Anàlisi de Classificadors per a la Detecció de Malalties Hepàtiques en el Dataset ILPD

Joel Márquez Álvarez
Grau en Ciència i Enginyeria de Dades
Universitat Politècnica de Catalunya
Barcelona, Spain
joel.marquez@estudiantat.upc.edu

Rebeca Torrecilla Domínguez
Grau en Ciència i Enginyeria de Dades
Universitat Politècnica de Catalunya
Barcelona, Spain
rebeca.torrecilla@estudiantat.upc.edu

***Abstract*—Aquest treball aborda el problema de classificació de pacients amb malaltia hepàtica a partir del dataset Indian Liver Patient (ILPD), presentat com una competició a la plataforma Kaggle InClass. El principal repte d'aquest conjunt de dades és el fort desequilibri entre les classes (pacients malalts i sans). Per resoldre'l, s'ha implementat un pipeline complet que inclou una anàlisi exploratòria de dades (EDA), enginyeria de característiques per reduir la multicol·linealitat, i un preprocessament robust amb transformació logarítmica i escalat mitjançant *RobustScaler* per minimitzar l'efecte dels outliers. La gestió del desequilibri s'ha realitzat amb la tècnica de sobremostreig SMOTE. S'han avaluat diversos models, incloent Regressió Logística, Extra Trees, Random Forest, SVC, KNN, i ensembles com *VotingClassifier* i *StackingClassifier*. Finalment, l'ExtraTreesClassifier, amb una optimització del llindar de decisió, ha demostrat ser el model més eficaç, assolint un F1-score (macro) de 0.7062 en validació, la qual cosa es tradueix en un resultat competitiu en el test públic del 0.6666 i una bona detecció de pacients sans i malalts.**

***Index Terms*—Aprenentatge Automàtic, Classificació, Malaltia Hepàtica, Desequilibri de Classes, SMOTE, ExtraTreesClassifier, Enginyeria de Característiques.**

I. Introducció

Les malalties hepàtiques constitueixen un problema de salut pública creixent a nivell mundial, causant aproximadament 2 milions de morts anualment [2]. En aquest context, el desenvolupament de models d'aprenentatge automàtic per a la detecció primerenca esdevé una eina crucial per donar suport als professionals mèdics, ja que permeten identificar patrons complexos i no lineals en dades clíniques que poden passar desapercebuts en una anàlisi convencional.

Aquest informe detalla el procés seguit per resoldre un problema de classificació binària proposat en una competició de Kaggle InClass [4]. L'objectiu és predir si un subjecte pateix una malaltia hepàtica basant-se en un conjunt de 10 característiques clíniques i demogràfiques provinents del dataset *Indian Liver Patient Dataset (ILPD)* del repositori de la UCI [1].

El dataset d'entrenament original ha sigut preprocessat pel professorat de l'assignatura. S'han eliminat les mostres amb valors perduts i convertit les variables categòriques en dades numèriques. Han reassignat els valors de la variable objectiu binària de l'antic format a un nou esquema 0 i 1, on 0 indica pacient malalt i 1 subjecte sa. Amb aquest preprocessat, el conjunt de dades d'entrenament conté 463 observacions i el de test 116. Un dels principals reptes d'aquest problema és el desequilibri de classes, ja que el nombre de pacients malalts és significativament superior al de subjectes sans, un factor que ha guiat tot el procés de modelatge. Aquest treball presenta l'anàlisi de dades, l'enginyeria de característiques, el pipeline de preprocessament, l'avaluació de models i els resultats dels experiments realitzats per assolir el millor rendiment possible.

II. Anàlisi i Enginyeria de Característiques

L'anàlisi exploratòria de dades (EDA) va ser fonamental per entendre la naturalesa de les dades i definir les estratègies de preprocessament i modelatge.

A. Distribució de les Dades

Com es pot observar a la Fig. 1, el dataset presenta un clar desequilibri de classes. La classe majoritària (Malalt, 0) representa aproximadament el 71% de les mostres. Això justifica la necessitat de tècniques de rebalanceig com SMOTE i mètriques d'avaluació robustes com l'F1-score (macro).

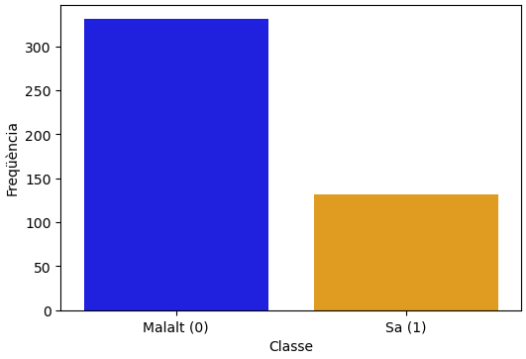

Fig. 1. Distribució de la variable objectiu (0: Malalt, 1: Sa).

L'anàlisi de les distribucions va revelar que diverses característiques bioquímiques com *TB, DB, Alkphos, Sgpt* i *Sgot* presenten un fort biaix cap a la dreta, indicant la presència de valors extrems (outliers). Com es mostra a la Fig. 2 amb l'exemple de la Bilirubina Total (TB), la majoria de valors es concentren a prop del zero, amb una cua llarga cap a la dreta que reflecteix els valors atípics, i amb un boxplot es

confirma la seva presència. Aquest patró va motivar l'ús de transformacions logarítmiques i escaladors robustos.

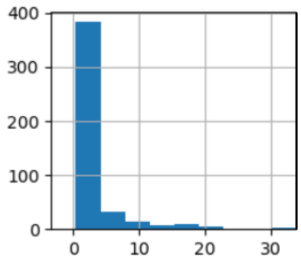

Fig. 2. Histograma de la Bilirubina Total (TB), exemple de distribució amb biaix a la dreta i outliers.

B. Enginyeria de Característiques

La matriu de correlació (vegeu Fig. A1 a l'Annex) va mostrar una forta multicol·linealitat entre variables com TB-DB (0.85), Sgpt-Sgot (0.91) i TP-ALB (0.80). Per abordar-ho i millorar el model, es van aplicar dues transformacions basades en pràctiques clíniques validades [5]:

- **Bilirubina Indirecta (IB)**: Es va calcular com $IB = TB - DB$, un paràmetre clínicament significatiu. Posteriorment, es va eliminar 'DB' per reduir la redundància.
- **Eliminació de Proteïnes Totals (TP)**: Es va eliminar 'TP' per la seva alta correlació amb 'ALB', conservant aquesta última per ser un indicador més específic de la funció hepàtica.

Aquestes decisions van millorar significativament el rendiment. Com es mostra a la Taula I, el F1-score en validació de tots els models va augmentar després d'aplicar aquests canvis. A més, en el test públic de Kaggle, el millor model va passar d'un F1 de 0.63414 a 0.66666.

Model	F1-Val (Abans)	F1-Val (Després)
ExtraTrees	0.6746	**0.7062**
Stacking	0.6743	0.6927
Voting	0.6741	0.6926
RandomForest	0.6683	0.6907
LogisticReg	0.6694	0.6628

TABLE I
COMPARATIVA DE F1-SCORE EN VALIDACIÓ ABANS I DESPRÉS DE L'ENGINYERIA DE CARACTERÍSTIQUES.

III. Mètodes de Classificació

El pipeline de modelatge es va dissenyar per abordar els reptes identificats a l'EDA, especialment els outliers i el desequilibri de classes.

A. Pipeline de Preprocessament i Gestió del Desequilibri

- **Transformació Logarítmica**: Es va aplicar la transformació `np.log1p` a les columnes amb biaix (TB, Alkphos, Sgpt, Sgot, IB) per normalitzar la seva distribució i reduir l'impacte dels valors extrems, utilitzant 'log1p' per la seva capacitat de gestionar valors nuls sense errors.
- **Escalat de Dades**: Es va utilitzar `RobustScaler`, que centra les dades utilitzant la mediana i les escala segons el rang interquartílic (IQR), fent-lo molt més robust als outliers.
- **SMOTE**: Per contrarestar el desequilibri, es va integrar aquesta tècnica [3] dins de cada pipeline. Consisteix a crear mostres sintètiques de la classe minoritària (Sa, 1) basant-se en els seus veïns més propers, i es va aplicar només a les dades d'entrenament de cada partició de la validació creuada per evitar *data leakage*.
- **Validació Creuada Estratificada**: Es va fer servir `StratifiedKFold` (amb 10 splits) en tots els processos d'optimització i avaluació.

B. Models de Classificació Avaluats

Es van avaluar diversos models per determinar el més adequat, buscant un equilibri entre simplicitat, interpretabilitat i potència predictiva:

- `LogisticRegression`: Com a model base lineal, simple i interpretable.
- `SVC`: Per la seva capacitat de trobar hiperplans de separació òptims.
- `KNeighborsClassifier`: Com a mètode no paramètric intuïtiu.
- `RandomForest` i `ExtraTrees`: Com a mètodes d'ensemble robustos, menys sensibles als outliers.
- `VotingClassifier` i `StackingClassifier`: Per combinar les fortaleses dels models individuals i millorar la generalització.

IV. Mètriques d'Avaluació

La selecció de les mètriques es va centrar en la seva capacitat per gestionar el desequilibri de classes inherent al problema.

- **F1-score (macro)**: És la mètrica oficial de la competició. La mitjana "macro" calcula l'F1-score per a cada classe de forma independent i després en fa la mitjana. Això otorga la mateixa importància a la classe minoritària i a la majoritària, proporcionant una visió equilibrada del rendiment.
- **AUC-ROC**: Avalua la capacitat del model per discriminar correctament entre classes, sent independent del llindar de decisió.
- **Precisió**: Indica la proporció de prediccions positives que són correctes. Per a la classe *Malalt*, una alta precisió minimitza els falsos positius.
- **Sensibilitat (Recall)**: Mesura la proporció de positius reals identificats. Un recall alt per a la classe *Malalt* és crucial per minimitzar els falsos negatius en aquest problema mèdic.

V. Experiments

El procés experimental es va centrar en l'optimització robusta de cada model i en la validació de les decisions preses.

A. Optimització d'Hiperparàmetres i Llindar de Decisió

Per a cada classificador, es va realitzar una cerca exhaustiva amb `GridSearchCV` (10-fold CV) optimitzant per l'F1_macro. Posteriorment, mitjançant `cross_val_predict`, es va buscar el llindar de decisió que maximitzava novament l'F1-score (macro), una tècnica crucial en problemes desequilibrats.

B. Anàlisi d'Experiments Descartats

Es van realitzar diversos experiments que, tot i ser teòricament prometedors, no van millorar els resultats i van ser descartats:

- **Ràtio SGOT/SGPT**: La creació d'aquest ràtio clínic [6] va empitjorar el rendiment (p. ex., F1 de validació del Stacking va baixar de 0.6927 a 0.6772).
- **StandardScaler**: L'ús de `StandardScaler` va oferir resultats consistentment inferiors en validació (p. ex., ExtraTrees va baixar de F1 0.7062 a 0.6964).
- **SMOTEENN**: Aquesta tècnica híbrida no va millorar els resultats (F1 de ExtraTrees va baixar a 0.6527).
- **class_weight**: Aquesta alternativa a SMOTE va resultar en un rendiment generalment inferior en validació i test.

VI. Resultats i Anàlisi

Els resultats obtinguts a la validació creuada, després de totes les optimitzacions, es presenten a la Taula II.

Model	F1 Macro	AUC (Sa)	Recall (Sa)	Recall (Malalt)
ExtraTrees	**0.7062**	0.7756	0.7121	0.7492
Stacking	0.6927	0.7689	0.5303	**0.8459**
Voting	0.6926	0.7769	0.6970	0.7372
RandomForest	0.6907	0.7599	0.5606	0.8218
LogisticReg	0.6628	0.7394	0.7727	0.6465
KNN	0.6562	0.7362	0.6667	0.6979
SVC	0.6474	0.7014	0.5833	0.7372

TABLE II
Resultats Finals dels Models en Validació Creuada.

Els models d'ensemble basats en arbres (`ExtraTrees`, `Stacking` i `Voting`) ofereixen el millor rendiment. Concretament, `ExtraTreesClassifier` aconsegueix l'F1-score més alt (0.7062) i el millor equilibri. La Fig. 3 mostra que les variables bioquímiques com les bilirubines, *Sgot* i *Alkphos* són les més influents. L'anàlisi dels coeficients de la Regressió Logística (Fig. A2 a l'Annex) ofereix una interpretació complementària.

És interessant observar el compromís (*trade-off*) de les sensibilitats de les dues classes. El `StackingClassifier`, tot i obtenir un excel·lent recall per a la classe malalta (0.8459), presenta un rendiment molt baix en el recall de la classe sana (0.5303). Aquest comportament implicaria un nombre significatiu de falsos positius. Per contra, la `LogisticRegression` mostra una tendència oposada.

L'`ExtraTreesClassifier` ofereix el millor equilibri entre el recall per a malalts (0.75) i per a sans (0.71), convertint-lo en la solució més robusta. La seva matriu de confusió (Fig. 4) visualitza aquest balanç.

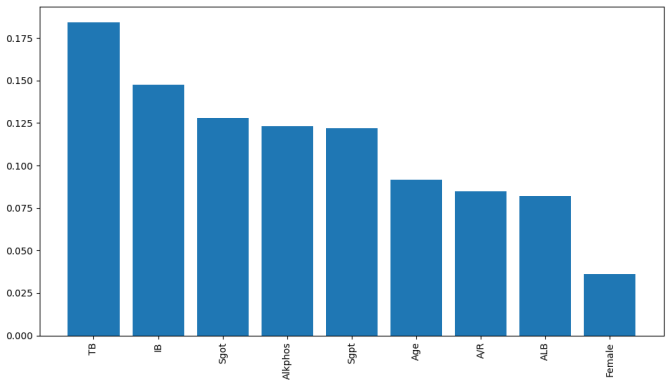

Fig. 3. Importància de les característiques segons l'ExtraTreesClassifier.

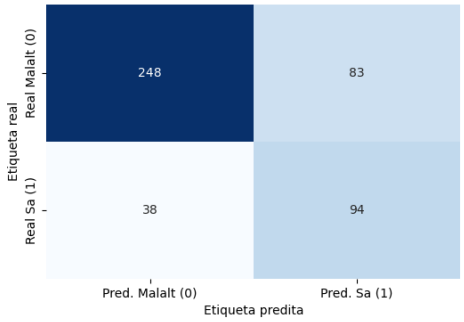

Fig. 4. Matriu de confusió del model ExtraTrees en validació creuada.

VII. Conclusions

En aquest treball s'ha treballat amb models de classificació per a la detecció de malalties hepàtiques. La clau del projecte ha sigut fer servir una bona metodologia, que ha abordat de manera efectiva el desequilibri de classes, la presència d'outliers i la multicol·linealitat.

L'estratègia de preprocessament, combinant enginyeria de característiques, transformació logarítmica i `RobustScaler`, juntament amb l'ús de SMOTE, va demostrar ser molt efectiva. A més, l'optimització del llindar de decisió va ser una tècnica útil per millorar la mètrica F1-score.

D'entre tots els models avaluats, l'`ExtraTreesClassifier` va destacar com la millor opció, obtenint un F1-score de 0.7062 en validació. Aquest model ofereix el millor equilibri per identificar correctament tant els pacients sans com els malalts, convertint-lo en una solució robusta i fiable. Com a prova final, cal veure el resultat en la part privada, que indicarà com de bé generalitza aquest model, i veurà si les mètriques de validació obtinguda són coherents sobre el test complet.

Agraïments

Volem agrair al professorat de l'assignatura la seva guia i el plantejament d'aquesta competició, que ens ha permès aplicar de forma pràctica els conceptes teòrics en un àmbit d'especial interès pels dos autors, la medicina.

Referències

[1] M. Lichman, *Indian Liver Patient Dataset (ILPD)*. Irvine, CA: University of California, School of Information and Computer Science, 2013. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29

[2] S. K. Asrani, H. Devarbhavi, J. Eaton, and P. S. Kamath, "Burden of liver diseases in the world," *Journal of Hepatology*, vol.70, no.1, pp.151–171, Jan.2019, doi:10.1016/j.jhep.2018.09.014. [Online]. Available: https://www.journal-of-hepatology.eu/article/S0168-8278(18)32388-2/abstract

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[4] J. Vidal, "Liver Patient Classification" Kaggle InClass, 2024. [Online]. Available: https://www.kaggle.com/t/1ccb38a0e7054e6dbe2c953c10cf0124

[5] A. Pan and S. Mukhopadhyay, "Liver Disease Detection: Evaluation of Machine Learning Algorithms Performances With Optimal Thresholds," *International Journal of Healthcare Information Systems and Informatics*, vol. 17, no. 2, 2022.

[6] M. Botros and K. A. Sikaris, "The De Ritis Ratio: The Test Of Time," *Clinical Biochemistry Reviews*, vol. 34, no. 3, pp. 117–130, Nov. 2013, PMID: 24353357; PMCID: PMC3866949

Appendix
Material Suplementari

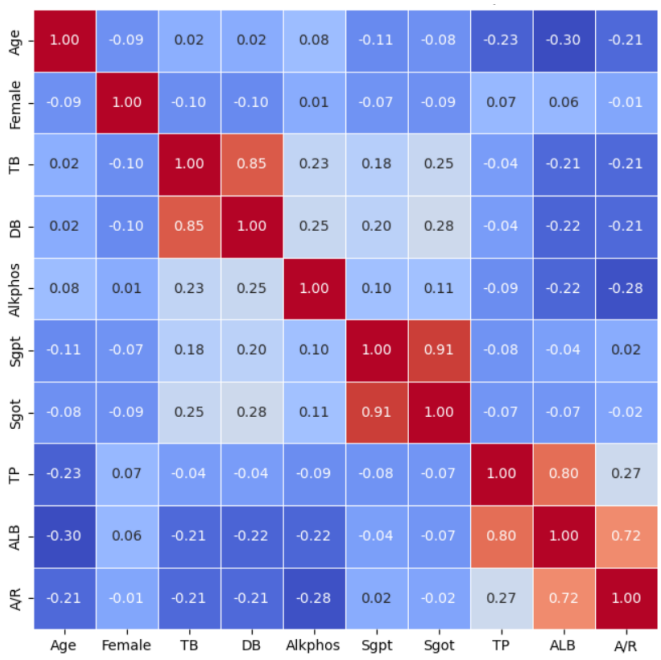

Fig. A1. Matriu de correlació de les característiques originals. S'observen correlacions fortes (properes a 1) entre TB-DB, Sgpt-Sgot i TP-ALB, justificant l'enginyeria de característiques.

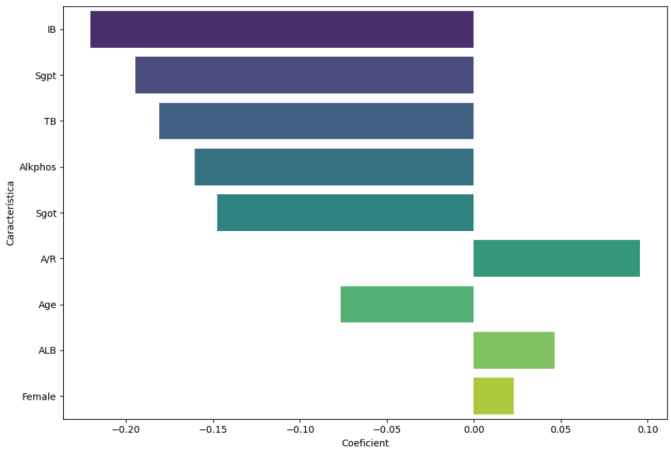

Fig. A2. Importància de les característiques segons els coeficients de la Regressió Logística. Els valors negatius s'associen amb la malaltia hepàtica, mentre que els positius amb la salut.