

UNIVERSITAT DE BARCELONA

Treball final de grau

GRAU D'ENGINYERIA INFORMÀTICA

Facultat de Matemàtiques i Informàtica
Universitat de Barcelona

Predicció de conflictes entre laboratoris i significança clínica de variants genètiques amb aprenentatge automàtic

Autora: Èlia Garcia Rovira

Director: Dr. Daniel Ortiz Martínez
Realitzat a: Departament de Matemàtiques i Informàtica

Barcelona, 10 de juny de 2026

Agraïments

En primer lloc, voldria expressar el meu agraïment al meu tutor, Daniel Ortiz Martínez, per la seva orientació i els seus consells al llarg de tot el projecte. Li agraeixo especialment la seva predisposició per resoldre els meus dubtes, així com el seu suport constant durant el desenvolupament d'aquest treball.

També vull donar les gràcies a la meva parella, pel seu suport i el seu amor incondicional en tot moment, per ser-hi sempre i per creure en mi fins i tot quan jo no ho feia. Gràcies igualment als meus amics, per acompanyar-me i animar-me durant aquesta etapa, i a la meva mare, per tot el que m'ha donat i per fer possible que avui arribi fins aquí.

Resum

Aquest treball de fi de grau presenta una anàlisi comparativa de mètodes d'aprenentatge automàtic per a la classificació clínica de variants genètiques a ClinVar. S'han implementat i avaluat diversos models de classificació en dos problemes complementaris: la predicció binària de classificacions conflictives entre laboratoris i la predicció multiclasse de la significança clínica (benigna, VUS i patogènica). Cada model s'ha entrenat sobre dues configuracions del conjunt de dades, amb i sense anotacions funcionals VEP, amb l'objectiu d'avaluar-ne l'impacte.

Els resultats mostren que el model Random Forest obté el millor rendiment en ambdues tasques, amb un ROC AUC de 0,827 en la classificació binària i un F1-score macro de 0,900 en la multiclasse. L'anàlisi d'interpretabilitat mitjançant valors SHAP identifica les freqüències al·lèliques i l'impacte molecular (IMPACT) com els predictors més rellevants. La comparació entre configuracions revela que les anotacions VEP aporten una millora modesta en la tasca binària, però significativa en la multiclasse, sobretot en el model MLP, mentre que els models basats en arbres es mantenen estables davant la reducció de variables.

Els resultats confirmen el valor de combinar models predictius amb tècniques d'interpretabilitat per donar suport a la classificació clínica de variants, alhora que mostren que és més difícil predir el desacord entre laboratoris que la significança clínica.

Paraules clau: ClinVar, Variants Genètiques, Aprenentatge Automàtic, Classificació Clínica, Interpretabilitat, SHAP, Random Forest, Anotacions VEP.

Abstract

This thesis presents a comparative analysis of machine learning methods for the clinical classification of genetic variants in ClinVar. Several classification models were implemented and evaluated on two complementary tasks: the binary prediction of conflicting classifications between laboratories and the multiclass prediction of clinical significance (benign, VUS, and pathogenic). Each model was trained on two dataset configurations, with and without VEP functional annotations, in order to assess their impact.

The results show that the Random Forest model achieves the best performance in both tasks, with a ROC AUC of 0.827 in the binary classification and a macro F1-score of 0.900 in the multiclass setting. Interpretability analysis using SHAP values identifies allele frequencies and molecular impact (IMPACT) as the most relevant predictors. The comparison between configurations reveals that VEP annotations provide a modest improvement in the binary task, but a significant one in the multiclass setting, particularly for the MLP model, while tree-based models remain robust to feature reduction.

Overall, the results confirm the value of combining predictive models with interpretability techniques to support the clinical classification of variants, while also showing that predicting disagreements between laboratories is more challenging than predicting clinical significance.

Keywords: ClinVar, genetic variants, machine learning, clinical classification, interpretability, SHAP, Random Forest, VEP annotations.

Resumen

Este trabajo de fin de grado presenta un análisis comparativo de métodos de aprendizaje automático para la clasificación clínica de variantes genéticas en ClinVar. Se implementaron y evaluaron diversos modelos de clasificación en dos problemas complementarios: la predicción binaria de clasificaciones conflictivas entre laboratorios y la predicción multiclase de la significancia clínica (benigna, VUS y patogénica). Cada modelo se entrenó sobre dos configuraciones del conjunto de datos, con y sin anotaciones funcionales VEP, con el objetivo de evaluar su impacto.

Los resultados muestran que el modelo Random Forest obtiene el mejor rendimiento en ambas tareas, con un ROC AUC de 0,827 en la clasificación binaria y un F1-score macro de 0,900 en la multiclase. El análisis de interpretabilidad mediante valores SHAP identifica las frecuencias alélicas y el impacto molecular (IMPACT) como los predictores más relevantes. La comparación entre configuraciones revela que las anotaciones VEP aportan una mejora modesta en la tarea binaria, pero significativa en la multiclase, especialmente en el modelo MLP, mientras que los modelos basados en árboles se mantienen robustos ante la reducción de variables.

En conjunto, los resultados confirman el valor de combinar modelos predictivos con técnicas de interpretabilidad para apoyar la clasificación clínica de variantes, al tiempo que muestran que predecir el desacuerdo entre laboratorios es más difícil que predecir la significancia clínica.

Palabras clave: ClinVar, variantes genéticas, aprendizaje automático, clasificación clínica, interpretabilidad, SHAP, Random Forest, anotaciones VEP.

Índex

1 Introducció **1**
- 1.1 Contextualització del problema . . . 1
- 1.2 Plantejament del problema i motivació . . . 3
- 1.3 Objectius . . . 3
- 1.4 Planificació del projecte . . . 4
- 1.5 Estructura del document . . . 4

2 Estat de l'Art **5**
- 2.1 Variants genètiques i el genoma humà . . . 5
- 2.2 ClinVar i la classificació de variants . . . 5
- 2.3 El problema de les classificacions conflictives . . . 6
- 2.4 Anotacions funcionals VEP i eines predictives . . . 6
  - 2.4.1 Conseqüències moleculars i IMPACT . . . 6
  - 2.4.2 SIFT, PolyPhen i CADD . . . 6
  - 2.4.3 Altres scores: BLOSUM62 i LoFtool . . . 7
- 2.5 Aprenentatge automàtic aplicat a variants genètiques . . . 7
- 2.6 Models de classificació . . . 7
  - 2.6.1 Regressió Logística . . . 8
    - 2.6.1.1 Ridge L2 . . . 8
    - 2.6.1.2 Lasso L1 . . . 8
  - 2.6.2 LinearSVC . . . 9
  - 2.6.3 Decision Tree . . . 9
  - 2.6.4 Random Forest . . . 10
  - 2.6.5 MLP . . . 10
- 2.7 Interpretabilitat dels models d'aprenentatge automàtic . . . 11
  - 2.7.1 Valors de Shapley . . . 11
  - 2.7.2 Integrated Gradients . . . 12

3 Materials i Mètodes **13**
- 3.1 Descripció del dataset i anàlisi exploratòria . . . 13
  - 3.1.1 Origen i construcció del dataset . . . 13
  - 3.1.2 Variables del dataset . . . 13
  - 3.1.3 Variable objectiu CLASS . . . 14
  - 3.1.4 Variable objectiu CLNSIG . . . 14
  - 3.1.5 Desbalanceig de classes i valors nuls . . . 15
    - 3.1.5.1 Classificació binària (CLASS) . . . 15
    - 3.1.5.2 Classificació multiclasse (CLNSIG) . . . 15
    - 3.1.5.3 Valors nuls . . . 16

3.2 Partició de dades . . . 16
3.3 Preprocessament . . . 17
3.3.1 Pipeline de preprocessament . . . 17
3.3.2 Dataset Bàsic vs Dataset Anotat . . . 17
3.4 Eines i entorn de desenvolupament . . . 18
3.4.1 Llenguatge de programació . . . 18
3.4.2 Llibreries de machine learning . . . 19
3.4.2.1 scikit-learn . . . 19
3.4.2.2 SHAP . . . 19
3.4.3 Llibreries de manipulació de dades . . . 19
3.4.3.1 pandas i NumPy . . . 19
3.4.4 Visualització . . . 20
3.4.4.1 Matplotlib i Seaborn . . . 20
3.4.5 Entorn de desenvolupament . . . 20
3.4.5.1 Visual Studio Code . . . 20
3.4.6 Reproductibilitat . . . 20
3.5 Mètriques d'avaluació . . . 21
3.5.1 ROC AUC (classificació binària) . . . 21
3.5.2 F1-Score (classificació multiclasse) . . . 21
3.5.3 Matriu de confusió . . . 22
3.5.4 SHAP normalitzada . . . 23

4 Implementació **24**
4.1 Configuració general . . . 24
4.2 Cerca d'hiperparàmetres per model . . . 25
4.3 Protocol d'avaluació . . . 25
4.4 Disponibilitat del codi . . . 26

5 Resultats i Discussió **27**
5.1 Classificació Binària . . . 27
5.1.1 Dataset Anotat . . . 27
5.1.2 Dataset Bàsic . . . 30
5.1.3 Comparació Anotat vs Bàsic . . . 32
5.2 Interpretabilitat Classificació Binària . . . 33
5.2.1 Models lineals . . . 34
5.2.2 Decision Tree . . . 36
5.2.3 Random Forest . . . 38
5.2.4 MLP . . . 39
5.2.5 Concordança entre models . . . 40
5.2.6 Comparació amb el dataset Bàsic . . . 41
5.3 Classificació Multiclasse . . . 42
5.3.1 Dataset Anotat . . . 43
5.3.2 Dataset Bàsic . . . 44
5.3.3 Comparació Anotat vs Bàsic . . . 46
5.4 Interpretabilitat Classificació Multiclasse . . . 47
5.4.1 Models lineals . . . 48
5.4.2 Decision Tree . . . 49
5.4.3 Random Forest . . . 51
5.4.4 MLP . . . 52

5.4.5 Concordança entre models . . . 52
5.4.6 Comparació amb el dataset Bàsic . . . 54
5.5 Discussió General . . . 55
5.5.1 Comparació entre prediccions . . . 55
5.5.2 Comparació amb treballs previs . . . 55
5.5.3 Aportació de les anotacions funcionals VEP . . . 56
5.5.4 Implicacions clíniques . . . 56

6 Conclusions **57**
6.1 Conclusions principals . . . 57
6.2 Limitacions . . . 58
6.3 Treball futur . . . 58

A Resultats SHAP per al dataset bàsic en el problema binari **62**
A.1 Models lineals . . . 62
A.2 Decision Tree . . . 62
A.3 Random Forest . . . 63
A.4 MLP . . . 64

B Resultats SHAP per al dataset bàsic en el problema multiclasse **65**
B.1 Models lineals . . . 65
B.2 Decision Tree . . . 66
B.3 Random Forest . . . 67
B.4 MLP . . . 68

Índex de figures

1.1 Representació del problema de classificació concordant (CLASS=0) i conflictiva (CLASS=1) entre laboratoris. Font: [3] 2

2.1 Valors i prediccions qualitatives de SIFT. Font: Ensembl [11]. 6
2.2 Valors i prediccions qualitatives de PolyPhen-2. Font: Ensembl [11]. . 7
2.3 Arquitectura d'un arbre de decisió. 9
2.4 Arquitectura d'un bosc aleatori. 10
2.5 Arquitectura d'una xarxa neuronal multicapa amb una capa oculta. . 11

3.1 Distribució de la variable objectiu CLASS. 15
3.2 Distribució de la variable objectiu CLNSIG. 16
3.3 Top 20 columnes amb més valors nuls. 16
3.4 Estructura d'una matriu de confusió binària. 22

4.1 Protocol d'avaluació aplicat a cada model i configuració. 26

5.1 Corbes ROC dels sis models al conjunt de test (dataset anotat, tasca binària). 28
5.2 Matriu de confusió del Random Forest (dataset anotat, tasca binària). 29
5.3 Matriu de confusió de l'Arbre de Decisió (dataset anotat, tasca binària). 29
5.4 Corbes ROC dels sis models al conjunt de test (dataset bàsic, tasca binària). 30
5.5 Matriu de confusió del Random Forest (dataset bàsic, tasca binària). 31
5.6 Matriu de confusió del Decision Tree (dataset bàsic, tasca binària). . 32
5.7 Comparació del Test ROC AUC entre el dataset Anotat i el dataset Bàsic per a cada model (tasca binària). 33
5.8 Valors SHAP per al model LinearSVC. 34
5.9 Valors SHAP per al model de regressió logística L2 (Ridge). 35
5.10 Valors SHAP per al model de regressió logística L1 (Lasso). 35
5.11 Estructura dels tres primers nivells de l'Arbre de Decisió (tasca binària, dataset anotat). 37
5.12 Valors SHAP signats del Random Forest (tasca binària, dataset anotat). 38
5.13 Valors SHAP signats del MLP (tasca binària, dataset anotat). 39
5.14 Importància SHAP normalitzada per model: top 15 variables més consistents entre els sis classificadors (tasca binària, dataset anotat). . 40
5.15 Importància SHAP normalitzada per model: top 15 variables més consistents entre els sis classificadors de tasca binària en dataset bàsic. 42
5.16 Matriu de confusió del Random Forest (dataset anotat, tasca multiclasse). 43
5.17 Matriu de confusió del MLP (dataset anotat, tasca multiclasse). . . . 44

5.18 Matriu de confusió del Random Forest (dataset bàsic, tasca multiclasse). 45
5.19 Matriu de confusió del MLP (dataset bàsic, tasca multiclasse). 46
5.20 Comparació del Test F1 macro entre el dataset Anotat i el dataset Bàsic per a cada model (tasca multiclasse). 47
5.21 Valors SHAP signats del LinearSVC per classe (tasca multiclasse, dataset anotat). 48
5.22 Valors SHAP signats de la Logística L2 Ridge per classe (tasca multiclasse, dataset anotat). 48
5.23 Valors SHAP signats de la Logística L1 Lasso per classe (tasca multiclasse, dataset anotat). 49
5.24 Estructura dels tres primers nivells del Decision Tree (tasca multiclasse, dataset anotat). 50
5.25 Valors SHAP signats del Random Forest per classe (tasca multiclasse, dataset anotat). 51
5.26 Valors SHAP signats del MLP per classe (tasca multiclasse, dataset anotat). 52
5.27 Importància SHAP normalitzada per model: top 15 variables més consistents entre els sis classificadors (tasca multiclasse, dataset anotat). 53
5.28 Importància SHAP normalitzada per model: top 15 variables més consistents entre els sis classificadors (tasca multiclasse, dataset bàsic). 54

A.1 Valors SHAP signats dels tres models lineals (dataset bàsic). 62
A.2 Valors SHAP signats del Decision Tree (dataset bàsic). 63
A.3 Estructura dels tres primers nivells del Decision Tree (dataset bàsic). 63
A.4 Valors SHAP signats del Random Forest (dataset bàsic). 64
A.5 Valors SHAP signats del MLP (dataset bàsic). 64

B.1 Valors SHAP signats del LinearSVC per classe (dataset bàsic, tasca multiclasse). 65
B.2 Valors SHAP signats de la Logística L2 Ridge per classe (dataset bàsic, tasca multiclasse). 66
B.3 Valors SHAP signats de la Logística L1 Lasso per classe (dataset bàsic, tasca multiclasse). 66
B.4 Valors SHAP signats del Decision Tree per classe (dataset bàsic, tasca multiclasse). 67
B.5 Estructura dels tres primers nivells del Decision Tree (dataset bàsic, tasca multiclasse). 67
B.6 Valors SHAP signats del Random Forest per classe (dataset bàsic, tasca multiclasse). 68
B.7 Valors SHAP signats del MLP per classe (dataset bàsic, tasca multiclasse). 68

Índex de taules

1.1 Planificació del treball 4

3.1 Distribució de les variables per tipus 14
3.2 Dimensions dels datasets per a cada tasca i configuració. 18
3.3 Dimensions dels datasets després del preprocessament i la partició train/test. 18
3.4 Versions de les principals eines i llibreries utilitzades. 21

4.1 Espais de cerca d'hiperparàmetres per a cada model. 25

5.1 Resultats de la classificació binària sobre el dataset anotat. 27
5.2 Resultats de la classificació binària sobre el dataset bàsic. 30
5.3 Resultats de la classificació multiclasse sobre el dataset anotat. 43
5.4 Resultats de la classificació multiclasse sobre el dataset bàsic. 45

Capítol 1

Introducció

Aquest capítol presenta el context del treball, la motivació que l'ha impulsat i els objectius que es persegueixen. S'exposa el problema de la classificació de variants genètiques en bases de dades clíniques, es justifica l'aplicació de tècniques d'aprenentatge automàtic per abordar-lo, i es descriu l'estructura del document.

1.1 Contextualització del problema

El genoma humà conté milions de variants genètiques, és a dir, posicions on la seqüència d'ADN d'una persona no coincideix amb la seqüència de referència [20]. La majoria d'aquestes variants no tenen cap efecte sobre la salut, però algunes poden estar associades a malalties hereditàries. Determinar si una variant és perjudicial o no és una feina crítica en la medicina clínica moderna, ja que d'aquesta interpretació en poden dependre decisions diagnòstiques i terapèutiques per al pacient.

ClinVar és una base de dades pública que recull interpretacions de variants genètiques enviades per laboratoris clínics de tot el món [19, 27]. Cada laboratori classifica les variants en una escala categòrica: benigna, probablement benigna, de significança incerta (VUS, de l'anglès *Variant of Uncertain Significance*), probablement patogènica i patogènica.

Aquest sistema, però, presenta un problema, laboratoris diferents poden arribar a conclusions diferents per a la mateixa variant. Quan una variant rep alhora classificacions de categories oposades, com ara benigna i patogènica, es parla d'una classificació conflictiva. La Figura 1.1 il·lustra la diferència entre una classificació concordant (CLASS=0), on tots els laboratoris coincideixen, i una de conflictiva (CLASS=1), on arriben a conclusions oposades.

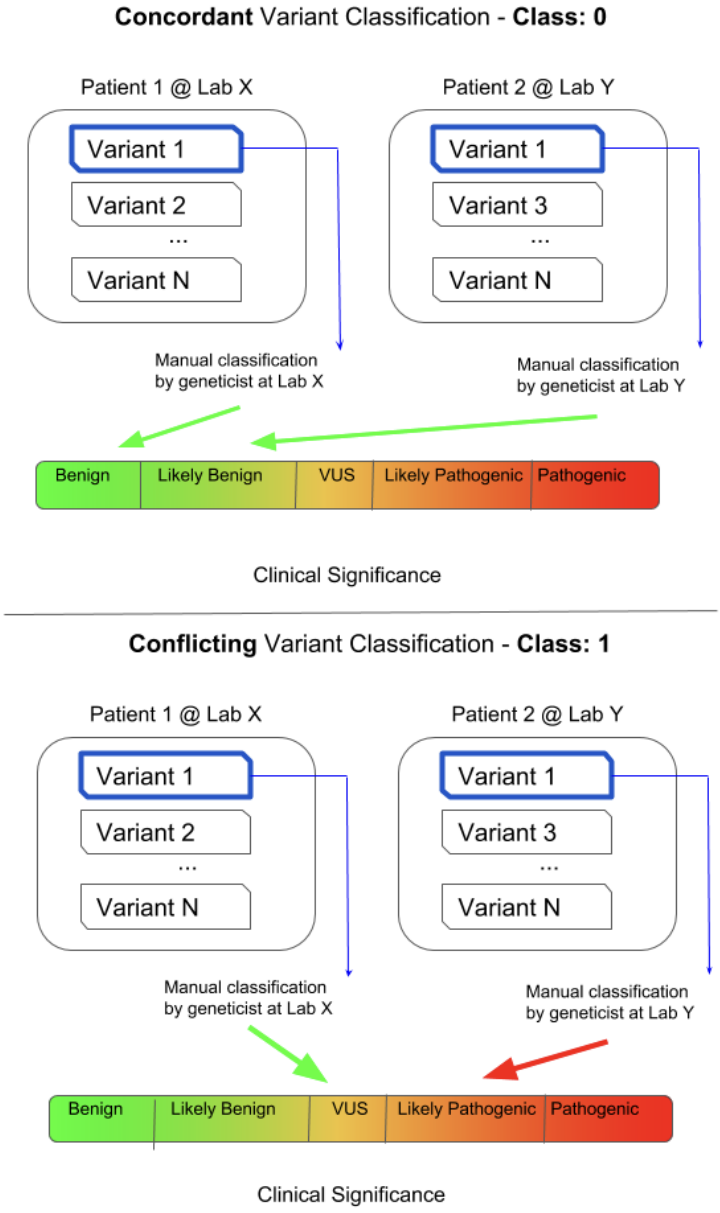


Figura 1.1: Representació del problema de classificació concordant (CLASS=0) i conflictiva (CLASS=1) entre laboratoris. Font: [3]

Aquestes diferències generen incertesa clínica i dificulten la presa de decisions mèdiques. En aquest context, es plantegen dues tasques complementàries de classificació amb dades de ClinVar. D'una banda, es planteja un problema de classificació binària per predir si una variant tindrà classificacions conflictives entre laboratoris. D'altra banda, es planteja un problema de classificació multiclasse per predir la significança clínica d'una variant, distingint entre les categories benigna, VUS i patogènica. En tots dos casos, s'avalua si les anotacions funcionals generades per eines bioinformàtiques externes, com ara predictors d'impacte funcional o puntuacions d'efecte perjudicial aporten valor predictiu addicional respecte a les dades clíniques bàsiques disponibles directament a ClinVar, com ara la freqüència al·lèlica o el tipus de variant.

1.2 Plantejament del problema i motivació

La presència de classificacions conflictives en ClinVar no és un fenomen aïllat. Quan un laboratori classifica una variant com a benigna i un altre com a patogènica, el metge que atén el pacient es troba davant d'una informació contradictòria sense eines clares per resoldre la discrepància. En casos on la decisió pot implicar un tractament, una cirurgia preventiva o un consell genètic familiar, aquesta variabilitat en la interpretació té un impacte real i directe sobre les persones.

Des d'una perspectiva d'aprenentatge automàtic, aquest problema presenta característiques interessants com un dataset gran amb variables clíniques i funcionals, desbalanceig de classes i la necessitat de models interpretables que permetin entendre quins factors contribueixen al conflicte.

A més, aquest treball també incorpora la tasca de predir la significança clínica d'una variant, distingint entre les categories benigna, VUS i patogènica. Aquesta segona predicció complementa la primera i permet avaluar si els mateixos models i característiques que detecten conflictes també són capaços de classificar correctament l'impacte clínic d'una variant. Les dues tasques ofereixen una visió més completa de la capacitat de l'aprenentatge automàtic per donar suport a la interpretació clínica de variants genètiques.

A escala personal, sempre m'ha interessat l'àmbit biològic i mèdic. La possibilitat de combinar eines informàtiques amb dades genòmiques per afrontar un problema amb impacte clínic real va ser el que em va motivar a escollir aquest tema. La combinació entre bioinformàtica i aprenentatge automàtic representa un dels camps amb més projecció i rellevància social de l'enginyeria actual.

1.3 Objectius

L'objectiu principal d'aquest treball és explorar la capacitat de diferents models d'aprenentatge automàtic per classificar variants genètiques a partir de dades de ClinVar. Aquest objectiu general es concreta en els següents objectius específics:

1. Construir i avaluar models de classificació binària per predir si una variant genètica presentarà classificacions conflictives entre laboratoris.

2. Construir i avaluar models de classificació multiclasse per predir la significança clínica d'una variant, distingint entre les categories benigna, VUS o patogènica.

3. Comparar el rendiment dels diferents models d'aprenentatge automàtic.

4. Avaluar l'impacte de les anotacions funcionals de VEP comparant resultats obtinguts amb el dataset complet respecte a un dataset sense aquestes anotacions.

5. Analitzar la interpretabilitat dels models mitjançant valors SHAP, identificant quines variables tenen més pes en les prediccions i si són consistents entre models.

1.4 Planificació del projecte

El treball s'ha desenvolupat al llarg de quatre mesos, des de febrer fins a juny de 2026, combinant sessions de treball autònom amb reunions periòdiques de seguiment amb el tutor. La Taula 1.1 resumeix les principals fases i la seva duració.

Fase	**Període**	**Tasques principals**
Exploració i preprocessament	2 setmanes	Anàlisi del dataset, pipeline de preprocessament i partició de dades
Classificació binària	4 setmanes	Entrenar models i obtenir resultats
Classificació multiclasse	3 setmanes	Entrenar models i obtenir resultats
Interpretabilitat	3 setmanes	SHAP per tots els models, anàlisi de features importants
Anàlisi Comparativa	1 setmana	Comparació models i impacte de les anotacions funcionals
Escriptura	3 setmanes	Redacció de la memòria

Taula 1.1: Planificació del treball

1.5 Estructura del document

Aquest document s'organitza en sis capítols.

- **Capítol 2 (Estat de l'Art):** Presenta els conceptes biològics fonamentals sobre variants genètiques i revisa els principals mètodes d'aprenentatge automàtic aplicats a la classificació de variants genètiques en entorns clínics.
- **Capítol 3 (Materials i Mètodes):** Descriu el dataset utilitzat, l'anàlisi exploratòria de les dades i el pipeline de preprocessament aplicat, això com les mètriques fetes servir.
- **Capítol 4 (Implementació):** Defineix la configuració experimental, incloent els models de classificació avaluats, la cerca d'hiperparàmetres i el protocol d'entrenament.
- **Capítol 5 (Resultats i Discussió):** Presenta els resultats obtinguts en les dues tasques de classificació, binària i multiclasse. Per a cada tasca es comparen els rendiments amb el conjunt de dades anotat i bàsic, i s'inclou l'anàlisi d'interpretabilitat amb valors SHAP. El capítol es tanca amb una discussió general que contextualitza els resultats i discuteix les implicacions clíniques.
- **Capítol 6 (Conclusions):** Resumeix les conclusions principals, les limitacions del treball i les línies de treball futur.

Capítol 2

Estat de l'Art

2.1 Variants genètiques i el genoma humà

El genoma humà conté aproximadament 3.000 milions de parells de bases [18], dels quals el 99.9% és idèntic entre individus. El 0.1% restant conté variacions anomenades variants genètiques, que poden contribuir a la predisposició a malalties hereditàries [1].

La interpretació clínica d'una variant segueix les directrius de l'American College of Medical Genetics and Genomics (ACMG), que defineix cinc categories: patogènica, probablement patogènica, de significança incerta (VUS), probablement benigna i benigna [32].

2.2 ClinVar i la classificació de variants

ClinVar és una base de dades pública mantinguda pel National Center for Biotechnology Information (NCBI) que recull interpretacions de variants genètiques enviades voluntàriament per laboratoris clínics, centres de recerca i hospitals de tot el món [27]. Conté milions de variants anotades i és una de les principals fonts de referència per a la interpretació clínica de variants genètiques.

Cada entrada a ClinVar associa una variant genètica amb una classificació clínica que descriu el seu efecte sobre la salut del portador, seguint les categories definides per l'ACMG:

- **Patogènica**: variant amb prou evidència per causar la malaltia.
- **Probablement patogènica**: variant amb forta sospita d'efecte patogènic, però amb evidència no concloent.
- **Significança incerta (VUS)**: variant amb evidència contradictòria o insuficient per assignar-la a una categoria definida.
- **Probablement benigna**: variant amb evidència parcial d'absència d'efecte clínic.
- **Benigna**: variant amb evidència suficient d'absència d'efecte patogènic.

2.3 El problema de les classificacions conflictives

Es considera que una variant té una classificació conflictiva quan ha rebut classificacions de dues o més de les tres grans categories (benigna, VUS i patogènica) per part de laboratoris diferents [19]. Aquestes diferències generen manca de consens clínic, ja que un mateix resultat genètic pot ser interpretat de manera oposada per laboratoris diferents, dificultant la presa de decisions.

2.4 Anotacions funcionals VEP i eines predictives

Les anotacions funcionals proporcionen informació predictiva sobre l'efecte biològic d'una variant i són àmpliament utilitzades com a *features* en models d'aprenentatge automàtic per a la classificació de patogenicitat. Ensembl VEP (Variant Effect Predictor) és una eina bioinformàtica desenvolupada per Ensembl que permet anotar variants genètiques amb informació funcional derivada de bases de dades i algoritmes externs [9]. Donat un fitxer de variants en format VCF, VEP identifica tots els transcrits afectats i prediu les conseqüències moleculars de cada variant sobre cada transcrit.

2.4.1 Conseqüències moleculars i IMPACT

VEP classifica cada variant segons el tipus de conseqüència molecular que produeix sobre el transcrit, seguint la terminologia de la Sequence Ontology (SO). Exemples de conseqüències són `missense_variant`, `synonymous_variant` o `frameshift_variant`. A partir d'aquestes conseqüències, VEP assigna un nivell d'impacte (IMPACT) que pot ser HIGH, MODERATE, LOW o MODIFIER, en ordre decreixent de severitat esperada [10]. Per exemple, un stop_gained es classifica com a HIGH, un missense_variant com a MODERATE i un synonymous_variant com a LOW.

2.4.2 SIFT, PolyPhen i CADD

- **SIFT (Sorting Intolerant From Tolerant)** prediu si una substitució aminoacídica és perjudicial per a la funció de la proteïna a partir de la conservació evolutiva de la posició [28]. Assigna una puntuació entre 0 i 1, on valors inferiors a 0,05 es classifiquen com a *deleterious* i la resta com a *tolerated* [11].

SIFT value	Qualitative prediction	Website display example
Less than 0.05	"Deleterious"	0.01
	"Deleterious - low confidence"	0.01
Greater than or equal to 0.05	"Tolerated"	0.8
	"Tolerated - low confidence"	0.8

Figura 2.1: Valors i prediccions qualitatives de SIFT. Font: Ensembl [11].

- **PolyPhen-2 (Polymorphism Phenotyping v2)** prediu l'efecte d'una substitució aminoacídica sobre l'estructura i funció de la proteïna combinant informació evolutiva i estructural [2]. La seva puntuació entre 0 i 1, representa

la probabilitat que el canvi sigui perjudicial, i les prediccions qualitatives són: *benign*, *possibly damaging* i *probably damaging* [11].

Polyphen value	Qualitative prediction	Website display example
greater than 0.908	"Probably Damaging"	0.95
greater than 0.446 and less than or equal to 0.908	"Possibly Damaging"	0.5
less than or equal to 0.446	"Benign"	0.25
unknown	"Unknown"	unknown

Figura 2.2: Valors i prediccions qualitatives de PolyPhen-2. Font: Ensembl [11].

- **CADD (Combined Annotation Dependent Depletion)** és una puntuació de perjudicialitat que integra múltiples anotacions funcionals i de conservació en una mètrica [31]. Es presenta en dues formes, CADD_RAW, que és el valor continu original, i CADD_PHRED, que és una transformació tipus Phred per facilitar la interpretació. Variants amb CADD_PHRED superior a 30 es consideren entre el 0,01% de variants més nocives del genoma humà [11].

2.4.3 Altres scores: BLOSUM62 i LoFtool

BLOSUM62 és una matriu de substitució d'aminoàcids molt usada en bioinformàtica que mesura la compatibilitat evolutiva entre dos aminoàcids [16]. Valors positius indiquen substitucions freqüents entre proteïnes homòlogues, i per tant probablement tolerades, mentre que valors negatius indiquen substitucions rares evolutivament i potencialment perjudicials.

LoFtool és un score que quantifica la intolerància d'un gen a variants amb pèrdua de funció (*loss-of-function*), calculat a partir de la proporció observada vs. esperada d'aquest tipus de variants en poblacions sanes [12]. Valors baixos indiquen gens més sensibles a aquest tipus de variants i, per tant, més probablement implicats en malalties.

2.5 Aprenentatge automàtic aplicat a variants genètiques

Diversos treballs han aplicat tècniques d'aprenentatge automàtic a la classificació de variants genètiques usant ClinVar com a font principal de dades. Nicora et al. [29] van combinar criteris ACMG amb anotacions funcionals en un model de regressió logística, assolint una precisió del 98%, tot i que assenyalen que la distribució de classes a ClinVar no reflecteix la realitat clínica. El treball més proper al present és el de Ghaedi et al. [14], que van predir específicament si una variant tindria classificacions conflictives entre laboratoris, obtenint un *accuracy* del 88% amb XGBoost.

2.6 Models de classificació

L'aprenentatge automàtic supervisat engloba un conjunt de tècniques que permeten construir models capaços de predir una variable objectiu a partir d'un conjunt de

característiques observades. En problemes de classificació, l'objectiu és assignar cada mostra a una de les categories predefinides, aprenent patrons que les diferencien a partir de dades d'entrenament.

Els models de classificació es poden agrupar entre models lineals, que assumeixen que la frontera de decisió entre classes és un hiperplà en l'espai de característiques,i els models no lineals, que són capaços de capturar relacions més complexes entre variables. En general, els models lineals tendeixen a ser més interpretables i menys propensos al sobreajustament, mentre que els no lineals poden assolir un rendiment superior en problemes amb relacions complexes, a costa d'una dificultat més elevada d'interpretació [25].

En aquest treball s'han avaluat sis models de classificació, combinant models lineals i no lineals, per tenir una visió general del comportament de les dades.

2.6.1 Regressió Logística

La regressió logística és un model lineal de classificació que estima la probabilitat de pertinença a una classe. A diferència de la regressió lineal clàssica, utilitza una funció sigmoide aplicada sobre una combinació lineal de les característiques per garantir que les prediccions es mantinguin sempre en un interval acotat entre 0 i 1. Tot i la seva utilitat, quan aquest model s'entrena amb un nombre molt gran de variables predictores, es torna massa flexible i tendeix a patir de sobreajustament, memoritzant el soroll de les dades en lloc dels patrons generals. Per solucionar-ho i millorar el rendiment davant de dades noves, s'introdueixen tècniques de regularització [7].

2.6.1.1 Ridge L2

La regularització Ridge aborda el sobreajustament afegint un terme de penalització a la funció de pèrdua basat en la suma dels quadrats dels coeficients. Aquesta tècnica és especialment útil per atenuar la inestabilitat quan hi ha variables altament correlacionades. La funció objectiu a minimitzar és:

$$\min_{w} \sum_{i=1}^{n} \mathcal{L}(y_i, \hat{y}_i) + \alpha \|w\|_2^2 \tag{2.1}$$

on α controla la intensitat de la regularització, w és el vector de coeficients i $\|w\|_2^2 = \sum_j w_j^2$ és la norma L2. Un valor d'α elevat genera coeficients més petits. L'efecte pràctic d'aquesta penalització és que redueix la magnitud de tots els coeficients de manera proporcional, apropant-los cap a zero, però sense arribar mai a eliminar-los completament [34].

2.6.1.2 Lasso L1

La regularització Lasso substitueix la penalització quadràtica per la norma L1, la qual es basa en la suma del valor absolut dels coeficients. La funció objectiu és:

$$\min_{w} \sum_{i=1}^{n} \mathcal{L}(y_i, \hat{y}_i) + \alpha \|w\|_1 \tag{2.2}$$

La gran diferència teòrica i pràctica de Lasso respecte a Ridge és la seva capacitat per simplificar l'equació. La geometria de la norma L1 força que els coeficients de

les variables amb menys poder predictiu es redueixin exactament a zero. Per tant, a diferència de Ridge, Lasso actua simultàniament com a mètode de regularització i com a mecanisme de selecció automàtica de característiques [34].

2.6.2 LinearSVC

La Màquina de Vectors de Suport Lineal (*Linear Support Vector Classifier*) és un model de classificació de dades linealment separables [8]. L'objectiu és trobar l'hiperplà òptim que maximitzi el marge de separació entre les diferents classes. L'equació d'aquest hiperplà és:

$$w \cdot x + b = 0 \tag{2.3}$$

La frontera de decisió es defineix pels vectors de suport, que són els punts de dades més propers a l'hiperplà. Maximitzar la distància entre aquests punts i l'hiperplà millora la capacitat de generalització del model, ja que un marge més gran redueix la sensibilitat a petites variacions en les dades d'entrada.

2.6.3 Decision Tree

L'arbre de decisió classifica les mostres mitjançant una seqüència de condicions binàries sobre les característiques, construint una estructura jeràrquica en forma d'arbre [6]. Cada node intern representa una condició sobre una característica, cada branca representa el resultat d'aquesta condició, i cada fulla representa una classe predita, tal com es mostra en la Figura 2.3.

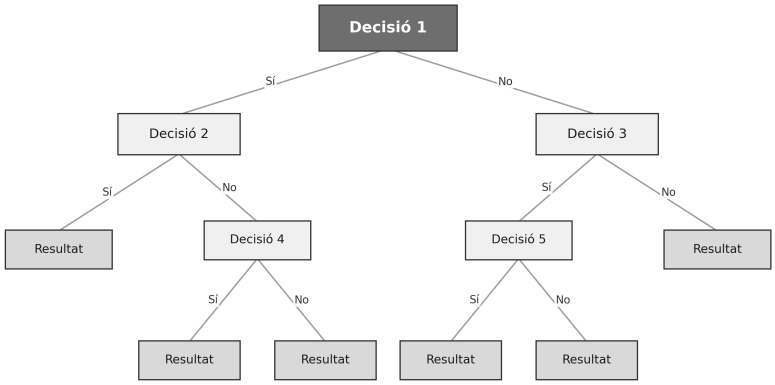


Figura 2.3: Arquitectura d'un arbre de decisió.

L'arbre s'entrena escollint en cada node la característica i el llindar que millor separa les classes. Els criteris de divisió més habituals són l'índex de Gini i l'entropia, que mesuren la puresa de les classes resultants en cada node. La profunditat màxima de l'arbre és un dels hiperparàmetres principals: arbres massa profunds tendeixen al sobreajustament, mentre que arbres massa superficials poden no capturar prou els patrons de les dades. El principal avantatge d'aquest model és la seva fàcil interpretabilitat, ja que es pot seguir el camí des de l'arrel fins a una fulla per entendre per què el model ha pres una decisió concreta.

2.6.4 Random Forest

El bosc aleatori és un mètode d'aprenentatge per conjunts que combina múltiples arbres de decisió per obtenir prediccions més precises [5]. Cada arbre s'entrena sobre una mostra aleatòria amb reemplaçament del conjunt d'entrenament (*bootstrapping*), i en cada divisió de node només es considera un subconjunt aleatori de característiques. Aquesta doble aleatorietat redueix la correlació entre arbres, cosa que disminueix la variància de les prediccions sense incrementar gaire el biaix, seguint el principi del *bagging*.

Per a tasques de classificació, la predicció final s'obté per vot majoritari entre tots els arbres, tal com s'il·lustra a la Figura 2.4.

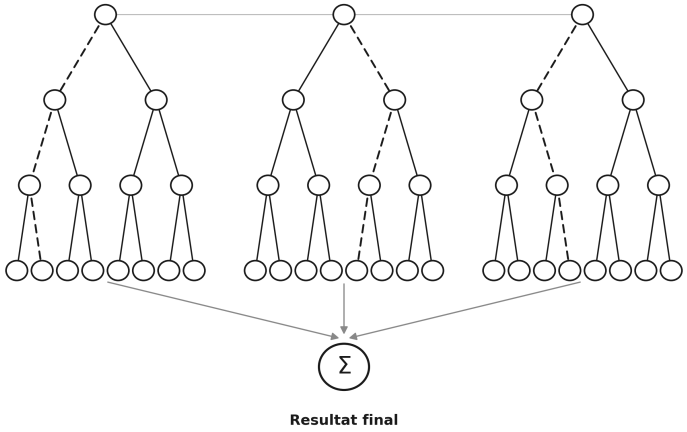


Figura 2.4: Arquitectura d'un bosc aleatori.

2.6.5 MLP

Una xarxa neuronal multicapa (*Multilayer Perceptron*) és un model no lineal format per capes de neurones interconnectades: una capa d'entrada, una o més capes ocultes i una capa de sortida, tal com es mostra a la Figura 2.5. Cada neurona calcula una combinació lineal de les seves entrades seguida d'una funció d'activació no lineal:

$$y = f\left(\sum_j w_j x_j + b\right) \tag{2.4}$$

on x_j són les entrades, w_j els pesos, b el biaix i f la funció d'activació. En aquest treball s'utilitza la funció ReLU (*Rectified Linear Unit*) [26], definida com:

$$f(x) = \max(0, x) \tag{2.5}$$

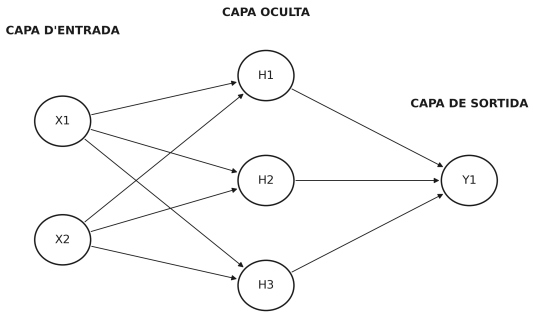


Figura 2.5: Arquitectura d'una xarxa neuronal multicapa amb una capa oculta.

L'entrenament consisteix a ajustar els pesos de la xarxa per minimitzar la funció de pèrdua mitjançant l'algorisme de retropropagació (*backpropagation*) [33]. Per evitar el sobreajustament, s'aplica la tècnica d'*early stopping*, que atura l'entrenament quan el rendiment en el conjunt de validació deixa de millorar durant un nombre determinat d'iteracions.

Per a la classificació binària, la capa de sortida conté una única neurona amb funció sigmoide, que retorna la probabilitat de pertinença a la classe positiva. Per a la classificació multiclasse, conté tantes neurones com classes (tres en el nostre cas: benigna, VUS i patogènica) amb funció *softmax*, que normalitza les sortides en una distribució de probabilitat sobre totes les classes.

2.7 Interpretabilitat dels models d'aprenentatge automàtic

A mesura que els models d'aprenentatge automàtic guanyen complexitat, esdevenen sovint caixes negres difícils d'interpretar. La interpretabilitat fa referència a la capacitat d'entendre i explicar per què un model pren una determinada decisió. En contextos clínics com el d'aquest treball, aquesta capacitat és tan important com el rendiment predictiu, ja que permet identificar quines variables biològiques o clíniques motiven cada predicció [25].

Els mètodes d'interpretabilitat es poden classificar en dues grans categories. Els mètodes intrínsecs són propis de models inherentment interpretables, com els coeficients de la regressió logística o l'estructura d'un arbre de decisió, i els mètodes agnòstics s'apliquen a qualsevol tipus de model un cop entrenat, independentment de la seva arquitectura interna.

2.7.1 Valors de Shapley

Els valors de Shapley provenen de la teoria de jocs cooperatius i permeten distribuir de manera justa la contribució de cada característica a una predicció concreta [25].

La idea central és calcular la contribució marginal de cada variable considerant totes les possibles combinacions amb les altres característiques. Per una variable i, el valor de Shapley es defineix com:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \left[v(S \cup \{i\}) - v(S) \right] \tag{2.6}$$

on N és el conjunt de totes les variables, S és un subconjunt de variables sense incloure la i-èsima, i $v(S)$ és la predicció del model utilitzant només les variables del subconjunt S.

Els valors de Shapley garanteixen tres propietats fonamentals que els fan especialment adequats per a la interpretacio de models [25]:

- **Eficiència**: les contribucions de totes les variables sumen la diferència entre la predicció del model i el valor base.
- **Simetria**: dues variables amb la mateixa contribució reben el mateix valor SHAP.
- **Absència** (*dummy*): una variable que no influeix en la predicció rep un valor SHAP de zero.

El principal inconvenient teòric dels valors de Shapley és el seu cost computacional, calcular-los exactament requereix avaluar el model en totes les combinacions possibles de variables, cosa que és impracticable per a conjunts de dades amb desenes o centenars de característiques. Per superar aquesta limitació, s'han desenvolupat aproximacions eficients i implementacions especialitzades segons el mena de model, com la llibreria SHAP [23], que s'utilitza en aquest treball i es detalla al Capítol 3.

2.7.2 Integrated Gradients

Els Integrated Gradients són un mètode d'interpretabilitat per a models diferenciables, com les xarxes neuronals, proposat per Sundararajan et al. [35]. El mètode calcula l'atribució de cada característica integrant els gradients del model al llarg d'un camí lineal entre una entrada de referència (*baseline*) i l'entrada real.

$$IG_i(x) = (x_i - x_i') \cdot \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i}\, d\alpha \tag{2.7}$$

on x' és el *baseline*, x és l'entrada real i F és la funció del model. Una propietat important és la completesa, que és la suma de totes les atribucions és igual a la diferència entre la predicció real i la del *baseline* [35].

Tot i que en aquest treball s'utilitza SHAP com a mètode principal d'interpretabilitat per la seva versatilitat i compatibilitat amb tots els models avaluats, els Integrated Gradients representen una alternativa sòlida per a models diferenciables i podrien explorar-se en treballs futurs centrats exclusivament en el MLP.

Capítol 3

Materials i Mètodes

Aquest capítol descriu el dataset utilitzat, l'anàlisi exploratòria realitzada, el pipeline de preprocessament aplicat i les mètriques d'avaluació seleccionades. També s'explica la partició de dades i les eines usades per al desenvolupament del projecte.

3.1 Descripció del dataset i anàlisi exploratòria

El dataset utilitzat en aquest treball prové de ClinVar, una base de dades pública de variants genètiques amb interpretacions clíniques aportades per laboratoris de tot el món. En aquesta secció es descriu l'origen i la construcció del dataset, les variables que el componen, les variables objectiu de cada tasca i les principals característiques observades durant l'anàlisi exploratòria.

3.1.1 Origen i construcció del dataset

El dataset utilitzat prové d'un repositori públic de Kaggle [3], basat en dades de ClinVar [19, 27] amb llicència CC0 (domini públic). El fitxer original és un VCF (Variant Call Format), el format estàndard per emmagatzemar variants genètiques, de ClinVar (GRCh37, abril de 2018) que ha estat processat mitjançant *scripts* públics disponibles a GitHub [4] per generar el fitxer final clinvar_conflicting.csv, que conté 65.188 variants i 46 columnes.

El procés de construcció combina dues fonts d'informació, les anotacions clíniques de ClinVar i les anotacions funcionals generades per Ensembl VEP. En primer lloc, s'extreuen del VCF original les variables clíniques i poblacionals. A continuació, es construeix la variable objectiu CLASS, s'assigna CLASS=1 a les variants amb el camp CLNSIGCONF no nul (classificació conflictiva) i CLASS=0 a la resta. Només es conserven les variants avaluades per múltiples *submitters*, és a dir múltiples laboratoris remitents, i s'eliminen les columnes que revelarien directament el *target*: CLNSIG, CLNSIGCONF i CLNREVSTAT. Finalment, s'incorporen les anotacions funcionals VEP descomponent el camp CSQ, que conté totes les anotacions funcionals generades per VEP concatenades.

3.1.2 Variables del dataset

El dataset conté 46 variables que es poden agrupar en tres categories segons el seu origen:

- **Dades brutes de la variant:** `CHROM`, `POS`, `REF` i `ALT`. Descriuen la localització i la naturalesa molecular del canvi genètic.
- **Anotacions clíniques de ClinVar:** variables com `AF_ESP` i `AF_EXAC`, (freqüències al·lèliques en diferents poblacions), `CLNDN` (nom de la malaltia associada), `CLNVC` (tipus de variant), i les variables objectiu `CLASS` per a la tasca binària i `CLNSIG` per a la tasca multiclasse.
- **Anotacions funcionals de VEP:** variables generades per Ensembl VEP com `Consequence`, `IMPACT`, `SIFT`, `PolyPhen`, `CADD_PHRED`, `CADD_RAW`, `BLOSUM62` i `LoFtool`, que descriuen l'efecte potencial de la variant sobre la proteïna.

La Taula 3.1 resumeix les estadístiques principals del dataset.

Tipus	**Variables**	**Percentatge**
Dades brutes de la variant	4	8,7%
Anotacions clíniques ClinVar	15	32,6%
Anotacions funcionals VEP	27	58,7%
Total	**46**	**100%**

Taula 3.1: Distribució de les variables per tipus

3.1.3 Variable objectiu CLASS

La variable CLASS indica si una variant ha rebut classificacions conflictives entre laboratoris (CLASS=1) o consistents (CLASS=0), i constitueix la variable objectiu pel primer dels dos problemes plantejats en aquest treball.

3.1.4 Variable objectiu CLNSIG

A diferència de la variable CLASS, la columna CLNSIG no està disponible al fitxer clinvar_conflicting.csv original, ja que el *script* de preprocessament de Kaggle l'elimina explícitament per evitar que reveli el *target* binari. Per al problema de classificació multiclasse, s'ha modificat el *script* de preprocessament original per conservar aquesta columna i generar un nou dataset partint dels mateixos fitxers VCF de ClinVar, aplicant els mateixos filtres que en la tasca binària per garantir la consistència entre ambdós problemes.

La columna CLNSIG conté la significança clínica assignada per ClinVar i presenta una gran varietat de valors. Tot i que la majoria de les variants es classifiquen amb una de les cinc categories ACMG estàndard (`Benign, Likely_benign, Uncertain_significance, Likely_pathogenic, Pathogenic`), una part conté etiquetes combinades amb modificadors addicionals (per exemple, `Benign,_risk_factor` o `Pathogenic/Likely_pathogenic`), així com valors no informatius o contradictoris (`not_provided, Conflicting_interpretations_of_pathogenicity`).

Per simplificar el problema i obtenir classes ben definides, els valors s'han agrupat en tres categories:

- **Classe 0 - Benigna:** agrupa `Benign`, `Likely_benign`, `Benign/Likely_benign` i les seves variants amb modificadors addicionals.
- **Classe 1 - VUS:** correspon únicament a `Uncertain_significance`.
- **Classe 2 - Patogènica:** agrupa `Pathogenic`, `Likely_pathogenic`, `Pathogenic/Likely_pathogenic` i les seves variants amb modificadors addicionals.

Les variants amb valors com `Conflicting_interpretations_of_pathogenicity`, `not_provided` o categories molt minoritàries (`drug_response`, `risk_factor`, etc.) s'han descartat per no pertànyer clarament a cap classe. És rellevant l'exclusió de les variants amb classificació conflictiva, aquestes corresponen a les variants amb CLASS=1 de la predicció binària, i per definició no es poden assignar a una única categoria clínica. Després d'aquest filtratge, el dataset resultant conté 57.691 variants, 7.497 menys que el conjunt utilitzat en la tasca binària.

3.1.5 Desbalanceig de classes i valors nuls

3.1.5.1 Classificació binària (CLASS)

La Figura 3.1 mostra la distribució de la variable objectiu `CLASS`. El 74,8% de les variants corresponen a CLASS=0 (sense conflicte) i el 25,2% a CLASS=1 (conflictives), fet que implica un desbalanceig que justifica l'ús del ROC AUC, una mètrica poc sensible al desbalanceig que s'explica en detall a la secció 3.5.1.

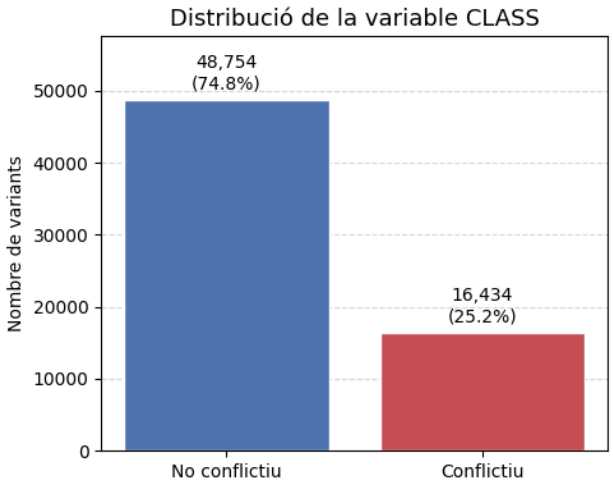


Figura 3.1: Distribució de la variable objectiu CLASS.

3.1.5.2 Classificació multiclasse (CLNSIG)

La Figura 3.2 mostra la distribució de les tres classes després de l'agrupació i el filtratge. S'observa un cert desbalanceig en el conjunt de dades, amb una major proporció de la classe *Benign* (46,3%), seguida de *VUS* (35,3%) i, finalment, la classe *Pathogenic*, que és la minoritària amb un 18,4%.

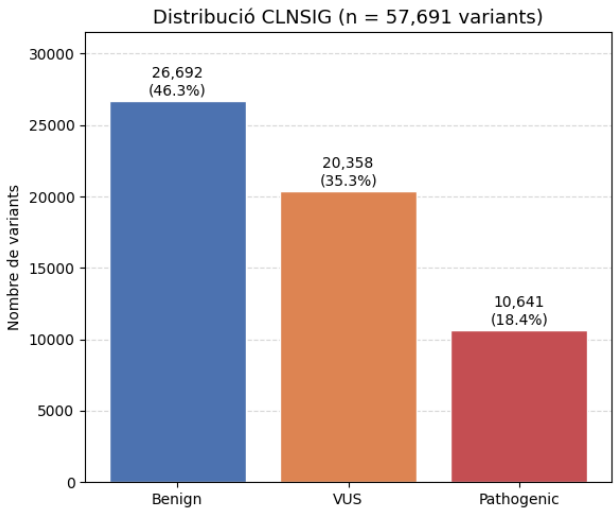


Figura 3.2: Distribució de la variable objectiu CLNSIG.

3.1.5.3 Valors nuls

L'anàlisi de valors nuls mostra que 9 columnes superen el 95% de valors absents: `MOTIF_SCORE_CHANGE`, `HIGH_INF_POS`, `MOTIF_POS`, `MOTIF_NAME`, `DISTANCE`, `SSR`, `CLNSIGINCL`, `CLNDISDBINCL` i `CLNDNINCL`. Aquestes columnes s'eliminen en la fase de preprocessament. La Figura 3.3 mostra el percentatge de valors nuls de les 20 columnes amb més absències.

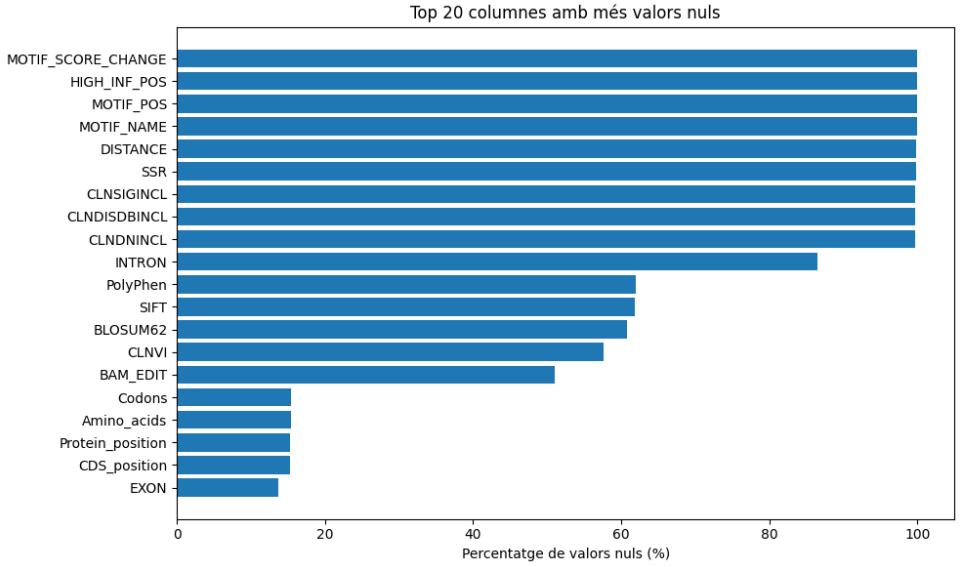


Figura 3.3: Top 20 columnes amb més valors nuls.

3.2 Partició de dades

El dataset s'ha dividit en dos conjunts de manera fixa per garantir comparacions justes entre tots els models:

- **Conjunt d'entrenament i validació (`X_train_val`) :** 90% de les dades,

utilitzat per entrenar els models i optimitzar els hiperparàmetres mitjançant validació creuada estratificada de 5 `folds`.

- **Conjunt de test (`X_test`):** 10% de les dades, reservat per l'avaluació final dels models.

La partició s'ha realitzat amb `train_test_split` de scikit-learn, fixant `random_state=42` per garantir la reproductibilitat i usant `stratify=y` per preservar la proporció de classes en ambdós conjunts.

3.3 Preprocessament

En aquesta secció es descriu el pipeline de preprocessament aplicat al dataset, incloent l'eliminació de columnes amb massa valors nuls, la imputació i normalització de variables, i la construcció dels dos datasets, anotat i bàsic, utilitzats per comparar l'aportació de les anotacions funcionals de VEP.

3.3.1 Pipeline de preprocessament

El pipeline de preprocessament és idèntic per a les dues tasques de classificació. Com a primer pas, s'eliminen les 9 columnes identificades a l'anàlisi de valors nuls amb més del 95% d'absències (MOTIF_SCORE_CHANGE, HIGH_INF_POS, MOTIF_POS, MOTIF_NAME, DISTANCE, SSR, CLNSIGINCL, CLNDISDBINCL i CLNDNINCL).

A continuació, s'aplica un pipeline diferenciat per tipus de variable, implementat amb `ColumnTransformer` de scikit-learn:

- **Variables numèriques:** imputació de valors nuls amb la mediana i normalització amb `StandardScaler`.
- **Variables categòriques:** imputació amb el valor més freqüent i codificació mitjançant `OneHotEncoder`, ignorant categories desconegudes al test i agrupant les categories amb menys de 20 aparicions en una categoria comuna infreqüent, per evitar la generació de columnes molt esparses que afegeixen soroll i augmenten innecessàriament la dimensionalitat.

Tot el pipeline s'ajusta sobre el conjunt d'entrenament `fit` sobre `X_train_val` i s'aplica al conjunt de test `transform` sobre `X_test`, evitant així la fuita d'informació (*data leakage*), és a dir, l'ús de qualsevol estadístic calculat sobre el conjunt de test durant l'entrenament, que podria sobreestimar el rendiment del model.

3.3.2 Dataset Bàsic vs Dataset Anotat

Una de les hipòtesis centrals d'aquest treball és avaluar si les anotacions funcionals generades per VEP aporten valor predictiu real per sobre de les dades clíniques bàsiques que ja proporciona ClinVar. Per respondre a aquesta pregunta de manera controlada, s'han construït dues configuracions de dataset per a cada tasca de classificació, diferenciades pel conjunt de variables incloses:

- **Dataset Anotat:** conté les 46 variables originals, incloent totes les anotacions funcionals de VEP.
- **Dataset Bàsic:** s'obté eliminant les 22 columnes provinents de VEP (`Consequence`, `IMPACT`, `SIFT`, `PolyPhen`, `CADD_PHRED`, `CADD_RAW`, `BLOSUM62`, entre d'altres), quedant-se únicament amb les variables clíniques de ClinVar i les dades brutes de la variant. Cal notar que de les 27 anotacions VEP, 5 (`MOTIF_NAME`, `MOTIF_POS`, `MOTIF_SCORE_CHANGE`, `HIGH_INF_POS` i `DISTANCE`) ja s'eliminen a la fase de neteja de valors nuls, per tant, la configuració Bàsica només cal que descarti les 22 columnes VEP restants.

La Taula 3.2 resumeix les dimensions de partida de cada configuració abans del preprocessament (incloent la columna *target*).

Tasca	**Dataset**	**Mostres**	**Columnes**
Binària	Anotat	65.188	46
Binària	Bàsic	65.188	24
Multiclasse	Anotat	57.691	47
Multiclasse	Bàsic	57.691	25

Taula 3.2: Dimensions dels datasets per a cada tasca i configuració.

Tots dos datasets passen pel mateix pipeline de preprocessament, garantint que qualsevol diferència de rendiment entre les configuracions Bàsic i Anotat sigui atribuïble exclusivament a la informació funcional de VEP.

Un cop aplicat el pipeline de preprocessament i la partició en conjunts d'entrenament i prova, les dimensions que reben els models són les que es mostren a la Taula 3.3.

Tasca	**Dataset**	**Entrenament**	**Prova**	**Variables**
Binària	Anotat	58.669	6.519	36
Binària	Bàsic	58.669	6.519	14
Multiclasse	Anotat	51.921	5.770	36
Multiclasse	Bàsic	51.921	5.770	14

Taula 3.3: Dimensions dels datasets després del preprocessament i la partició train/test.

3.4 Eines i entorn de desenvolupament

Aquesta secció descriu el conjunt d'eines, llibreries i entorn de desenvolupament utilitzat per implementar els experiments d'aquest treball.

3.4.1 Llenguatge de programació

Python 3.9.6 ha estat el llenguatge principal utilitzat en tot el projecte. És el llenguatge de referència en ciència de dades i bioinformàtica gràcies a la seva simplicitat.

3.4.2 Llibreries de machine learning

Les llibreries de *machine learning* han estat el nucli tècnic del projecte, proporcionant tant els algoritmes de classificació com les eines d'interpretabilitat.

3.4.2.1 scikit-learn

scikit-learn [30] ha estat la llibreria central del projecte. S'ha utilitzat per implementar tots els models de classificació, el pipeline de preprocessament, la cerca d'hiperparàmetres amb `GridSearchCV` i `RandomizedSearchCV`, la validació creuada estratificada amb `StratifiedKFold` de 5 `folds`, i el càlcul de mètriques d'avaluació com ROC AUC i F1-Score.

3.4.2.2 SHAP

La llibreria SHAP [23] s'ha utilitzat per calcular els valors d'interpretabilitat en tots els models. SHAP (*SHapley Additive exPlanations*) és una implementació eficient dels valors de Shapley, presentats al Capítol 2, que ofereix implementacions especialitzades segons el tipus de model:

- `LinearExplainer` per als models lineals (LinearSVC, Logística L1 i L2). Aprofita que la contribució de cada variable es pot calcular directament a partir dels coeficients apresos i la diferència entre el valor observat i el valor mitjà a les dades, cosa que permet obtenir els valors SHAP de manera exacta i amb un cost computacional molt baix.
- `TreeExplainer` per al Decision Tree i el Random Forest [22]. Calcula els valors SHAP de manera exacta en temps polinòmic, aprofitant l'estructura de l'arbre per evitar el càlcul combinatori complet (que seria exponencial). En el cas del Random Forest s'ha utilitzat l'opció `approximate=True` per reduir encara més el cost computacional.
- `KernelExplainer` per al MLP. Mètode agnòstic al model que aproxima els valors de Shapley mitjançant una regressió lineal ponderada sobre coalicions (subconjunts) de variables. El seu principal inconvenient és el cost computacional elevat, motiu pel qual s'utilitza amb un subconjunt reduït de mostres com a *background*. En aquest treball s'ha usat un mostreig aleatori de 100 mostres del conjunt d'entrenament i 200 del conjunt de test, amb `random_state=42` per garantir la reproductibilitat.

3.4.3 Llibreries de manipulació de dades

Les llibreries de manipulació de dades han permès carregar, transformar i processar el dataset de manera eficient.

3.4.3.1 pandas i NumPy

pandas [24] s'ha utilitzat per a la càrrega, manipulació i anàlisi del dataset. NumPy [15] ha proporcionat el suport per a operacions numèriques i matricials, especialment en el processament dels valors SHAP.

3.4.4 Visualització

Les eines de visualització s'han utilitzat per generar totes les figures i gràfics del treball.

3.4.4.1 Matplotlib i Seaborn

Matplotlib [17] i Seaborn [36] s'han utilitzat per generar totes les figures del treball: distribucions de classes, matrius de confusió, corbes ROC, gràfics de valors nuls i visualitzacions SHAP.

3.4.5 Entorn de desenvolupament

3.4.5.1 Visual Studio Code

Tot el codi s'ha desenvolupat i executat en Visual Studio Code amb la integració nativa de Jupyter Notebooks. El projecte s'ha organitzat en quatre notebooks que separen cada combinació de tasca i configuració de dataset:

- `class_annotated`: classificació binària (CLASS) sobre el dataset amb anotacions VEP. Desa els resultats finals en un fitxer `.pkl` per a la comparació posterior amb la configuració bàsica.
- `class_basic`: classificació binària (CLASS) sobre el dataset sense anotacions VEP. Carrega els resultats de `class_annotated` per generar les comparacions Anotat vs Bàsic.
- `clnsig_annotated`: classificació multiclasse (CLNSIG) sobre el dataset amb anotacions VEP. Desa els resultats finals en un fitxer `.pkl` per a la comparació posterior amb la configuració bàsica.
- `clnsig_basic`: classificació multiclasse (CLNSIG) sobre el dataset sense anotacions VEP. Carrega els resultats de `clnsig_annotated` per generar les comparacions Anotat vs Bàsic.

Cada notebook conté el preprocessament, l'entrenament dels models i l'anàlisi d'interpretabilitat corresponent. Aquesta organització modular facilita la reproductibilitat i la comparació directa entre configuracions, ja que cada experiment està aïllat en un fitxer independent.

3.4.6 Reproductibilitat

La Taula 3.4 resumeix les versions principals de les llibreries utilitzades. Tots els experiments s'han executat amb `random_state=42` per garantir la reproductibilitat dels resultats.

Llibreria	Versió
Python	3.9.6
scikit-learn	1.6.1
pandas	2.3.3
NumPy	2.0.2
SHAP	0.49.1
Matplotlib	3.9.4
Seaborn	0.13.2

Taula 3.4: Versions de les principals eines i llibreries utilitzades.

3.5 Mètriques d'avaluació

La selecció de les mètriques d'avaluació condiciona directament les conclusions que es poden extreure d'un model. En problemes amb desbalanceig de classes, mètriques estàndard com l'*accuracy* poden donar una imatge enganyosa del rendiment, ja que el model prediu sempre la classe majoritària, podent obtenir un `accuracy` alt tot i ser inservible. En aquesta secció es descriuen les mètriques utilitzades en aquest treball. El ROC AUC per a la classificació binària, el F1-Score macro per a la classificació multiclasse, i la matriu de confusió com a eina complementària per analitzar els errors de cada model.

3.5.1 ROC AUC (classificació binària)

La corba ROC (Receiver Operating Characteristic) representa la relació entre la taxa de vertaders positius (sensibilitat) i la taxa de falsos positius (1 - especificitat) per a tots els possibles llindars de classificació [13]. L'àrea sota aquesta corba (AUC, Area Under the Curve) resumeix el rendiment global del model en un únic valor entre 0 i 1, on 0,5 correspon a un classificador aleatori i 1,0 a un classificador perfecte. El ROC AUC equival a la probabilitat que el model assigni una puntuació més alta a una mostra positiva aleatòria que a una negativa aleatòria [13].

S'ha escollit el ROC AUC com a mètrica principal per a la classificació binària per a dues raons. Primer, és robust davant el desbalanceig entre classes, ja que no depèn d'un llindar fix de classificació. Segon, permet comparar el poder discriminatiu de models de naturalesa molt diferent, com és en el cas d'aquest treball on es comparen models lineals amb arbres o xarxes neuronals.

3.5.2 F1-Score (classificació multiclasse)

El F1-Score és la mitjana harmònica entre precisió (*precision*) i la sensibilitat (*recall*):

$$\text{Precisió} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN} \tag{3.1}$$

$$F1 = 2 \cdot \frac{\text{Precisió} \times \text{Recall}}{\text{Precisió} + \text{Recall}} \tag{3.2}$$

La precisió mesura la proporció de prediccions positives correctes, i el recall la proporció de positius reals identificats pel model.

En problemes multiclasse amb desbalanceig, s'utilitza el F1-Score macro, que calcula el F1 per a cada classe independentment i en fa la mitjana no ponderada:

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{C} \sum_{c=1}^{C} \mathrm{F1}_c \tag{3.3}$$

on C és el nombre de classes. Aquest enfocament dona el mateix pes a totes les classes independentment de la seva freqüència, penalitzant especialment el mal rendiment en les classes minoritàries.

S'ha escollit aquesta mètrica perquè penalitza els models que ignoren les classes minoritàries i perquè és la mètrica estàndard per a problemes multiclasse amb classes desbalancejades en l'aprenentatge automàtic aplicat a genòmica [29].

3.5.3 Matriu de confusió

La matriu de confusió resumeix les prediccions d'un model comparant les classes reals amb les classes predites. Per a la classificació binària té la forma de la Figura 3.4, on:

- **TP (True Positive)**: variants conflictives classificades correctament.
- **TN (True Negative):** variants no conflictives classificades correctament.
- **FP (False Positive):** variants no conflictives classificades errròniament com a conflictives.
- **FN (False Negative):** variants conflictives classificades errròniament com a no conflictives.

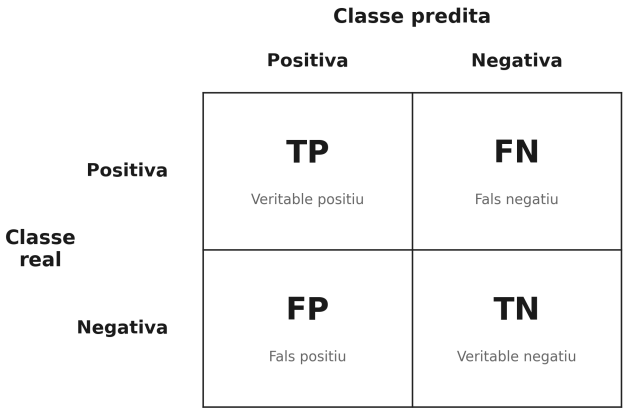


Figura 3.4: Estructura d'una matriu de confusió binària.

Per a la classificació multiclasse, la matriu s'estén a tres classes (benigna, VUS i patogènica), cosa que permet identificar quins tipus de confusió cometen els models entre categories adjacents.

3.5.4 SHAP normalitzada

Per comparar la importància de les variables entre models d'estructura diferents, els valors SHAP absoluts no són directament comparables, ja que cada model genera SHAP amb escales i magnituds diferents. Per fer la comparació possible, s'aplica una normalització per mostra.

Per a cada mostra j del conjunt de test, es calcula la contribució relativa de cada variable i com:

$$\tilde{\phi}_{i,j} = \frac{|\phi_{i,j}|}{\sum_{k=1}^{n} |\phi_{k,j}|} \tag{3.4}$$

on $\phi_{i,j}$ és el valor SHAP de la variable i a la mostra j, i n és el nombre total de variables. D'aquesta manera, per a cada mostra els valors absoluts normalitzats sumen 1, i cada cel·la representa la fracció de la contribució total que aporta cada variable a la predicció d'aquella mostra.

La importància normalitzada final per a cada variable s'obté fent la mitjana entre totes les mostres del conjunt de test:

$$I_i = \frac{1}{N} \sum_{j=1}^{N} \tilde{\phi}_{i,j} \tag{3.5}$$

on N és el nombre de mostres. El resultat I_i és un valor entre 0 i 1 que permet comparar la importància relativa de cada variable entre models diferents, independentment de les escales SHAP de cada classificador.

Capítol 4

Implementació

Aquest capítol descriu la configuració experimental utilitzada per entrenar i avaluar els models de classificació. El mateix protocol s'aplica a les quatre combinacions experimentals, cada una de les dues tasques, binària i multiclasse, sobre cadascuna de les dues configuracions de dataset, anotat i bàsic.

4.1 Configuració general

Tots els experiments comparteixen la configuració següent per garantir la reproductibilitat i la comparació justa entre models:

- **Llavor aleatòria:** `random_state=42` en tots els models i particions.
- **Validació creuada:** `StratifiedKFold` amb 5 `folds` i `shuffle=True`, preservant la proporció de classes en cada `fold`.
- **Cerca d'hiperparàmetres:** `GridSearchCV` per a la majoria dels models i `RandomizedSearchCV` per al Random Forest en totes les tasques, així com per al MLP en la tasca multiclasse, optimitzant el ROC AUC en la tasca binària i el F1-Score macro en la multiclasse.
- **Gestió del desbalanceig:** `class_weight="balanced"` en tots els models que ho suporten. El MLP queda fora d'aquesta configuració ja que MLPClassifier de scikit-learn no admet aquest paràmetre. Aquesta limitació es té en compte a l'hora d'interpretar els resultats.

4.2 Cerca d'hiperparàmetres per model

La Taula 4.1 resumeix els espais de cerca d'hiperparàmetres per a cada model.

Model	**Hiperparàmetre**	**Valors explorats**
LinearSVC	`C`	0.001, 0.01, 0.1, 1.0, 10.0
	`class_weight`	None, balanced
Logistic Regression L1/L2	`C`	0.001, 0.01, 0.1, 1.0, 10.0
	`class_weight`	None, balanced
Decision Tree	`max_depth`	5, 10, 20, None
	`min_samples_split`	2, 10, 20
	`min_samples_leaf`	1, 5, 10
	`class_weight`	None, balanced
Random Forest	`n_estimators`	100, 200, 300
	`max_depth`	None, 10, 20, 30
	`min_samples_split`	2, 5, 10
	`min_samples_leaf`	1, 2, 5
	`max_features`	sqrt, 0.3
	`class_weight`	None, balanced
MLP	`hidden_layer_sizes`	(64,), (128,), (64,32)
	`alpha`	0.001, 0.005
	`learning_rate_init`	0.0001, 0.0005, 0.001

Taula 4.1: Espais de cerca d'hiperparàmetres per a cada model.

4.3 Protocol d'avaluació

El protocol d'avaluació segueix els passos següents per a cada model i cada combinació de tasca i dataset, tal com s'il·lustra a la Figura 4.1:

1. **Cerca d'hiperparàmetres:** S'aplica `GridSearchCV` o `RandomizedSearchCV` sobre el conjunt `X_train_val`, amb validació creuada estratificada de 5 folds. La mètrica d'optimització és el ROC AUC per a la tasca binària i el F1-Score macro per a la multiclasse.

2. **Reentrenament:** El millor model trobat es reentrena automàticament sobre tot el conjunt `X_train_val` gràcies al paràmetre `refit=True`.

3. **Avaluació final:** El model reentrenat s'avalua una única vegada sobre el conjunt de test `X_test`, que no ha participat en cap etapa anterior, garantint una estimació imparcial del rendiment real.

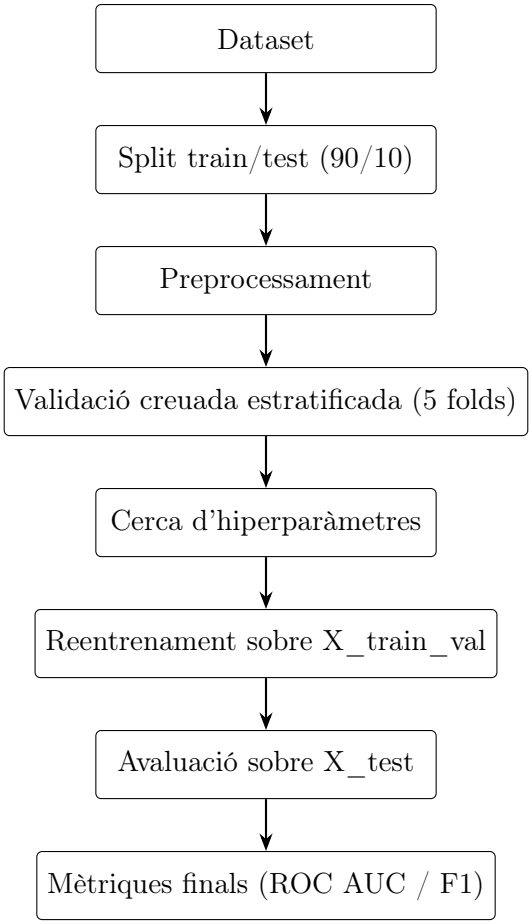


Figura 4.1: Protocol d'avaluació aplicat a cada model i configuració.

4.4 Disponibilitat del codi

Tot el codi del projecte és públicament accessible al repositori de GitHub del projecte, incloent-hi els notebooks per a les dues tasques de classificació.

`https://github.com/eliagarcia/tfg-genetic-variants`

Capítol 5

Resultats i Discussió

Aquest capítol presenta els resultats obtinguts per als sis models de classificació avaluats en les dues tasques del treball, la classificació binària de conflictes entre laboratoris (CLASS) i la classificació multiclasse de la significança clínica (CLNSIG). Per cadascuna de les tasques, es comparen els resultats obtinguts amb el dataset anotat que inclou les anotacions funcionals VEP, i amb el dataset bàsic, que conté únicament les variables clíniques de ClinVar. Finalment, es presenta una discussió general comparant els resultats obtinguts.

5.1 Classificació Binària

En aquesta secció es presenten els resultats de la classificació binària, on l'objectiu és predir si una variant rebrà classificacions conflictives entre laboratoris. Primer es mostren els resultats obtinguts amb el dataset anotat, després el dataset bàsic, i finalment es compara el rendiment entre les dues configuracions per avaluar l'aportació de les anotacions funcionals VEP.

5.1.1 Dataset Anotat

La Taula 5.1 resumeix els resultats obtinguts pels sis models sobre el dataset anotat.

Model	**CV AUC**	**Test AUC**	**Millors hiperparàmetres**
LinearSVC	0.742	0.743	C=0.01
MLP	0.759	0.763	hidden=(64,32), lr=0.001
Logística L2	0.749	0.749	C=0.1
Logística L1	0.743	0.747	C=1.0
Decision Tree	0.758	0.720	max_depth=10, class_weight=balanced
Random Forest	0.814	**0.827**	n_est=100, max_depth=30, max_feat=0.3

Taula 5.1: Resultats de la classificació binària sobre el dataset anotat.

El Random Forest obté el millor rendiment amb un Test AUC de 0.827, seguit del MLP amb 0.763. Els models lineals (LinearSVC, Logística L1 i L2) presenten un rendiment similar entre ells, al voltant de 0.74–0.75. El Decision Tree és el model amb menor rendiment al conjunt de test, amb un Test AUC de 0.720, tot i haver assolit un CV AUC de 0.758. Aquesta diferència entre validació i test de gairebé

0,04 punts, suggereix una certa inestabilitat del model davant variacions del conjunt de dades, inherent als arbres de decisió únics.

La Figura 5.1 mostra les corbes ROC de tots els models al conjunt de test. S'observa la superioritat del Random Forest, que presenta una corba per sobre de la resta.

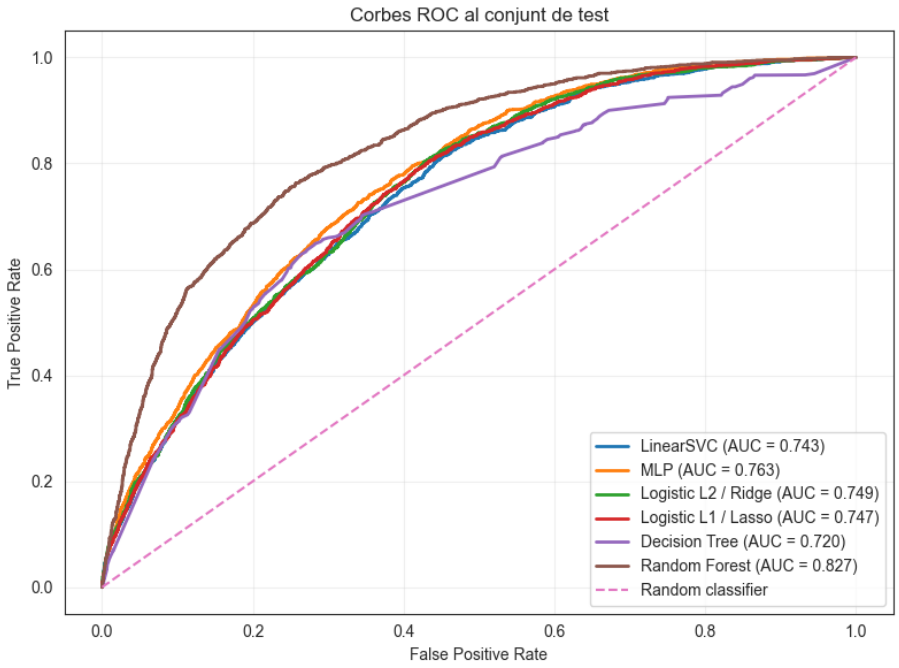


Figura 5.1: Corbes ROC dels sis models al conjunt de test (dataset anotat, tasca binària).

Les Figures 5.2 i 5.3 mostren les matrius de confusió del millor i pitjor model, respectivament. El Random Forest classifica correctament 4.665 variants no conflictives i 475 conflictives, però presenta un nombre considerable de falsos negatius (1.168), és a dir, variants realment conflictives que el model no identifica com a tals. Aquest comportament reflecteix la dificultat de detectar la classe minoritària (`CLASS=1`, conflictives, 25%) davant la classe majoritària (`CLASS=0`, no conflictives, 74%).

El Decision Tree, en canvi, mostra un patró d'errors diferent: identifica més variants conflictives correctament (1.032 vs 475 del Random Forest), però a costa d'un nombre molt superior de falsos positius (1.285 vs 211 del Random Forest). Aquesta major sensibilitat a la classe minoritària es contraresta amb una pèrdua significativa de precisió global, la qual cosa explica el seu menor AUC.

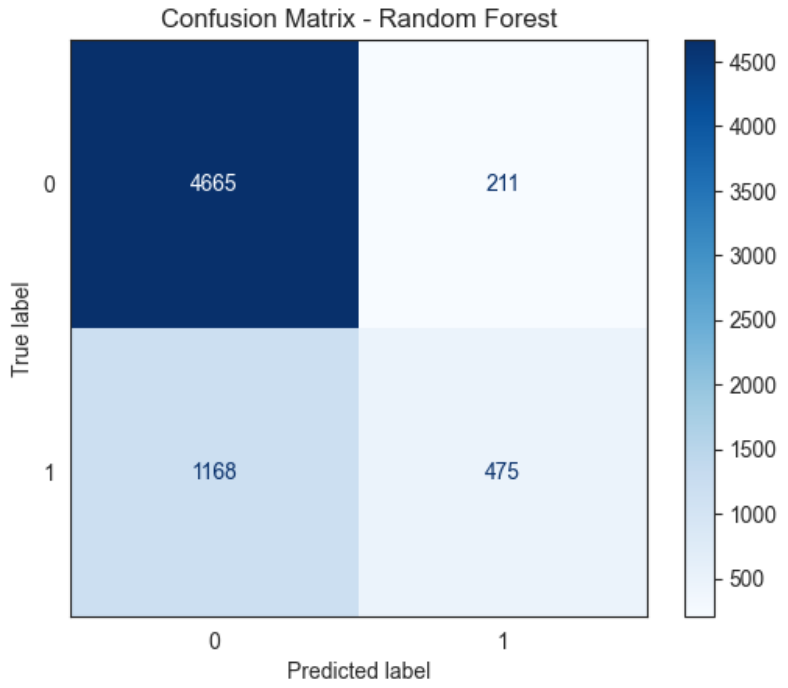


Figura 5.2: Matriu de confusió del Random Forest (dataset anotat, tasca binària).

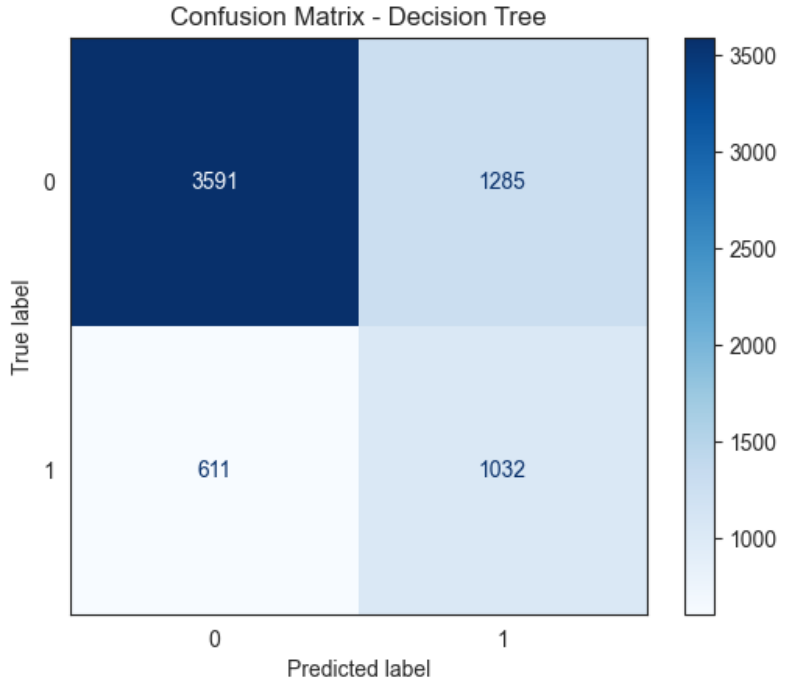


Figura 5.3: Matriu de confusió de l'Arbre de Decisió (dataset anotat, tasca binària).

Considerant els resultats obtinguts, els models no lineals superen els models lineals en aquesta tasca, fet que apunta que la relació entre les variables del dataset i la presència de conflictes no és estrictament lineal. El Random Forest destaca per la seva capacitat d'aprofitar interaccions complexes entre variables funcionals i clíniques, sent el millor model avaluat.

5.1.2 Dataset Bàsic

La Taula 5.2 resumeix els resultats obtinguts pels sis models sobre el dataset bàsic, que exclou les anotacions funcionals de VEP.

Model	CV AUC	Test AUC	Millors hiperparàmetres
LinearSVC	0.740	0.737	C=0.01
MLP	0.766	0.757	hidden=(64,), lr=0.001
Logística L2	0.746	0.743	C=0.1
Logística L1	0.744	0.743	C=1.0
Decision Tree	0.743	0.711	max_depth=10
Random Forest	0.791	**0.799**	n_est=200, max_depth=20

Taula 5.2: Resultats de la classificació binària sobre el dataset bàsic.

De manera similar al cas anterior, Random Forest continua sent el millor model amb Test AUC de 0.799, seguit del MLP amb 0.757. Els models lineals (LinearSVC, L1 i L2) obtenen resultats al voltant de 0.74, i el Decision Tree torna a ser el model amb menor rendiment amb 0.711.

La Figura 5.4 mostra les corbes ROC dels sis models al conjunt de test amb el dataset bàsic. La distribució qualitativa de les corbes és pràcticament idèntica a la del dataset anotat, amb el Random Forest per sobre dels altres.

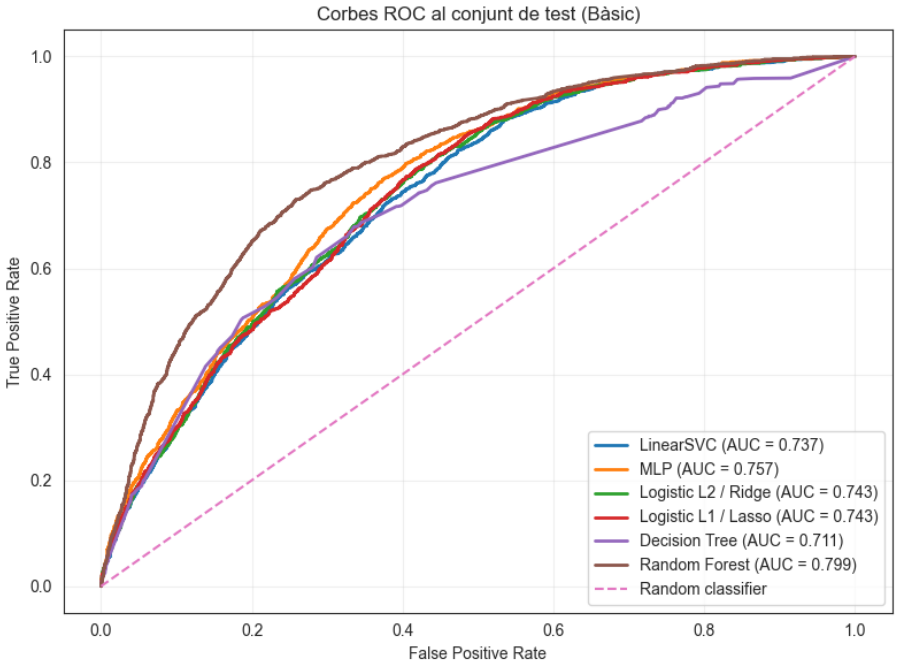


Figura 5.4: Corbes ROC dels sis models al conjunt de test (dataset bàsic, tasca binària).

Les Figures 5.5 i 5.6 mostren les matrius de confusió del millor i pitjor model. El Random Forest classifica correctament 4.682 variants no conflictives i 369 de conflictives, però presenta 1.274 falsos negatius, variants conflictives que el model

classifica com a no conflictives. Aquest nombre és superior al de la configuració anotada (1.168 falsos negatius i 475 conflictives detectades), indicant que les anotacions funcionals VEP ajuden el model a identificar millor la classe minoritària. En canvi, els falsos positius es mantenen baixos amb 194 davant dels 211 del dataset anotat, mantenint la bona especifitat del model.

El Decision Tree mostra un patró d'errors diferent, identificant 961 variants conflictives correctament, gairebé el triple que el Random Forest amb 369, però amb un cost molt superior en falsos positius (1.274 davant 194 del Random Forest). Aquesta major sensibilitat a la classe minoritària es contraresta amb una pèrdua de precisió global, fet que explica el seu menor AUC.

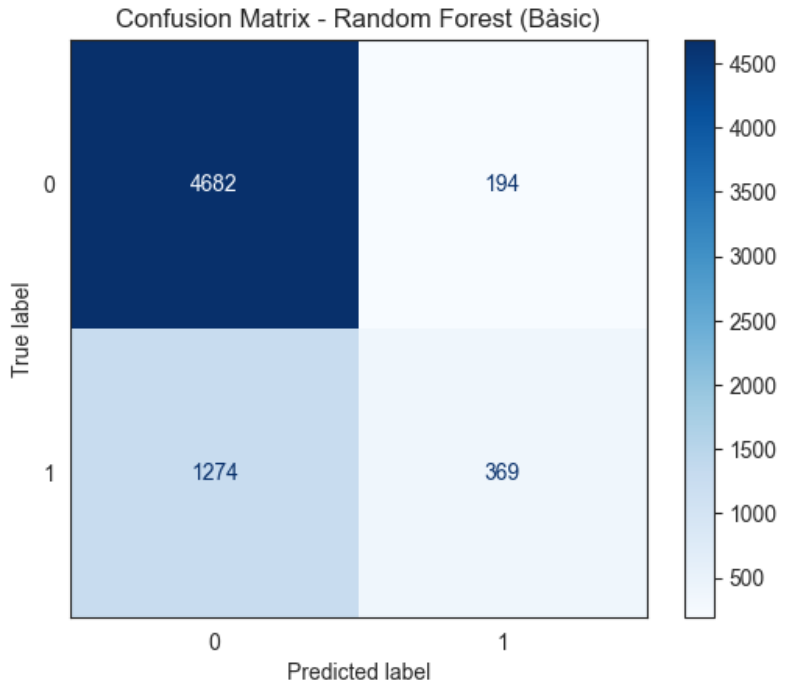


Figura 5.5: Matriu de confusió del Random Forest (dataset bàsic, tasca binària).

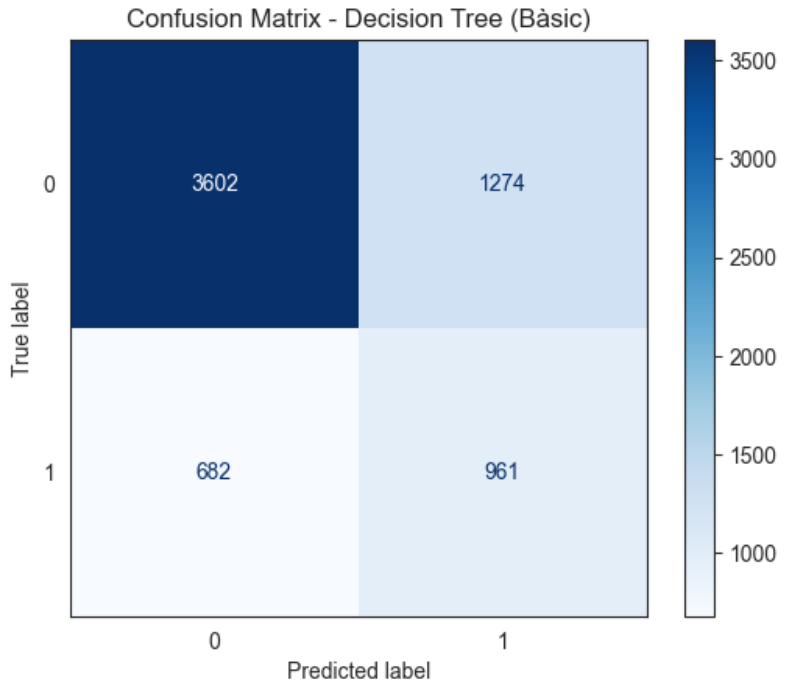


Figura 5.6: Matriu de confusió del Decision Tree (dataset bàsic, tasca binària).

Els resultats sobre el dataset bàsic confirmen el mateix patró observat al dataset anotat, el Random Forest és el model més adequat per a aquesta tasca, els models no lineals superen els models lineals, i el Decision Tree continua presentant una caiguda entre validació i test. Tanmateix, tots els models registren un rendiment lleugerament inferior respecte a la configuració anotada, fet que s'analitza en detall a la subsecció següent.

5.1.3 Comparació Anotat vs Bàsic

Per quantificar l'aportació de les anotacions funcionals VEP a la tasca de classificació binària, s'ha comparat el rendiment dels sis models entrenats sobre el dataset anotat i sobre el dataset bàsic. La Figura 5.7 mostra els valors de ROC AUC al conjunt de test per a cada model i configuració.

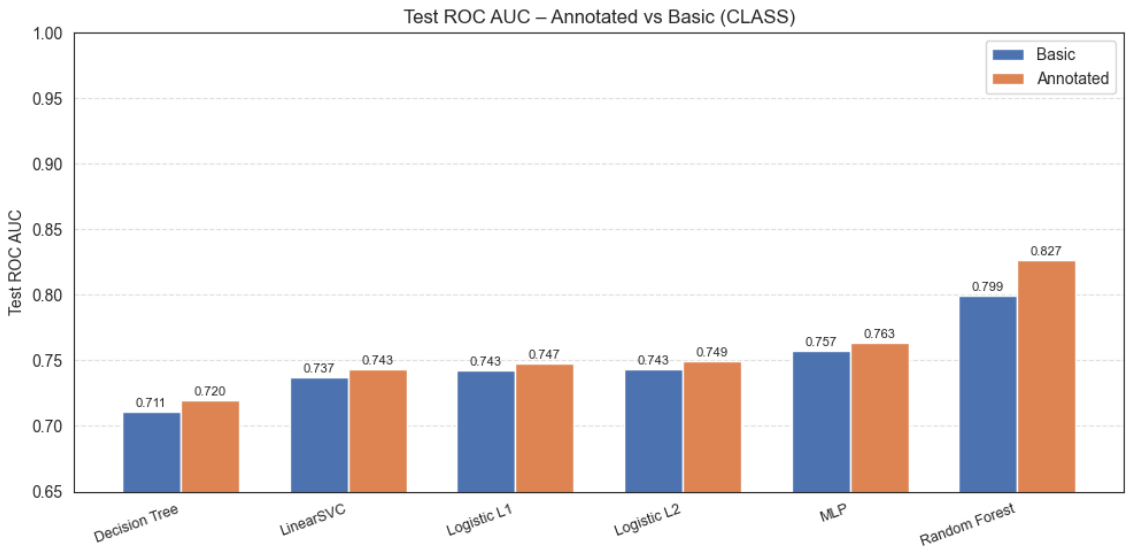


Figura 5.7: Comparació del Test ROC AUC entre el dataset Anotat i el dataset Bàsic per a cada model (tasca binària).

S'observa un patró clar, tots els models milloren amb el dataset anotat, però la magnitud de la millora és desigual. Les diferències absolutes oscil·len entre 0,004 punts d'AUC, en el cas del model de regressió logística L1 i 0,028 punts pel Random Forest, fet que representa una millora relativa d'entre l'1% i el 3,5%.

El Random Forest és el model que més aprofita les anotacions funcionals, passant d'un 0.799 a 0.827. Aquest comportament és coherent amb la naturalesa del model, els arbres aleatoris són especialment efectius capturant interaccions no lineals entre variables i les anotacions VEP aporten informació funcional complementària que els models lineals no poden explotar amb la mateixa eficàcia.

El MLP, també no lineal, segueix una tendència similar, amb una millora modesta de +0.006 punts. Els models lineals experimenten millores entre +0.004 i +0.006. Aquest resultat suggereix que la informació addicional aportada per VEP no és linealment separable respecte a la variable objectiu, i que calen models capaços de capturar relacions complexes per aprofitar-la plenament.

El Decision Tree també millora lleugerament amb +0.009, però el seu rendiment absolut continua sent el més baix dels sis models en ambdues configuracions.

Amb tot això podem dir que el Random Forest aprofita les anotacions funcionals VEP entre 4 i 7 vegades més que els models lineals, fet que confirma que el valor predictiu de les anotacions funcionals resideix en les interaccions no lineals entre variables, més que no pas en relacions amb la variable objectiu.

5.2 Interpretabilitat Classificació Binària

Aquesta secció presenta l'anàlisi d'interpretabilitat dels sis models de classificació entrenats. L'objectiu és identificar quines variables del dataset contribueixen més a les prediccions de cada model i, sobretot, detectar quines són consistentment importants entre models diferents. Aquesta informació és especialment rellevant en el context biomèdic, on entendre per què un model pren una decisió és tan important com el rendiment predictiu.

Per a cada model s'han calculat els valors SHAP [21] al conjunt de test, utilitzant l'*explainer* adequat segons el tipus de model. Pels models lineals s'ha fet servir el `LinearExplainer`, per al Decision Tree i Random Forest s'ha utilitzat el `TreeExplainer` i per al MLP el `KernelExplainer`.

L'anàlisi d'interpretabilitat per al problema binari s'ha realitzat tant sobre el dataset anotat com sobre el dataset bàsic. Primer es presenten els resultats per al dataset anotat, que conté el conjunt complet de variables clíniques i funcionals, i ofereix els millors rendiments, després els resultats per al dataset bàsic que permeten comparar com canvia la importància de les variables quan s'eliminen les anotacions funcionals VEP.

Per cada model es presenta un gràfic que mostra la contribució direccional, on el signe del valor SHAP indica el sentit de l'efecte, els valors positius (en blau) empenyen la predicció cap a la classe conflictiva, mentre que els valors negatius (en vermell) l'empenyen cap a la classe no conflictiva.

5.2.1 Models lineals

Les Figure 5.8, 5.9 i 5.10 mostren els valors SHAP signats dels tres models lineals avaluats, LinearSVC, Logística L2 Ridge i Logística L1 Lasso.

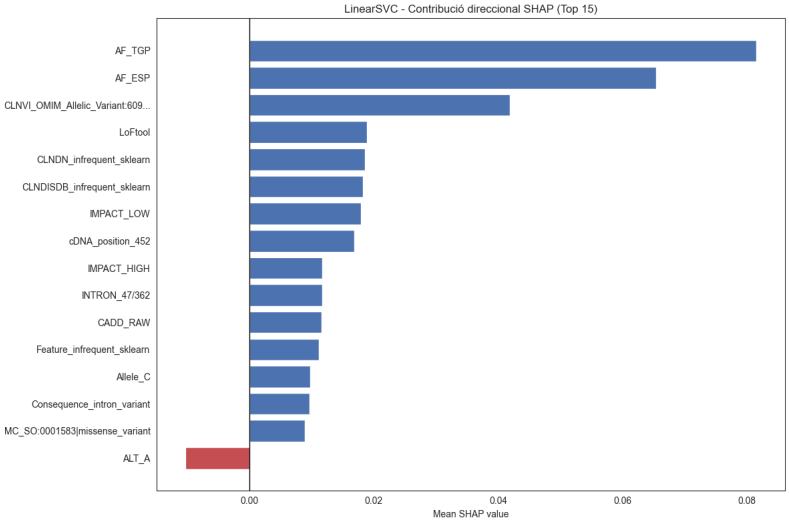


Figura 5.8: Valors SHAP per al model LinearSVC.

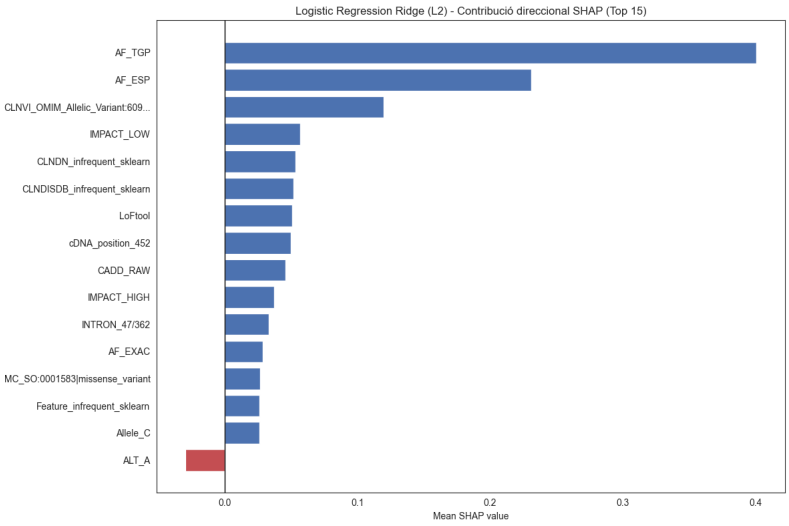


Figura 5.9: Valors SHAP per al model de regressió logística L2 (Ridge).

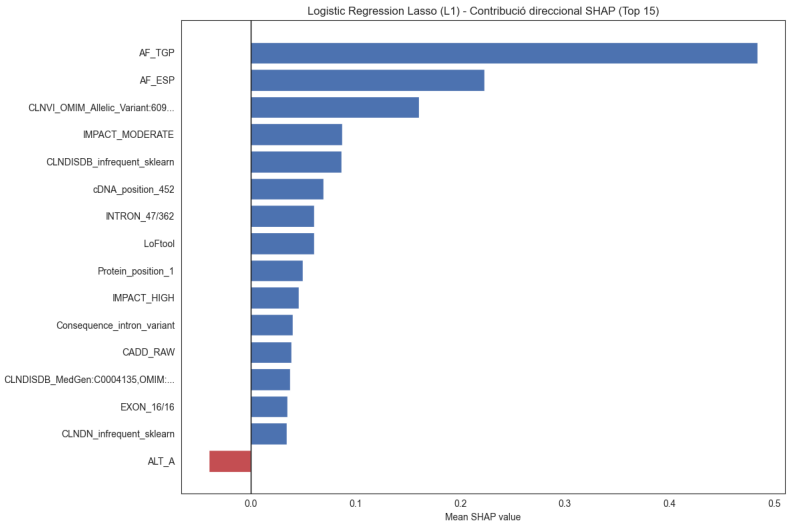


Figura 5.10: Valors SHAP per al model de regressió logística L1 (Lasso).

En tots tres casos, les variables més rellevants són les freqüències al·lèliques, `AF_TGP` i `AF_ESP`. Les freqüències al·lèliques mesuren la proporció en què apareix una variant determinada dins d'una població de referència, i són un dels predictors clínics més utilitzats [20, 19]. Les variants comunes tendeixen a ser benignes, mentre que les variants rares tenen més probabilitat de ser patogèniques [19]. Tots dos predictors mostren una contribució clarament positiva cap a la classe Conflict, indicant que les variants amb freqüències al·lèliques superiors a la mitjana tendeixen a generar més classificacions conflictives entre laboratoris. Aquest comportament és coherent amb la naturalesa del problema, les variants comunes a la població general apareixen tant en individus sans com en pacients amb malalties, fet que dificulta la seva interpretació clínica i genera divergències d'opinió entre laboratoris.

5.2.2 Decision Tree

A diferència dels models lineals, el Decision Tree permet visualitzar directament la lògica de classificació apresa pel model. La Figura 5.11 mostra els tres primers nivells de l'arbre òptim trobat per validació creuada.

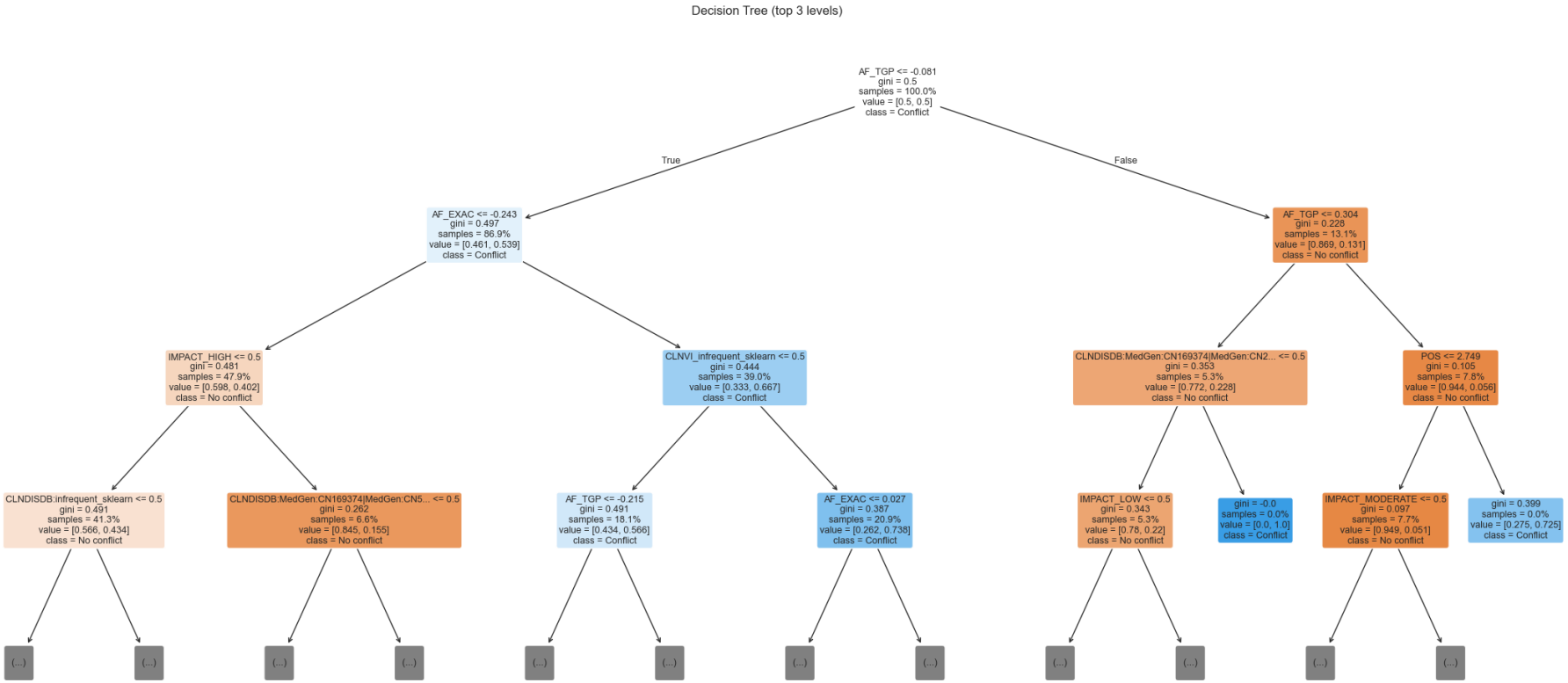


Figura 5.11: Estructura dels tres primers nivells de l'Arbre de Decisió (tasca binària, dataset anotat).

La primera divisió que fa el model es basa en la variable `AF_TGP`, confirmant així que la freqüència al·lèlica és el predictor més discriminatiu entre les variants conflictives i les no conflictives. Quan `AF_TGP` és baix, variants rares, l'arbre tendeix a explorar branques associades a la classe Conflict, mentre que valors elevats porten cap a la classe No conflict. A nivells inferiors apareixen de manera recurrent altres freqüències al·lèliques, `AF_EXAC`, i de nou `AF_TGP`, categories IMPACT (HIGH, LOW, MODERATE), que classifiquen la severitat predita de la conseqüència molecular de la variant [10], i variables d'identificació clínica com `CLNVI_infrequent_sklearn`, una categoria que agrupa identificadors de variants poc freqüents al dataset, i `CLNDISDB_MedGen:CN169374`, un identificador específic d'una base de dades de malalties.

5.2.3 Random Forest

La Figura 5.12 mostra els valors SHAP signats del Random Forest, el millor model de la tasca binària.

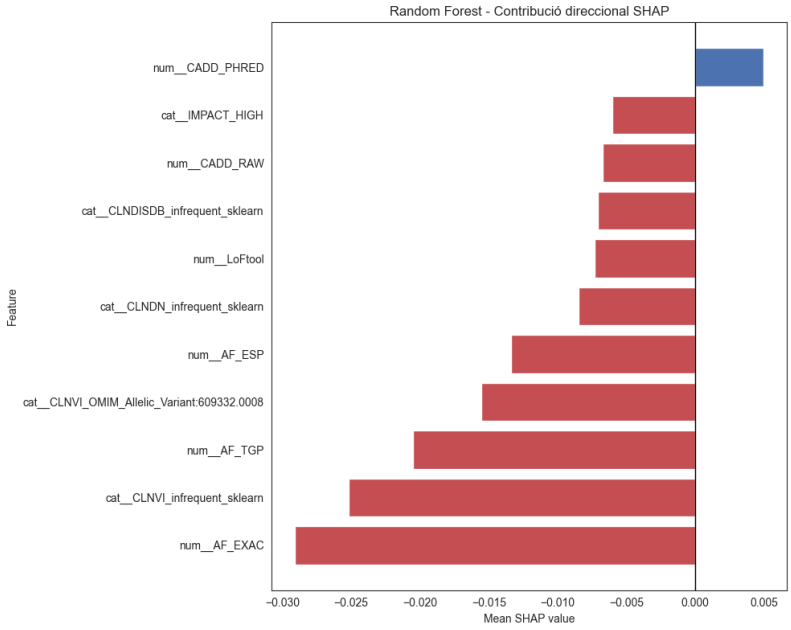


Figura 5.12: Valors SHAP signats del Random Forest (tasca binària, dataset anotat).

A diferència dels models lineals, on els valors SHAP eren positius cap a la classe Conflict, en el Random Forest la majoria de les variables principals presenten contribucions negatives cap a la classe No conflict. La raó d'aquesta diferència no és la naturalesa del model sinó la configuració d'entrenament. El Decision Tree, també basat en arbres, utilitza `class_weight=balanced` i mostra contribucions majoritàriament positives cap a Conflict, mentre que Random Forest s'ha entrenat sense ajust de pesos. En absència de balancing, el model ha après a recolzar predominantment la classe majoritària (No conflict, 74,8% de les mostres).

Les variables més rellevants tornen a ser les freqüències al·lèliques `AF_EXAC`, `AF_TGP` i `AF_ESP`. També apareixen amb pes elevat `CLNVI_infrequent_sklearn`, que agrupa identificadors de variants poc freqüents al dataset, i

CLNVI_OMIM_Allelic_Variant:609332.0008, un identificador específic d'una variant documentada al registre OMIM. També destaquen puntuació de perjudicialitat com CADD_RAW, una mètrica que combina diverses anotacions funcionals per estimar l'impacte d'una variant, i LoFtool, que quantifica la intolerància d'un gen a variants amb pèrdua de funció. Finalment, la categoria IMPACT_HIGH de VEP indica variants amb conseqüències moleculars potencialment greus.

L'única variable amb contribució positiva cap a la classe Conflict és CADD_PHRED, indicant que valors elevats d'aquesta puntuació de perjudicialitat, associats a variants potencialment més perjudicials, tendeixen a generar més conflictes entre laboratoris. Aquest comportament té sentit biològic, ja que les variants amb impacte funcional incert són les que poden rebre interpretacions divergents [32].

5.2.4 MLP

La Figura 5.13 mostra els valors SHAP signats del MLP a la tasca binària.

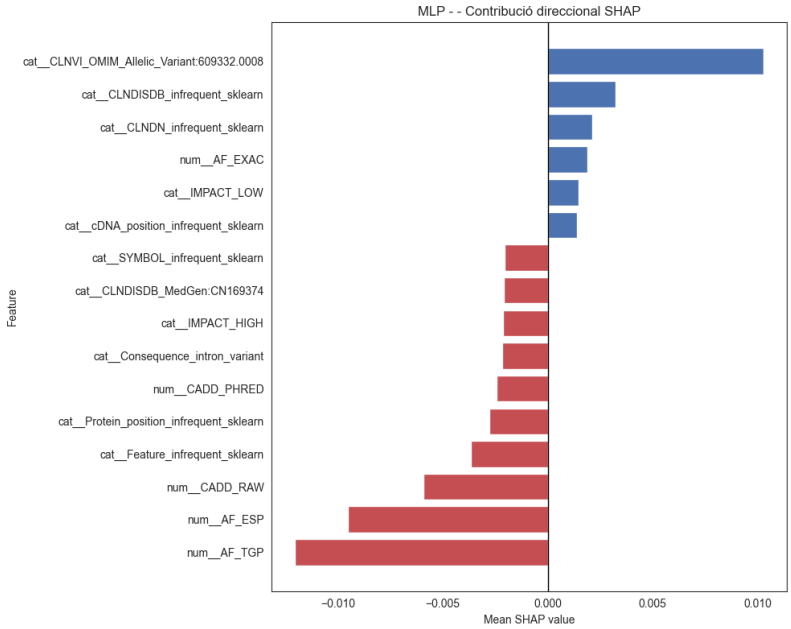


Figura 5.13: Valors SHAP signats del MLP (tasca binària, dataset anotat).

El MLP mostra un comportament diferent del de la resta de models. En comparació amb Random Forest, on les contribucions principals eren negatives cap a No conflict i dels models lineals, on predominaven les contribucions positives cap a la classe Conflict, el MLP presenta una distribució més equilibrada entre ambdues direccions.

La variable amb contribució positiva més alta cap a la classe Conflict és CLNVI_OMIM_Allelic_Variant:609332.0008, un identificador específic d'una variant documentada al registre OMIM. També apareixen amb contribució positiva la categoria CLNDISDB_infrequent_sklearn que agrupa identificadors de bases de dades de malalties poc freqüents, i IMPACT_LOW que indica variants amb conseqüències moleculars de baixa severitat predita.

En el costat negatiu, cap a la classe No conflict, destaquen com en els altres models les freqüències al·lèliques AF_TGP i AF_ESP, i puntuacions de perjudicialitat com

`CADD_RAW` i `CADD_PHRED`. En aquest model la categoria `IMPACT_HIGH` apareix amb contribució negativa, suggerint que el MLP associa variants amb impacte molecular alt a la classe No conflict.

Aquest patró diferent en el MLP és coherent amb la no-linearitat del model, al contrari dels altres classificadors, el MLP pot capturar interaccions complexes entre variables, fet que fa que algunes contribucions difereixen entre els models més senzills.

5.2.5 Concordança entre models

Per identificar les variables que els sis models consideren de manera consistent, s'ha calculat la importància SHAP normalitzada de cada característica per a cadascun dels classificadors, seguint el procediment descrit a la Secció 3.5.4. La Figura 5.14 mostra el *heatmap* comparatiu de les 15 variables més consistents.



Figura 5.14: Importància SHAP normalitzada per model: top 15 variables més consistents entre els sis classificadors (tasca binària, dataset anotat).

La variable `AF_TGP` apareix com la més important en quatre dels sis models, amb importàncies normalitzades que oscil·len entre 0.066 en el cas de MLP i 0.160 en el cas de Decision Tree. Aquesta consistència confirma que la freqüència al·lèlica al 1000 Genomes Project és el predictor més estable per a la detecció de classificacions conflictives, independentment del tipus de model utilitzat.

Seguidament, `AF_ESP` i `CLNVI_OMIM_Allelic_Variant:609332.0008` mostren també una presència consistent entre tots els models, amb importàncies normalitzades al voltant de 0.04–0.07 i 0.03–0.07 respectivament. La resta del

top 8 inclou CLNDISDB_infrequent_sklearn, CADD_RAW, IMPACT_HIGH, LoFtool i CLNDN_infrequent_sklearn totes amb un component uniforme entre classificadors.

El *heatmap* permet destacar dos patrons importants. Primer, la variable AF_EXAC mostra una importància desigual entre models, rep un valor molt baix entre 0.006 i 0.012 en els models lineals i MLP, però apareix amb valors molt elevats en el Decision Tree amb 0.232 i el Random Forest amb 0.124. Aquest comportament indica que els models basats en arbres han trobat útil aquesta variable per realitzar les divisions en l'estructura interna, mentre que els altres prefereixen la informació continguda en AF_TGP i AF_ESP.

Segon, variables com IMPACT_MODERATE i MC_SO:0001819|synonymous_variant mostren importàncies desiguals entre els models lineals. En Logística Lasso els assigna més pes que a la Ridge, reflectint la propietat de Lasso de seleccionar conjunts diferents de variables a través de la regularització.

Els resultats mostren un nucli de variables consistentment importants, les freqüències al·lèliques, l'identificador OMIM, les puntuacions de perjudicialitat com CADD i LoFtool, i la categoria IMPACT_HIGH, que aporten informació estable per a la predicció de conflictes. Aquest acord entre models de naturalesa molt diferent reforça la credibilitat dels resultats i identifica quins factors són els més informatius per a la tasca de classificació binària.

5.2.6 Comparació amb el dataset Bàsic

L'anàlisi d'interpretabilitat s'ha repetit també sobre el dataset bàsic. Com els resultats són en gran part similars als del dataset anotat, es presenten en detall a l'Apèndix A. A continuació, es comenta el més destacat i diferències més rellevants.

La Figura 5.15 mostra el *heatmap* de consens entre els sis models per al dataset bàsic.

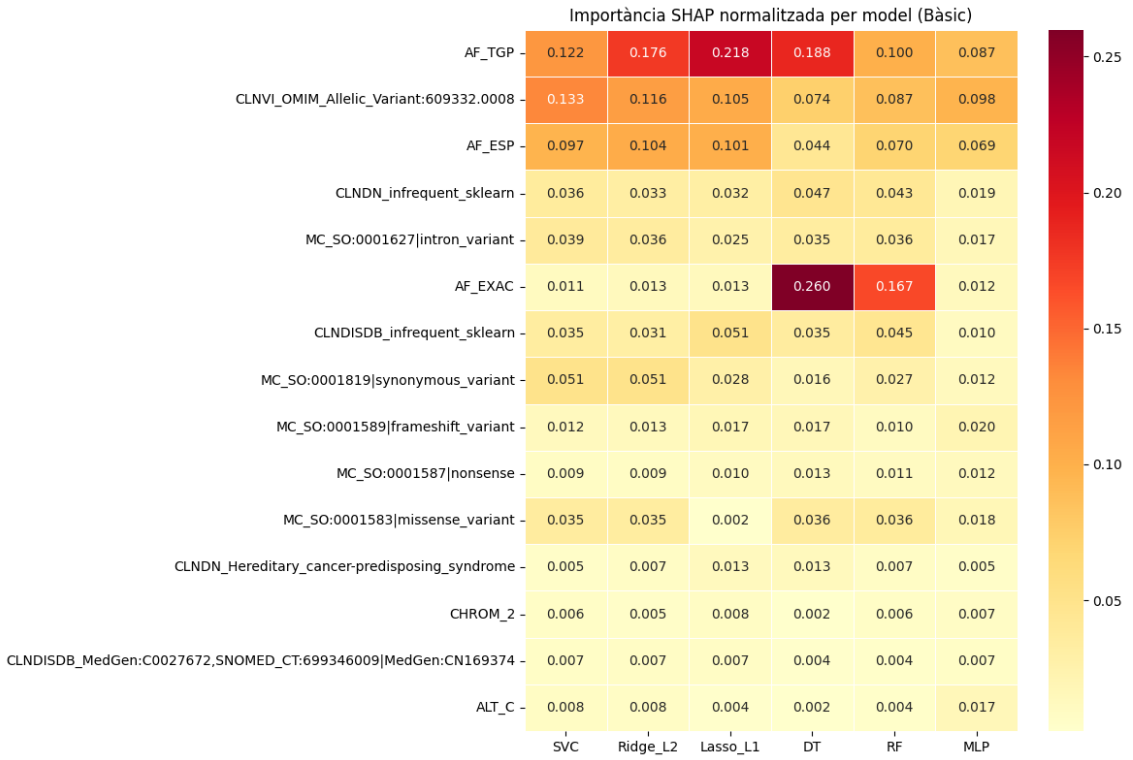


Figura 5.15: Importància SHAP normalitzada per model: top 15 variables més consistents entre els sis classificadors de tasca binària en dataset bàsic.

Les variables més consistents segueixen sent les freqüències al·lèliques `AF_TGP`, `AF_ESP`, `AF_EXAC` i `CLNVI_OMIM_Allelic_Variant:609332.0008`, mostrant que el bloc principal de predictors es manté independentment del conjunt de variables disponibles. Tanmateix, `AF_TGP` rep encara més pes en el dataset bàsic en gairebé tots els models, els models es recolzen més en les variables clíniques disponibles.

La diferència més destacable és l'aparició de noves variables relacionades amb el tipus de conseqüència molecular, `MC_SO:0001627|intron_variant`, `MC_SO:0001819|synonymous_variant`, `MC_SO:0001589|frameshift_variant` i `MC_SO:0001587|nonsense`. En el dataset anotat, aquesta informació era redundant amb les categories d'IMPACT de VEP i altres anotacions funcionals, en el dataset bàsic amb la falta d'aquestes variables, els models ho compensen aprofitant les anotacions de tipus de variant del camp MC, que descriuen el tipus de conseqüència molecular de la variant.

5.3 Classificació Multiclasse

En aquesta secció es presenten els resultats de la classificació multiclasse, on l'objectiu és predir la significança clínica d'una variant en tres categories: benigna, VUS i patogènica. Seguint la mateixa estructura que en la secció anterior, primer es mostren els resultats obtinguts amb el dataset anotat, després amb el dataset bàsic i finalment es compara el rendiment entre les dues configuracions per avaluar l'aportació de les anotacions funcionals VEP en aquesta tasca.

5.3.1 Dataset Anotat

La Taula 5.3 resumeix els resultats obtinguts pels sis models sobre el dataset anotat.

Model	CV F1 macro	Test F1 macro	Millors hiperparàmetres
LinearSVC	0.859	0.868	C=0.1
MLP	0.870	0.862	hidden=(64,32), lr=0.0005
Logística L2	0.862	0.868	C=0.1
Logística L1	0.854	0.862	C=10.0
Decision Tree	0.887	0.884	max_depth=20
Random Forest	0.904	**0.900**	n_est=100, max_features=0.3

Taula 5.3: Resultats de la classificació multiclasse sobre el dataset anotat.

A diferència del cas binari, els valors d'F1 macro són més elevats, situant-se entre 0.862 (MLP) i 0.900 (Random Forest) al conjunt de test. El Random Forest torna a ser el millor model, seguit del Decision Tree amb un F1 macro de 0.884. Els models lineals i el MLP obtenen resultats similars, al voltant de 0.86–0.87.

És interessant el comportament del Decision Tree, que en la tasca binària era el pitjor model i en la multiclasse esdevé el segon millor. A més, en aquesta tasca presenta una diferència mínima entre CV i Test (0.003 punts), molt inferior a l'observada en la tasca binària (0.038 punts). Això apunta que les categories clíniques tenen fronteres més estables i fàcilment separables mitjançant regles condicionals que les del problema binari.

La Figura 5.16 mostra la matriu de confusió del Random Forest, el millor model.

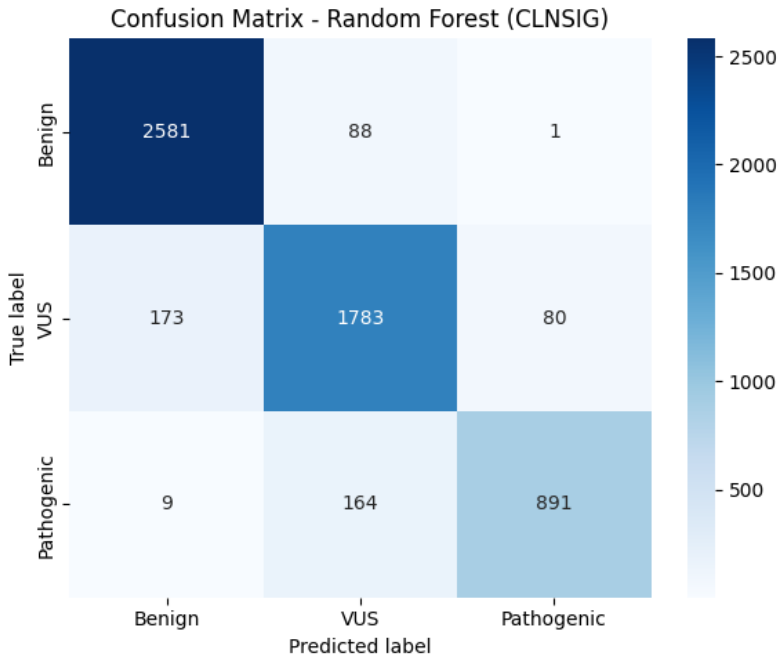


Figura 5.16: Matriu de confusió del Random Forest (dataset anotat, tasca multiclasse).

S'observa una bona discriminació entre les tres classes, 2.581 de 2.670 variants benignes són classificades correctament (97%), 1.783 de 2.036 VUS (88%) i 891 de

1.064 patogèniques (84%). La confusió més alta es produeix entre VUS i les classes adjacents, fet esperable ja que VUS és, per definició, una categoria intermèdia amb límits menys clars.

En contrast, la Figura 5.17 mostra la matriu de confusió del MLP, el model amb pitjor rendiment en aquesta tasca.

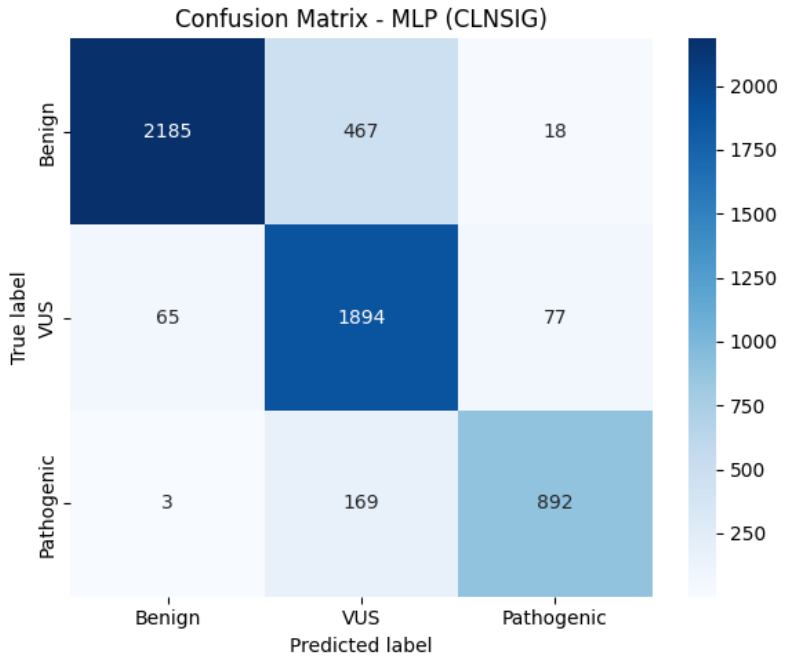


Figura 5.17: Matriu de confusió del MLP (dataset anotat, tasca multiclasse).

El MLP tendeix a classificar moltes variants com a VUS amb recall de 0.93, però amb una precisió baixa del 0,75, a costa d'un recall més baix per la classe benigna de 0,82. Aquest comportament és coherent amb l'absència de `class_weight=balanced` al MLP, en problemes desbalancejats amb una classe central VUS, el model tendeix a refugiar-se en aquesta classe quan no està segur, l'estratègia que minimitza la pèrdua però que penalitza el F1 macro.

Aquests resultats indiquen que els models basats en arbres són els més adequats per a aquesta tasca de classificació multiclasse, superant tant els models lineals com el MLP.

5.3.2 Dataset Bàsic

La Taula 5.4 resumeix els resultats obtinguts pels sis models sobre el dataset bàsic.

Model	CV F1 macro	Test F1 macro	Millors hiperparàmetres
LinearSVC	0.809	0.808	C=0.01, class_weight=balanced
MLP	0.759	0.742	hidden=(64,), lr=0.0001
Logística L2	0.816	0.820	C=0.1
Logística L1	0.789	0.804	C=0.01, class_weight=balanced
Decision Tree	0.871	0.869	max_depth=20, min_samples_split=20
Random Forest	0.884	**0.883**	n_est=100, max_depth=None, max_feat=0.3

Taula 5.4: Resultats de la classificació multiclasse sobre el dataset bàsic.

El Random Forest continua sent el millor model amb un F1 macro de 0.883, seguit pel Decision Tree amb 0.869. Els models lineals se situen al voltant de 0.80 i 0.82, mentre que MLP cau fins a 0.742, sent el pitjor model d'aquesta configuració.És especialment destacable que el MLP cau de 0.862 al dataset anotat a 0.742 al dataset bàsic, confirmant que aquesta arquitectura depèn de l'aportació funcional de les anotacions VEP.

Les Figures 5.18 i 5.19 mostren les matrius de confusió del millor i pitjor model.

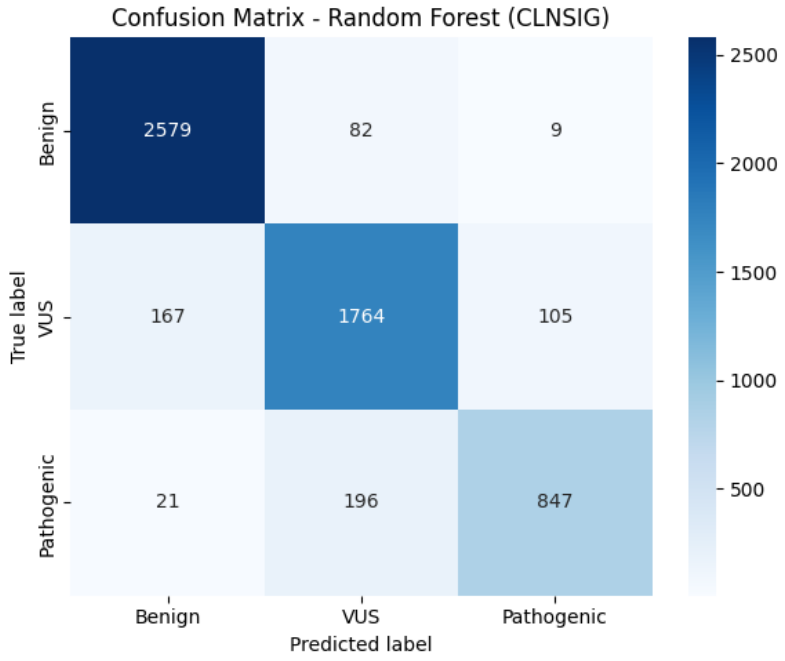


Figura 5.18: Matriu de confusió del Random Forest (dataset bàsic, tasca multiclasse).

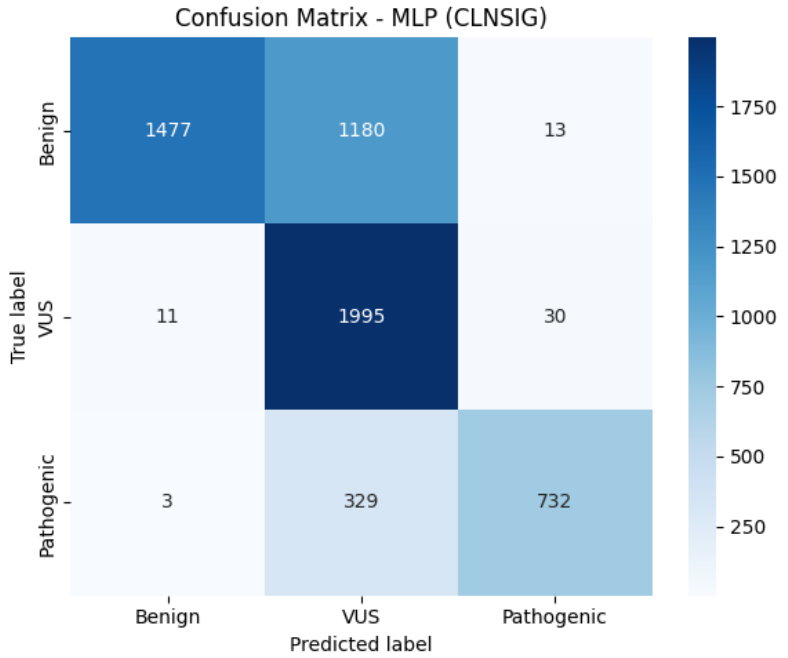


Figura 5.19: Matriu de confusió del MLP (dataset bàsic, tasca multiclasse).

El Random Forest manté una bona discriminació fins i tot sense les anotacions VEP, amb recalls del 97%, 87% i 80% per a Benign, VUS i Pathogenic respectivament. En canvi, el MLP mostra un patró desequilibrat, identifica només el 55% de les benignes i el 69% de les patogèniques, però classifica gairebé totes les VUS correctament (98%). Aquest comportament reflecteix dues limitacions, primer la reducció de variables dificulta la separació entre categories, i segon l'absència de `class_weight=balanced` que, com ja s'ha comentat, no suportat pel `MLPClassifier`, accentua el biaix cap a la classe intermèdia. Tot i això, els errors entre classes oposades (Benign ↔ Pathogenic) es mantenen baixos en tots dos models (30 i 16 casos respectivament), confirmant que la separabilitat entre els extrems de l'escala clínica és estable.

5.3.3 Comparació Anotat vs Bàsic

Per poder comparar l'aportació de les anotacions funcionals VEP en el problema de classificació multiclasse, s'ha comparat el rendiment dels sis models entrenats sobre el dataset anotat i sobre el dataset bàsic. La Figura 5.20 mostra els valors de F1 macro al conjunt de test per a cada model i configuració.

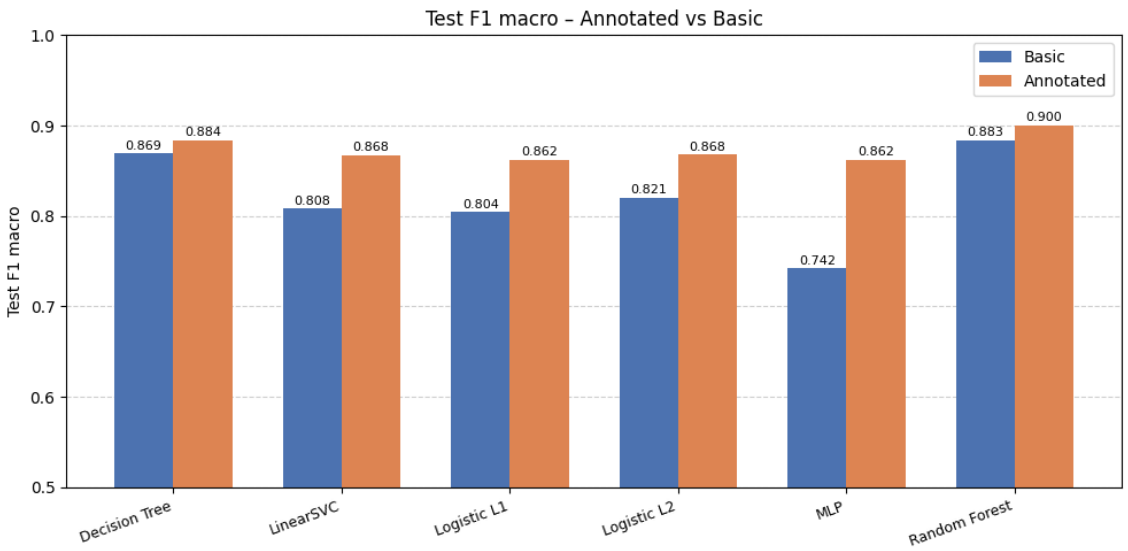


Figura 5.20: Comparació del Test F1 macro entre el dataset Anotat i el dataset Bàsic per a cada model (tasca multiclasse).

A diferència del cas binari, on les diferències eren més petites, en predicció multiclasse l'aportació de les anotacions funcionals és més marcada. Tots els models obtenen un F1 macro superior amb el dataset anotat, però la magnitud de la millora varia segons el tipus de model.

El MLP és el model que més guanya amb les anotacions funcionals, passant de 0.742 a 0.862. Aquesta diferència és notable i indica que el MLP és sensible a la quantitat i qualitat de les variables d'entrada. Amb el dataset bàsic té problemes per definir fronteres clares entre les tres classes, però amb les variables funcionals VEP recupera un rendiment competitiu.

Els models lineals també presenten millores. La logística L1 millora +0.058, la LinearSVC +0.060 i la logística L2 +0.047. Aquestes diferències són una mica més grans que les observades en el problema binari, on els models lineals només milloraven entre 0.004 i 0.006 punts.

En canvi, els models basats en arbres són els que menys necessiten les anotacions funcionals. El Random Forest millora únicament +0.017 i el Decision Tree +0.015. Aquest comportament indica que els arbres són capaços d'extreure la major part del senyal predictiu a partir de les variables clíniques bàsiques de ClinVar, i que les anotacions funcionals només aporten una millora marginal.

Aquesta diferència de comportament entre models és el resultat més rellevant d'aquesta comparació, el Random Forest i el Decision Tree ofereixen rendiments competitius fins i tot sense les anotacions funcionals, mentre que la resta de models depenen d'aquestes per aconseguir un millor rendiment.

5.4 Interpretabilitat Classificació Multiclasse

L'anàlisi d'interpretabilitat per a la tasca multiclasse presenta una diferència respecte al problema binari, ara cada model produeix valors SHAP separats per a cada una de les tres classes, Benign, VUS i Pathogenic, cosa que permet identificar quines variables empenyen cap a cada categoria clínica. Primer s'analitzen els resultats per

al dataset anotat i després el del dataset bàsic.

A diferència del cas binari, on el signe del valor SHAP indicava una única direcció, en el problema multiclasse cada model genera un conjunt de valors SHAP per a cada classe. Per això es presenta un gràfic amb un panell per classe. Dins de cada panell, el signe s'interpreta respecte a la classe corresponent, els valors positius (en blau) indiquen que la variable empeny la predicció cap a aquella classe, mentre que els valors negatius (en vermell) indiquen que l'empeny en contra. Així, una mateixa variable pot contribuir de manera positiva en una classe i negativa en una altra.

5.4.1 Models lineals

Les Figures 5.21, 5.22 i 5.23 mostren els valors SHAP signats dels tres models lineals avaluats, descomposats per classe.

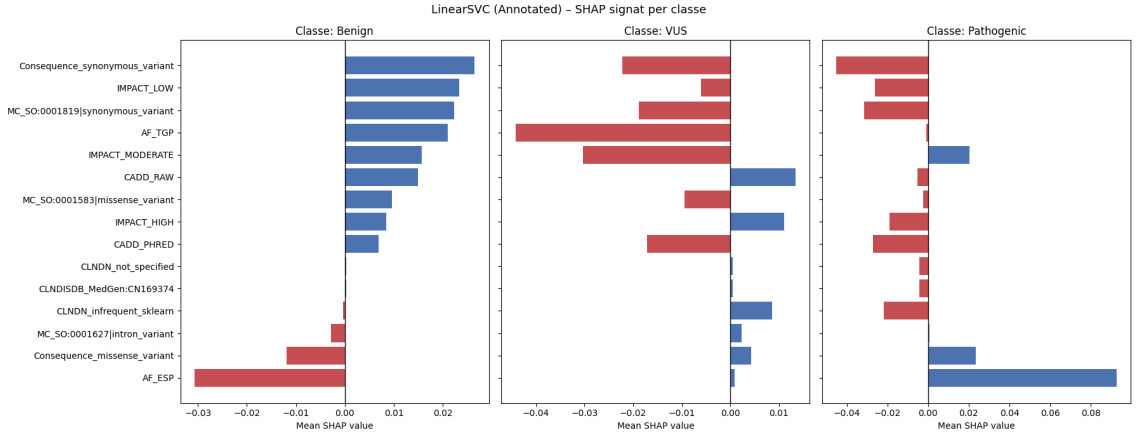


Figura 5.21: Valors SHAP signats del LinearSVC per classe (tasca multiclasse, dataset anotat).

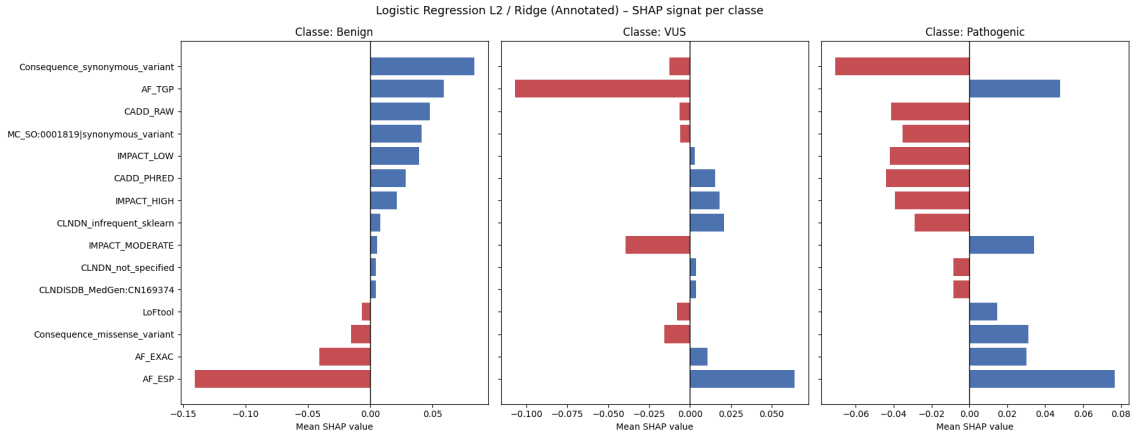


Figura 5.22: Valors SHAP signats de la Logística L2 Ridge per classe (tasca multiclasse, dataset anotat).

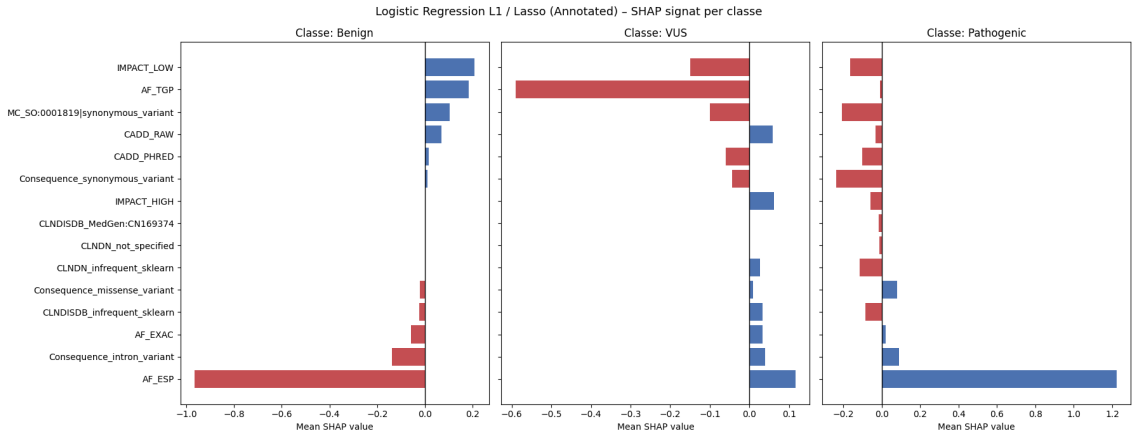


Figura 5.23: Valors SHAP signats de la Logística L1 Lasso per classe (tasca multiclasse, dataset anotat).

En tots tres models s'observa que les freqüències al·lèliques (`AF_TGP`, `AF_ESP` i `AF_EXAC`) tenen un paper destacat en la separació de classes. De manera predominant, valors elevats d'aquestes variables empenyen cap a la classe Benign, mentre que valors baixos contribueixen cap a Pathogenic, fet coherent amb el coneixement clínic [32]: les variants comunes a la població tendeixen a ser benignes, mentre que les rares tenen més probabilitat de ser patogèniques. La Logística L1 mostra el patró més extrem, conseqüència de la seva tendència a concentrar la importància en un subconjunt reduït de variables.

També destaquen variables relacionades amb el tipus de conseqüència molecular com `Consequence_synonymous_variant` i `MC_SO:0001819|synonymous_variant` que empenyen cap a la classe Benign, fet coherent ja que les variants sinònimes no alteren la seqüència proteica i solen tenir un impacte clínic limitat. En canvi les categories IMPACT amb HIGH, MODERATE i LOW tenen un comportament diferent segons cada model, contribuint diferentment a les tres classes en funció de la severitat predita de la conseqüència molecular [10].

5.4.2 Decision Tree

A diferència de la predicció binària, on l'arbre dividia inicialment per la freqüència al·lèlica, en la predicció multiclasse la primera divisió es basa en `IMPACT_HIGH` com es pot veure en la Figura 5.24. Aquest canvi reflecteix la naturalesa diferent del problema, la severitat predita de la conseqüència molecular és el criteri més discriminatiu per separar variants patogèniques de la resta.

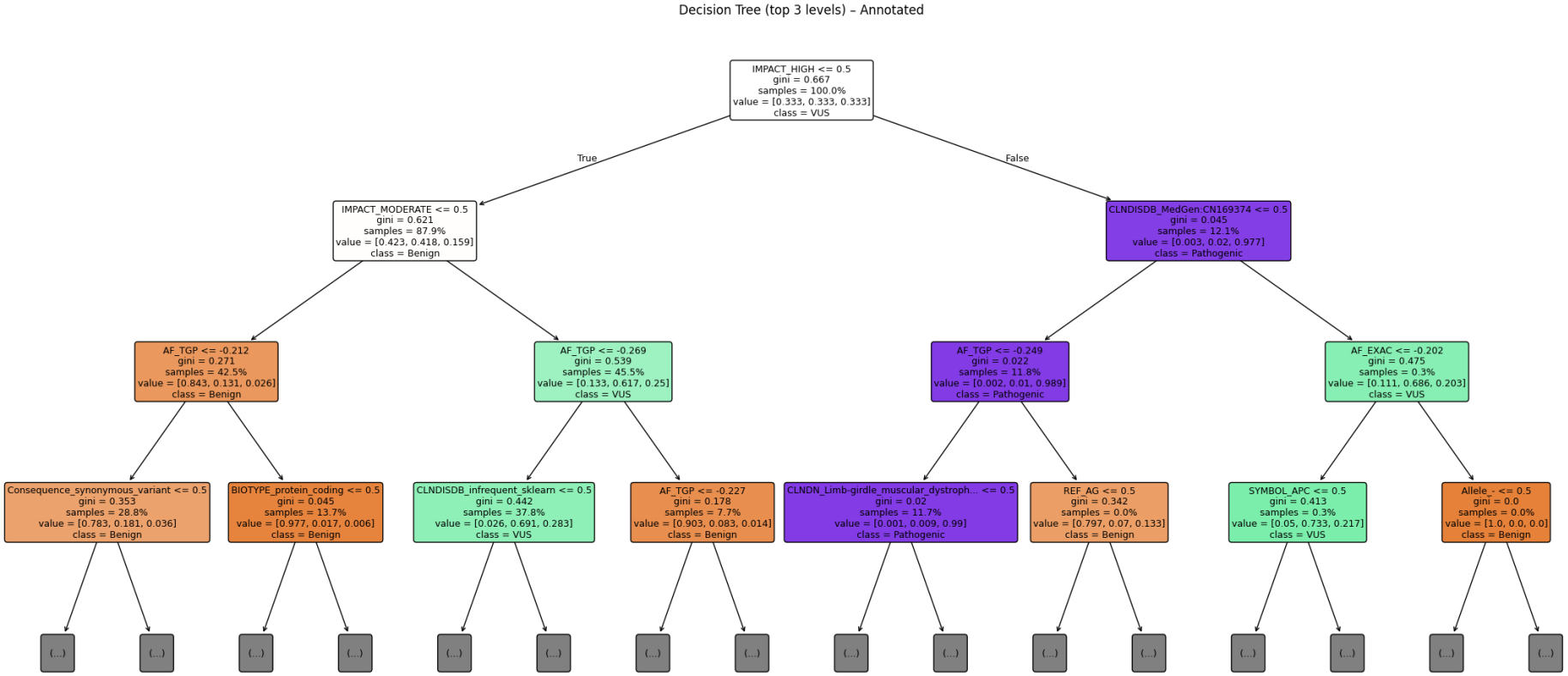


Figura 5.24: Estructura dels tres primers nivells del Decision Tree (tasca multiclasse, dataset anotat).

Quan la variant té IMPACT_HIGH (és a dir, la branca dreta de l'arbre, on la condició IMPACT_HIGH $\leq$ 0.5 és falsa), el node resultant ja està format per variants patogèniques amb un 97.7% de les mostres d'aquesta branca pertanyen a la classe Pathogenic. L'arbre encara la subdivideix per l'identificador clínic CLNDISDB_MedGen:CN169374 per afinar la separació. En canvi, quan IMPACT_HIGH no es compleix, branca esquerra, l'arbre divideix per IMPACT_MODERATE i després per la freqüència al·lèlica AF_TGP.

L'anàlisi visual de l'arbre revela que combinant les categories d'IMPACT amb les freqüències al·lèliques i alguns identificadors clínics específics, el model pot separar les tres classes amb una puresa elevada en nivells profunds, explicant així el bon rendiment del Decision Tree.

5.4.3 Random Forest

La Figura 5.25 mostra els valors SHAP signats del Random Forest descomposats per classe.

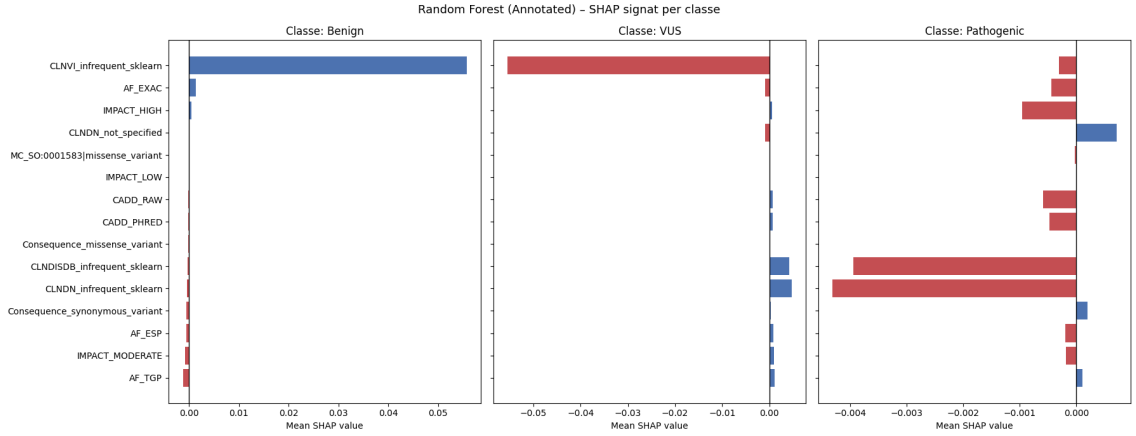


Figura 5.25: Valors SHAP signats del Random Forest per classe (tasca multiclasse, dataset anotat).

El comportament del Random Forest difereix de l'observat en els models lineals. La variable que contribueix més és CLNVI_infrequent_sklearn, que agrupa identificadors de variants poc freqüents al dataset, amb una contribució positiva cap a la classe Benign i negativa cap a la classe VUS. Aquest patró indica que el model ha après que les variants documentades amb identificadors poc freqüents, és a dir, sense un codi CLNVI específic conegut, tendeixen a ser benignes.

També destaquen les variables CLNDISDB_infrequent_sklearn i CLNDN_infrequent_sklearn, que mostren un comportament similar, contribució positiva cap a VUS i negativa cap a Pathogenic. La presència recurrent d'aquestes categories agrupades, que provenen del comportament de l'OneHotEncoder, suggereix que el Random Forest utilitza intensament la informació de quina cobertura clínica té cada variant com a senyal predictiu.

Pel que fa a les variables biològiques, la importància global del Random Forest, calculada com a mitjana absoluta entre classes, confirma que les variables més rellevants són IMPACT_MODERATE, AF_TGP, CADD_RAW i IMPACT_HIGH.

5.4.4 MLP

La Figura 5.26 mostra els valors SHAP signats del MLP descomposats per classe.

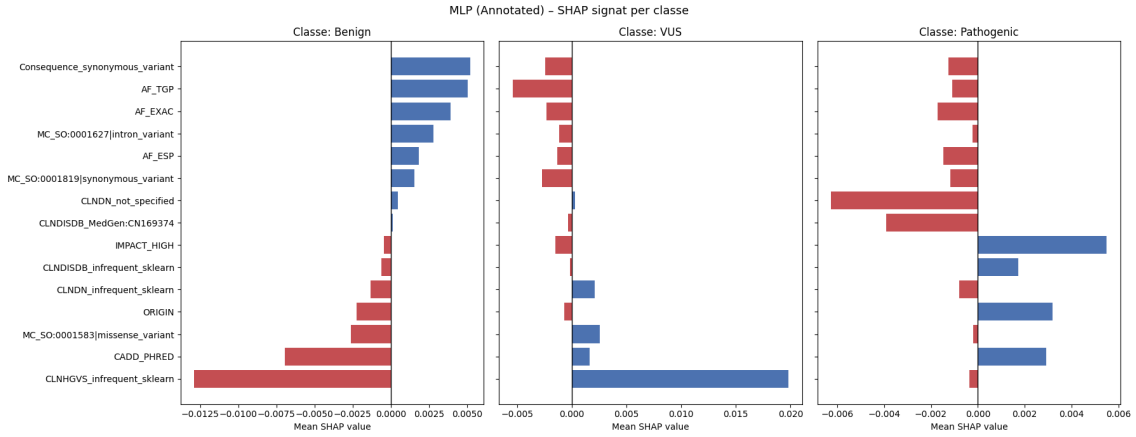


Figura 5.26: Valors SHAP signats del MLP per classe (tasca multiclasse, dataset anotat).

Com es pot veure en la Figura 5.26 la variable `Consequence_synonymous_variant` contribueix positivament a la classe Benign, fet esperable ja que, com s'ha comentat anteriorment, les variants sinònimes no alteren la seqüència proteica i tendeixen a ser benignes. Les freqüències al·lèliques `AF_TGP`, `AF_ESP` i `AF_EXAC` també empenyen cap a Benign quan són elevades, confirmant el patró vist en els models lineals, les variants comunes a la població solen ser benignes.

Per a la classe Pathogenic, les variables més rellevants són `IMPACT_HIGH`, amb una contribució positiva i `ORIGIN`, que indica l'origen de la submissió. Variants amb impacte molecular alt com canvis de fase de lectura o pèrdua del codó d'inici solen ser patogèniques. També cal destacar el patró oposat amb `CLNDN_not_specified` i `CLNDISDB_MedGen:CN169374` que contribueixen negativament a Pathogenic.

Per a la classe VUS, la variable més influent és `CLNHGVS_infrequent_sklearn` que agrupa identificadors HGVS poc freqüents en el dataset. Aquest comportament indica que el MLP associa la falta de documentació clínica detallada amb la incertesa en la classificació, fet consistent amb la definició de VUS com a variants amb evidència insuficient.

5.4.5 Concordança entre models

De la mateixa manera que en el problema binari, per identificar les variables importants entre tots els models s'ha calculat la importància SHAP normalitzada per a cada característica i model. La Figura 5.27 mostra el *heatmap* resultant de les 15 variables més consistents.



Figura 5.27: Importància SHAP normalitzada per model: top 15 variables més consistents entre els sis classificadors (tasca multiclasse, dataset anotat).

La variable més consistent torna a ser `AF_TGP` amb importàncies normalitzades entre 0.065 en LinearSVC i 0.145 en Decision Tree confirmant el paper central de les freqüències al·lèliques a la classificació clínica de variants. `AF_ESP` segueix com a segona variable més rellevant, amb una importància especialment elevada en Regressió Logística Lasso amb 0.251.

`IMPACT_HIGH` apareix com la tercera variable més consistent, amb un valor especialment alt al Decision Tree amb 0.182. Aquest comportament confirma el que ja s'havia observat a l'arbre de Decision Tree, aquestes variables era la primera divisió i, per tant, concentra un gran poder discriminatiu.

També cal destacar variables relacionades amb el tipus de conseqüència molecular com `Consequence_synonymous_variant` i puntuacions de perjudicialitat com `CADD_PHRED` i `CADD_RAW`, que apareixen amb importància moderada en tots els models.

Comparant amb el consens de la tasca binària, s'observen diferències en l'ordre de prioritat. Mentre que en el problema binari la variable més rellevant era `AF_TGP` seguida de `AF_ESP` i `CLNVI_OMIM_Allelic_Variant:609332.0008`, en la multiclasse la tercera posició passa a ser `IMPACT_HIGH`, indicant que la severitat predita de la conseqüència molecular és més útil per distingir entre les tres categories clíniques que no pas detectar conflictes entre laboratoris. La consistència entre models reforça la fiabilitat d'aquests resultats i mostra que els classificadors de naturalesa molt diferent identifiquen els mateixos predictors clau.

5.4.6 Comparació amb el dataset Bàsic

L'anàlisi d'interpretabilitat sobre el dataset bàsic mostra similituds amb les del dataset anotat, però amb diferències degudes a l'absència de les anotacions funcionals VEP. Els resultats detallats per a cada model es presenten a l'Apèndix B; en aquesta secció es comenten únicament les diferències més importants.

La Figura 5.28 mostra el *heatmap* de consens entre els sis models per al dataset bàsic.

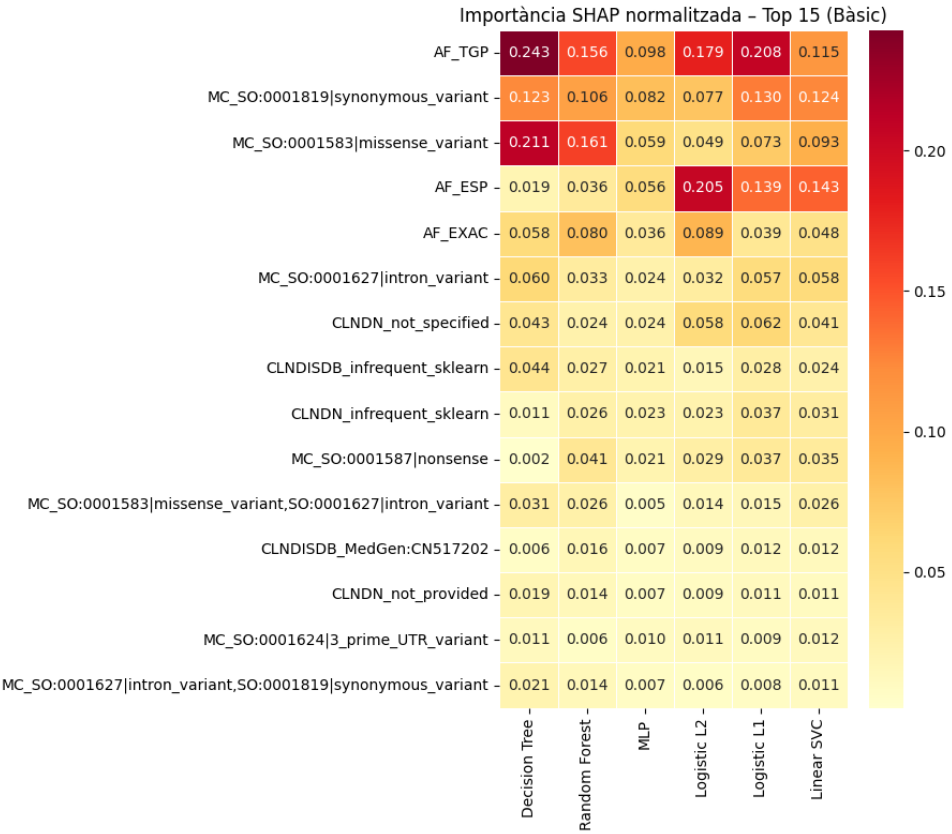


Figura 5.28: Importància SHAP normalitzada per model: top 15 variables més consistents entre els sis classificadors (tasca multiclasse, dataset bàsic).

La variable més consistent continua sent la freqüència al·lèlica `AF_TGP`, la seva importància de la qual augmenta respecte el dataset anotat, com per exemple en el Decision Tree passa de 0.145 a 0.243. En absència d'anotacions funcionals com `CADD`, `IMPACT` o `LoFtool`, els models compensen aprofitant més les variables clíniques disponibles.

També destaca l'aparició com a variables més importants de categories relacionades amb el tipus de conseqüència molecular del camp MC, com `MC_SO:0001819|synonymous_variant`, `MC_SO:0001583|missense_variant` i `MC_SO:0001587|nonsense`. En particular, `MC_SO:0001583|missense_variant` assoleix valors elevats en el Decision Tree i Random Forest, actuant com a substitut de les categories d'IMPACT de VEP. Els models infereixen la severitat de la conseqüència molecular a partir del camp MC del VCF original, que era redundant amb les anotacions de VEP.

Tot i aquestes diferències, el bloc de les variables més importants com les fre-

qüències al·lèliques i el tipus de conseqüència molecular es manté estable en les dues configuracions, fet que reforça la robustesa del senyal predictiu i confirma que les variables bàsiques de ClinVar ja contenen gran part de la informació necessària per a la classificació multiclasse.

5.5 Discussió General

Aquesta secció posa en context els resultats obtinguts en les dues tasques de classificació i els compara amb treballs previs, també es discuteixen les implicacions clíniques dels resultats.

5.5.1 Comparació entre prediccions

Els resultats mostren que els models proposats assoleixen un rendiment acceptable en les dues tasques. En la classificació binària, el Random Forest assoleix un ROC AUC de 0,827, valor que indica una capacitat discriminativa raonable en un problema desbalancejat. En la classificació multiclasse, el mateix model obté un F1-Score macro de 0,900, mostrant que separa bé les tres categories clíniques. Cal subratllar, però, que aquests valors corresponen a mètriques diferents i no són directament comparables.

La predicció de conflictes depèn no només de les característiques biològiques de la variant, sinó també de factors externs com els criteris d'interpretació de cada laboratori o l'evidència disponible en el moment de la submissió [19], factors que limiten el rendiment assolible per models basats només amb característiques biològiques. La significança clínica, en canvi, està més relacionada amb propietats mesurables com l'impacte funcional o la freqüència poblacional.

5.5.2 Comparació amb treballs previs

Abans d'entrar en la comparació, cal subratllar que les particions de dades, els criteris de filtratge i les versions de ClinVar utilitzades en els treballs citats no són idèntics als d'aquest treball. La discussió s'ha d'entendre com una referència orientativa en problemes similars.

El treball més proper al present és el de Ghaedi et al. [14], que també va realitzar la predicció de conflictes a ClinVar mitjançant *XGBoost*. Els autors reporten al conjunt de test un *accuracy* del 88% amb valors ponderats de precisió, recall i F1-Score de 0,84, 0,88 i 0,85. La comparació directa amb el ROC AUC de 0,827 no és immediata, però l'observació conjunta confirma la dificultat d'aquesta tasca.

Nicora et al. [29] van assolir una precisió del 98% combinant criteris ACMG (descrits a la Secció 2.2) amb anotacions funcionals en un model de regressió logística. Aquest resultat és bastant superior al d'aquest treball, però els mateixos autors adverteixen que la distribució de classes a ClinVar pot no reflectir la realitat clínica i que la seva mètrica de precisió podria estar inflada pel desbalanceig de classes. En el present treball, l'ús del F1 macro com a mètrica principal mitiga aquest problema, donant resultats més conservadors però més representatius del rendiment real.

5.5.3 Aportació de les anotacions funcionals VEP

La comparació entre els conjunts de dades anotat i bàsic revela un patró asimètric, en el problema binari la millora aportada per les anotacions funcionals és modesta, entre +0.004 i +0.028 punts d'AUC, mentre que en el problema de multiclasse aquesta millora és major fins a +0.120 punts. Aquest comportament és coherent amb la naturalesa de les anotacions funcionals. Predictors com SIFT [11], PolyPhen [2] i CADD [31] estan dissenyats específicament per estimar l'impacte funcional d'una variant sobre la proteïna, informació directament relacionada amb la significança clínica però només indirectament amb la presència de conflictes entre laboratoris.

Els models basats en arbres són els que menys depenen d'aquestes anotacions, mantenint un bon rendiment amb el dataset bàsic. Aquest comportament evidencia que les variables clíniques de ClinVar ja contenen una part important del senyal predictiu, i que els arbres són especialment efectius extraient aquest senyal. Els models lineals i el MLP, en canvi, necessiten les anotacions funcionals per assolir un rendiment competitiu.

5.5.4 Implicacions clíniques

Des d'una, perspectiva clínica, els resultats tenen dues lectures. D'una banda la capacitat de predir significança clínica amb un F1 macro de 0.900 obre la possibilitat d'utilitzar aquests models com a eines de suport a la decisió en laboratoris de diagnòstic genètic, especialment per a variants amb poques submissions o classificacions discrepants. D'altra banda, la dificultat per predir conflictes suggereix que aquest problema requereix d'informació addicional no continguda al dataset, com el context històric de les submissions o l'evidència experimental disponible per a cada variant.

En cap cas els models presentats haurien de substituir la interpretació humana experta, sinó complementar-la especialment per prioritzar variants que mereixen una revisió manual més acurada o per detectar discrepàncies interpretatives entre laboratoris abans que es propaguin a informes clínics.

Capítol 6

Conclusions

Aquest treball ha presentat un estudi sistemàtic de l'aplicació de tècniques d'aprenentatge automàtic a la classificació clínica de variants genètiques, abordant dos problemes complementaris, tant la predicció de conflicte entre laboratoris com la predicció de significança clínica. S'han comparat sis models de classificació sobre dues configuracions del dataset de ClinVar, i s'ha aplicat un anàlisi d'interpretabilitat amb valors SHAP per identificar els factors més rellevants en les prediccions. Els resultats obtinguts demostren la viabilitat de combinar models predictius amb tècniques d'interpretabilitat per complementar la interpretació humana de variants genètiques.

6.1 Conclusions principals

Les conclusions principals d'aquest estudi són:

1. **Superioritat dels models no lineals.** El Random Forest ha estat el millor model en totes les configuracions. Els models lineals han mostrat un rendiment inferior demostrant que no són problemes linealment separables.

2. **Contribució asimètrica de les anotacions VEP.** L'aportació de les anotacions funcionals VEP varien segons la tasca i el model. En la predicció binària, el dataset anotat aporta una millora consistent però moderada. En canvi, en la predicció multiclasse l'efecte és més asimètric, en el MLP cau sense anotacions VEP, mentre que el Random Forest quasi no pateix variació. Tot això indica que les xarxes neuronals depenen molt més de la riquesa informativa de les variables d'entrada, mentre que els arbres són més robustos a configuracions amb menys variables.

3. **Consistència en les variables més rellevants.**L'anàlisi d'interpretabilitat amb SHAP ha revelat que les variables més importants són consistents entre models molt diferents. Les freqüències al·lèliques són el factor més determinant, reforçant la fiabilitat dels patrons identificats, més enllà de les particularitats algorísmiques de cada classificador.

4. **Rellevància metodològica de la interpretabilitat.** L'aplicació de SHAP no només ha permès identificar les variables més importants, sinó que també ha aportat valor de validació, la concordança entre els valors SHAP de models diferents dona una verificació independent de l'aprenentatge real del sistema.

6.2 Limitacions

Cal destacar diverses limitacions del treball. En primer lloc, el dataset utilitzat correspon a una versió de ClinVar d'abril de 2018, i és possible que la distribució actual de classificacions hagi canviat. En segon lloc, l'agrupació de la variable CLNSIG en tres categories implica una pèrdua d'informació respecte als nivells originals, tot i que aquesta simplificació era necessària per evitar un desbalanceig extrem entre classes.

6.3 Treball futur

A partir dels resultats obtinguts i de les limitacions detectades es proposen tres línies principals de treball futur:

1. **Actualització del dataset.** Replicar l'anàlisi amb una versió més recent de ClinVar permetria avaluar com han evolucionat les distribucions de classes i de conflictes des de 2018, així com aprofitar el volum més gran de variants classificades disponibles avui dia.

2. **Incorporació de gradient boosting.** Afegir XGBoost i altres algorismes de gradient boosting a la comparació permetria contrastar el Random Forest amb mètodes que han demostrat un rendiment superior en problemes similars.

3. **Tractament diferenciat de la classe VUS.** Explorar enfocaments alternatius per a la categoria VUS, com tècniques de classificació ordinal o models específicament dissenyats per categories d'incertesa, que podrien reflectir millor la naturalesa clínicament gradual d'aquesta categoria.

Bibliografia

[1] 1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, and G. R. Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[2] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev. A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249, 2010.

[3] Kevin Arvai. Clinvar conflicting interpretations, 2018.

[4] Kevin Arvai. clinvar-kaggle: scripts to process clinvar vcf, 2018.

[5] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[6] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

[7] Computational Genomics with R. Logistic regression and regularization, 2023. Accessed: 2026-05-27.

[8] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[9] Ensembl. Ensembl variant effect predictor (vep), 2025.

[10] Ensembl. Ensembl variation - calculated variant consequences, 2025.

[11] Ensembl. Ensembl variation - pathogenicity predictions, 2025.

[12] J. Fadista, N. Oskolkov, O. Hansson, and L. Groop. Loftool: a gene intolerance score based on loss-of-function variants in 60,706 individuals. *Bioinformatics*, 33(4):471–474, 2017.

[13] T. Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[14] Hamid Ghaedi, Scott K. Davey, and Harriet Feilotter. Variant classification discordance: Contributing factors and predictive models. *The Journal of Molecular Diagnostics*, 26(2):115–126, 2024.

[15] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, et al. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020.

[16] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

[17] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[18] E. S. Lander, L. M. Linton, B. Birren, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[19] M. J. Landrum, S. Chitipiralla, G. R. Brown, et al. Clinvar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1):D1062–D1067, 2018.

[20] M. Lek, K. J. Karczewski, E. V. Minikel, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291, 2016.

[21] S. Lundberg and S. Lee. Shap: Shapley additive explanations, 2017.

[22] S. M. Lundberg, G. Erion, H. Chen, et al. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020.

[23] S. M. Lundberg and S. I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

[24] W. McKinney. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, pages 56–61, 2010.

[25] Christoph Molnar. *Interpretable Machine Learning*. 3 edition, 2022.

[26] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010.

[27] National Center for Biotechnology Information. Clinvar, 2024. Accedit el juny de 2026.

[28] P. C. Ng and S. Henikoff. Sift: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814, 2003.

[29] G. Nicora, S. Zucca, I. Limongelli, R. Bellazzi, and P. Magni. A machine learning approach based on acmg/amp guidelines for genomic variant classification and prioritization. *Scientific Reports*, 12(1):2517, 2022.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[31] P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, and M. Kircher. Cadd: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, 47(D1):D886–D894, 2019.

[32] S. Richards, N. Aziz, S. Bale, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genetics in Medicine*, 17(5):405–424, 2015.

[33] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[34] Scikit-learn developers. Linear models: Ridge regression and classification, 2025. Accessed: 2026-05-27.

[35] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th ICML*, 2017.

[36] M. L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.

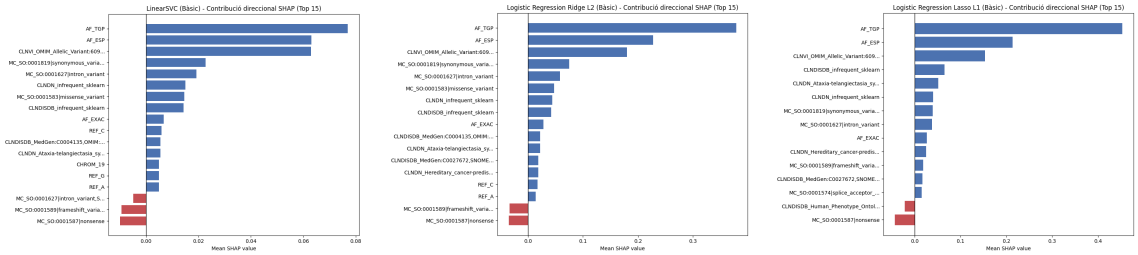
Apèndix A

Resultats SHAP per al dataset bàsic en el problema binari

Aquest apèndix recull els resultats detallats de l'anàlisi d'interpretabilitat sobre el dataset bàsic per la predicció de classificació binària. Els resultats segueixen el mateix patró qualitatiu que els del dataset anotat vist en el Capítol 5, per la qual cosa es mostren aquí com a referència complementària.

A.1 Models lineals

Els tres models lineals (LinearSVC, Ridge L2 i Lasso L1) mostren un patró molt similar sobre el dataset bàsic, amb AF_TGP, AF_ESP i CLNVI_OMIM_Allelic_Variant:609332.0008 com a variables més rellevants i contribucions positives cap a la classe Conflict.



(a) LinearSVC (b) Logística L2 (Ridge) (c) Logística L1 (Lasso)

Figura A.1: Valors SHAP signats dels tres models lineals (dataset bàsic).

A.2 Decision Tree

L'Arbre de Decisió sobre el dataset bàsic divideix novament a la primera capa per AF_TGP, confirmant el paper central d'aquesta variable. En els nivells inferiors, les divisions s'aprofiten d'AF_EXAC i de categories CLNVI i CLNDN.

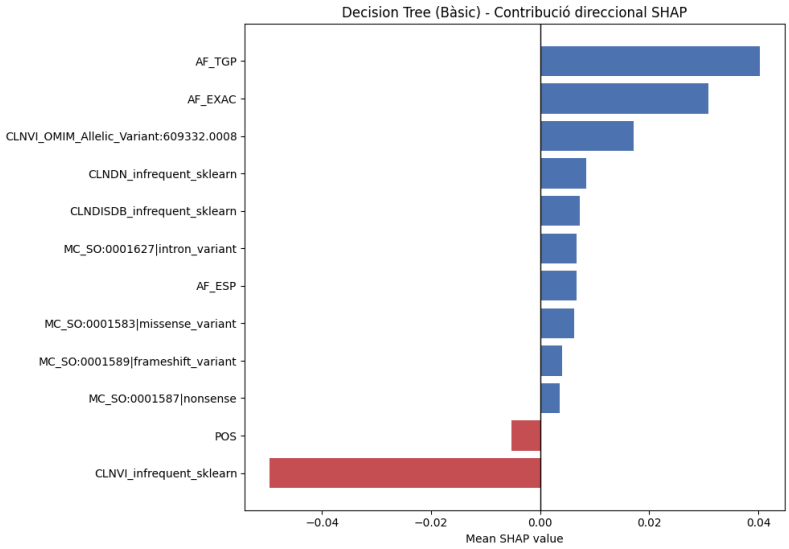


Figura A.2: Valors SHAP signats del Decision Tree (dataset bàsic).

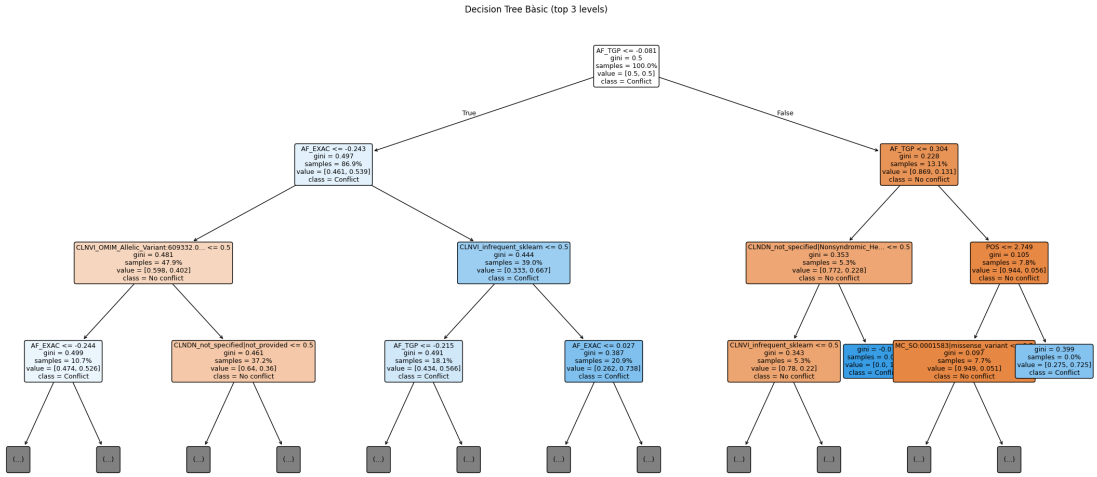


Figura A.3: Estructura dels tres primers nivells del Decision Tree (dataset bàsic).

A.3 Random Forest

El Random Forest sobre el dataset bàsic mostra un patró dominat per les variables de freqüència al·lèlica, especialment AF_EXAC, AF_TGP i AF_ESP, que presenten contribucions negatives cap a la classe Conflict.

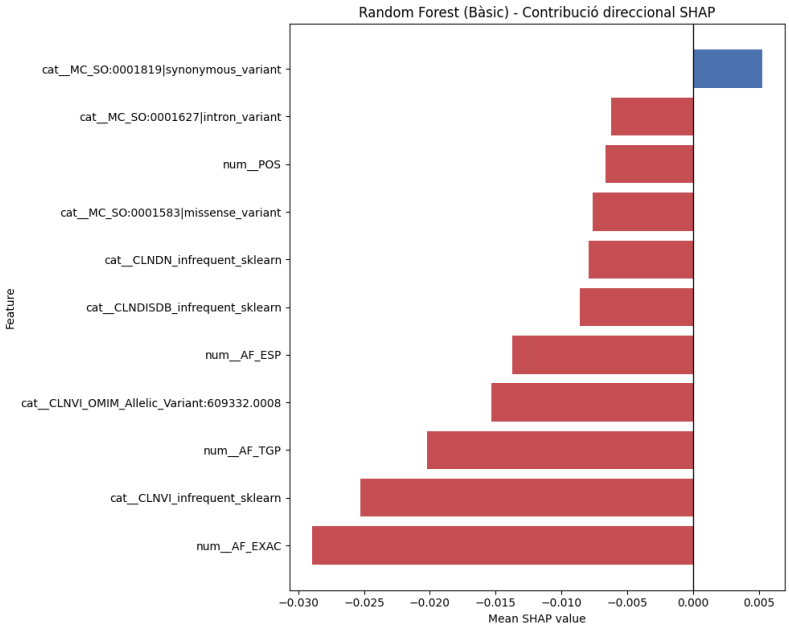


Figura A.4: Valors SHAP signats del Random Forest (dataset bàsic).

A.4 MLP

El MLP sobre el dataset bàsic mostra un patró diferent als altres models: la variable CLNHGVS_infrequent_sklearn (categoria que agrupa identificadors HGVS poc freqüents) emergeix com la més important, fet que no s'observa als altres models ni a la configuració anotada. Aquest comportament reflecteix la capacitat del MLP de capturar patrons no lineals diferents segons les variables disponibles.

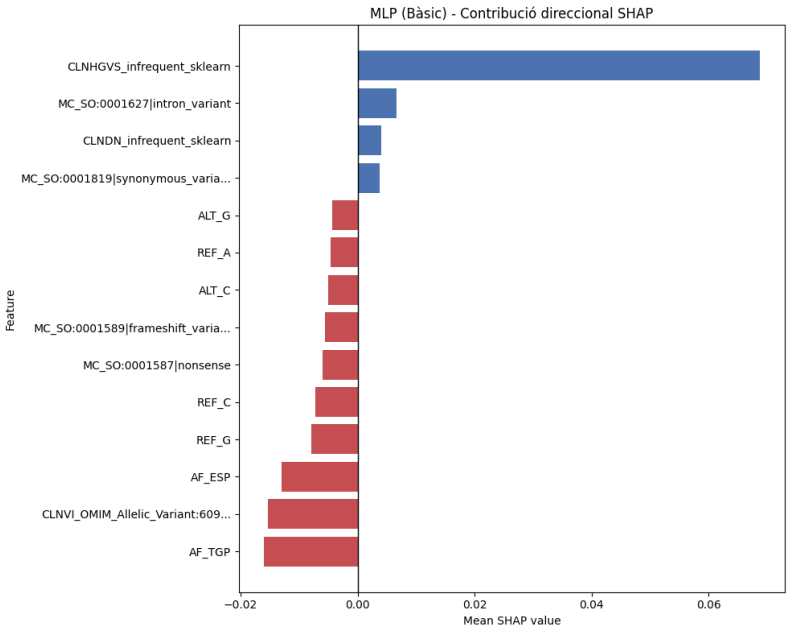


Figura A.5: Valors SHAP signats del MLP (dataset bàsic).

Apèndix B

Resultats SHAP per al dataset bàsic en el problema multiclasse

Aquest apèndix recull els resultats detallats de l'anàlisi d'interpretabilitat sobre el dataset bàsic per a la tasca de classificació multiclasse.

B.1 Models lineals

Els tres models lineals sobre el dataset bàsic multiclasse mostren patrons similars als observats al dataset anotat, amb les freqüències al·lèliques i el tipus de conseqüència molecular com a variables principals. Tanmateix, la contribució de variables MC és més gran en absència d'anotacions VEP.

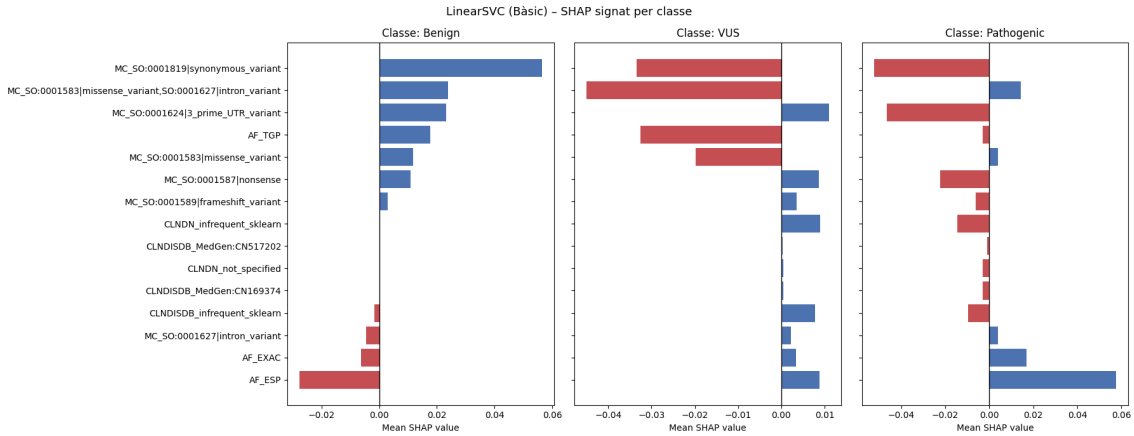


Figura B.1: Valors SHAP signats del LinearSVC per classe (dataset bàsic, tasca multiclasse).

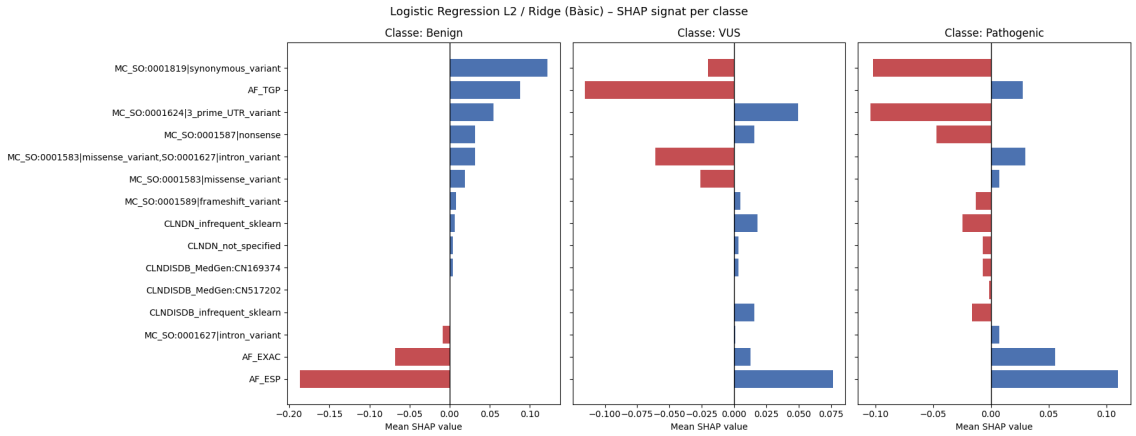


Figura B.2: Valors SHAP signats de la Logística L2 Ridge per classe (dataset bàsic, tasca multiclasse).

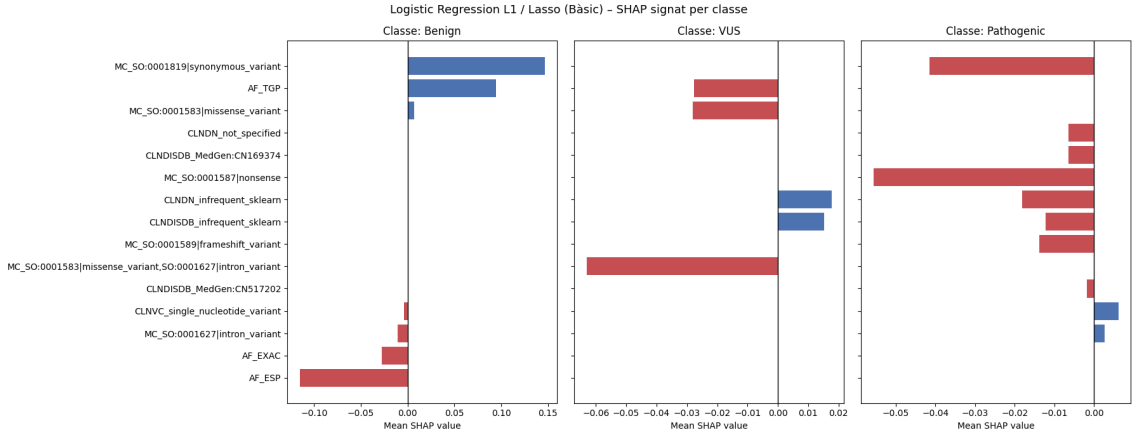


Figura B.3: Valors SHAP signats de la Logística L1 Lasso per classe (dataset bàsic, tasca multiclasse).

B.2 Decision Tree

L'Arbre de Decisió sobre el dataset bàsic continua dividint per AF_TGP a la primera capa, però els nivells inferiors aprofiten més intensament les variables MC.

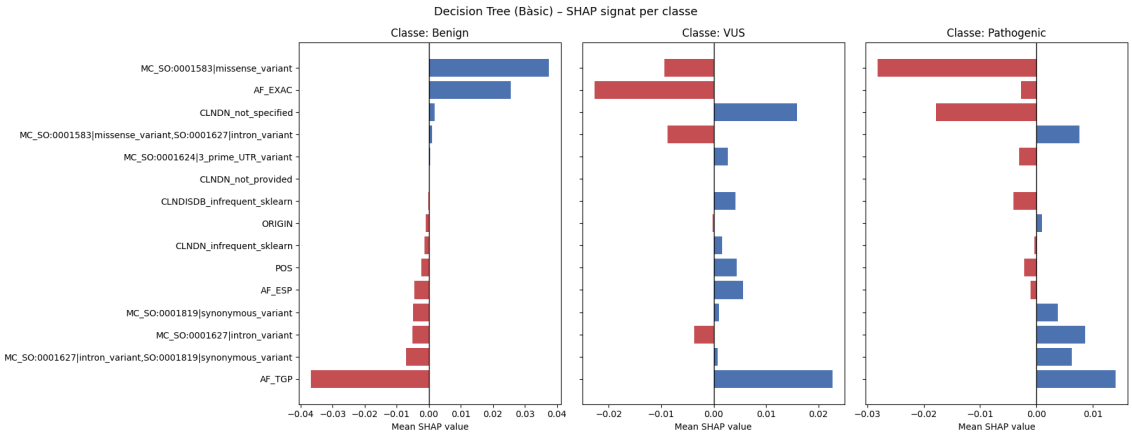


Figura B.4: Valors SHAP signats del Decision Tree per classe (dataset bàsic, tasca multiclasse).

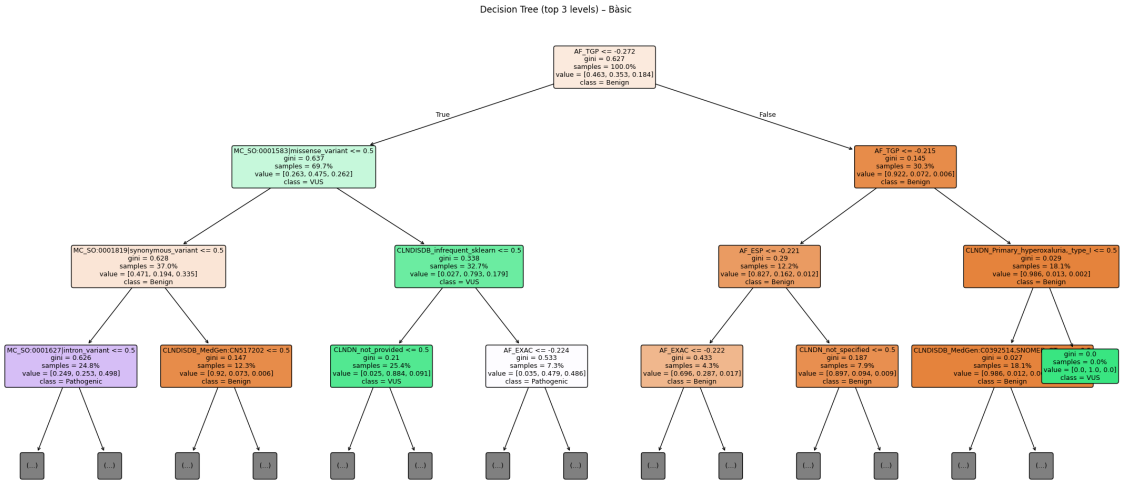


Figura B.5: Estructura dels tres primers nivells del Decision Tree (dataset bàsic, tasca multiclasse).

B.3 Random Forest

El Random Forest sobre el dataset bàsic distribueix la importància entre un conjunt més ampli de variables, amb les freqüències al·lèliques i les variables MC com a contribuïdores principals en absència d'anotacions VEP.

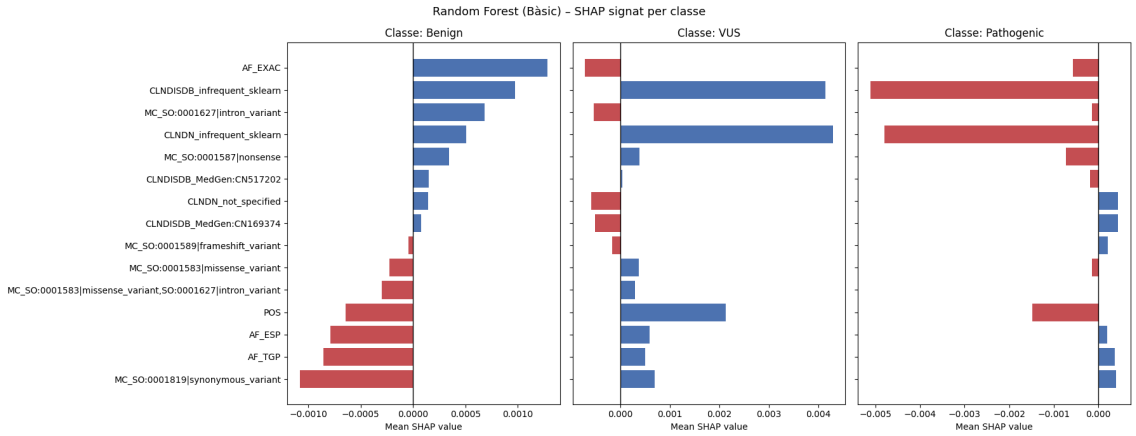


Figura B.6: Valors SHAP signats del Random Forest per classe (dataset bàsic, tasca multiclasse).

B.4 MLP

El MLP sobre el dataset bàsic manté les freqüències al·lèliques com a variables més influents, amb una distribució de la importància entre classes coherent amb la resta de models.

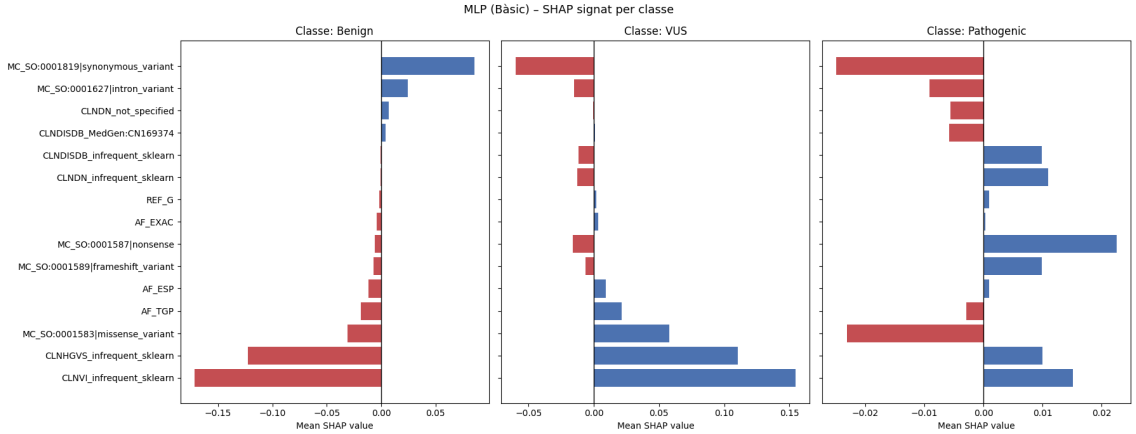


Figura B.7: Valors SHAP signats del MLP per classe (dataset bàsic, tasca multiclasse).