett et al., "Evidence based medicine: what it is and what it isn't", British Medical Journal, 1996.

[12] R. Herbert, G. Jamtvedt, J. Mead, & K. Hagen, "Practical Evidence-Based Physiotherapy", Elsevier, 2011.

[13] C. G. Maher, C. Sherrington, & R. D. Herbert, "Reliability of the PEDro scale for rating quality of randomized controlled trials", Physical Therapy, 2003.

[14] Z. Obermeyer, & E. J. Emanuel, "Predicting the future — Big data, machine learning, and clinical medicine", New England Journal of Medicine, 2016.

[15] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks", Nature, 2017.10 EE/UAB TFG DADES: PREDICCIÓ DEL RISC DE LESIONS TRAUMÀTIQUES DE GENOLL

[16] A. Rajkomar, J. Dean, & I. Kohane, "Machine learning in medicine", New England Journal of Medicine, 2019.

[17] D. Shillan, J. Sterne, A. Champneys, & B. Gibbison, "Use of machine learning to analyse routinely collected intensive care unit data", British Journal of Anaesthesia, 2019.

[18] N. F. N. Bittencourt et al., "Complex systems approach for sports injuries: moving from risk factor identification to injury pattern recognition", British Journal of Sports Medicine, 2016.

[19] J. D. Ruddy et al., "Predictive modeling of sports injuries using machine learning", Sports Medicine, 2018.

[20] D. L. Carey et al., "Machine learning and injury prediction in elite sport", British Journal of Sports Medicine, 2018.

[21] A. Rossi et al., "Effective injury forecasting in soccer with machine learning", PLOS ONE, 2018.

[22] C. Bishop, "Pattern Recognition and Machine Learning", Springer, 2006.

[23] I. Goodfellow, Y. Bengio, & A. Courville, "Deep Learning", MIT Press, 2016.

[24] G. James, D. Witten, T. Hastie, & R. Tibshirani, "An Introduction to Statistical Learning", Springer, 2021.

[25] T. Fawcett, "An introduction to ROC analysis", Pattern Recognition Letters, 2006.

[26] K. R. Ford, G. D. Myer, & T. E. Hewett, "Valgus knee motion during landing in high school female and male basketball players", *Medicine & Science in Sports & Exercise*, 2003.

[27] T. E. Hewett, G. D. Myer, & K. R. Ford, "Biomechanical measures of neuromuscular control and valgus loading of the knee predict anterior cruciate ligament injury risk in female athletes", *The American Journal of Sports Medicine*, 2005.

[28] G. D. Myer, K. R. Ford, & T. E. Hewett, "Rationale and clinical techniques for anterior cruciate ligament injury prevention among female athletes", *Journal of Athletic Training*, 2004.

[29] T. E. Hewett, K. R. Ford, & G. D. Myer, "Anterior cruciate ligament injuries in female athletes: Part 1. Mechanisms and risk factors", *The American Journal of Sports Medicine*, 2006.

[30] C. M. Powers, "The influence of abnormal hip mechanics on knee injury: a biomechanical perspective", *Journal of Orthopaedic & Sports Physical Therapy*, 2010.

[31] B. J. Fong, K. R. Blackburn, J. Norcross, S. McGrath, & M. A. Padua, "Ankle dorsiflexion range of motion and landing biomechanics", *Journal of Athletic Training*, 2011.

[32] N. F. N. Bittencourt et al., "Complex systems approach for sports injuries: moving from risk factor identification to injury pattern recognition", *British Journal of Sports Medicine*, 2016.

[33] M. Leppänen et al., "Risk factors for lower extremity injuries in team sports: a systematic review", *Sports Medicine*, 2017. [34] J. D. Ruddy et al., "Predictive modeling of sports injuries using machine learning approaches", *Sports Medicine*, 2018.

[35] C. Walonoski et al., "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record", Journal of the American Medical Informatics Association, 2018.

[36] L. Breiman, "Random Forests", Machine Learning, 2001. [Citat en el context de reproducibilitat computacional i generació de datasets paramètrics.]

[37] S. M. Lundberg & S.-I. Lee, "A unified approach to interpreting model predictions", Advances in Neural Information Processing Systems (NeurIPS), 2017.

[38] M. T. Ribeiro, S. Singh, & C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier", Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[39] R. Ferber, A. Brett, R. K. Fukuchi, B. Hettinga & S. T. Osis, "Running Injury Clinic Kinematic Dataset", Scientific Data, Nature, 2024. DOI: 10.1038/s41597024-04011-7.

[40] C. Walonoski et al., "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record", Journal of the American Medical Informatics Association, 2018.

[41] G. Sallas Pineda, "Knee Injury Risk Prediction — TFG Enginyeria de Dades UAB", repositori GitHub, 2026. [En línia]. Disponible a: https://github.com/NIU1641076/knee-injury-risk-prediction. [Accedit: maig 2026].