Classificació supervisada del patrimoni cultural del Garraf

Yulia Chernykh

28 de maig de 2023

Resum

Aquest estudi se centra en el desenvolupament d'un classificador supervisat de patrimoni cultural utilitzant tècniques de Machine Learning. L'objectiu és classificar els elements de patrimoni en categories comunes de l'àmbit, com ara patrimoni material, immaterial, moble, immoble i documental. S'han fet servir diferents models, en concret el model SVC, Regressió Logística multinomial, el de Random Forest i Gradient Boosting. Per últim, s'ha comparat el seu rendiment. El model Gradient Boosting ha mostrat el millor resultat, amb una precisió de 0.91, un recall de 0.97 i una mitjana del cross validation de 0.98.

1 Introducció

El patrimoni cultural d'una regió és un aspecte fonamental per a la seva identitat i desenvolupament. La classificació i comprensió dels diferents elements patrimonials pot proporcionar informació valuosa per a la presa de decisions relacionades amb la conservació, promoció i gestió adequada d'aquests recursos. En aquest context, l'aplicació de tècniques de Machine Learning pot ajudar a identificar patrons i tendències en conjunts de dades complexes, facilitant la tasca de classificació. En aquest estudi, s'ha abordat el problema de classificar el patrimoni cultural utilitzant tècniques de classificació supervisada multiclasse per tal d'agilitzar les tasques de catalogació dels elements.

2 State of the Art

Els mètodes més comuns usats en la classificació multiclasse inclouen el Support Vector Classifier (SVC), la Regressió Logística multinomial, el Decision Tree, el Random Forest, el Gradient Boosting Classifier i les xarxes neuronals (ANN). Aquests algoritmes han mostrat resultats prometedors en la classificació de conjunts de dades similars. La qualitat de les dades, l'equilibri de classes, la dimensionalitat, la quantitat de característiques rellevants o la presència de soroll són factors que condicionaran el rendiment dels models.

3 Metologia

3.1 Obtenció de les dades

Les dades es van extreure de la web de dades obertes de la Diputació de Barcelona mitjançant una crida a l'API per obtenir el dataset de Patrimoni del Garraf en format JSON. Es va convertir en una taula i es va reduir el conjunt de dades original de 37 variables a 9 de rellevants. Les variables seleccionades van ser 'Títol', 'Edat', 'Municipi', 'Protecció', 'Tipologia', 'Titularitat', 'Ús actual', 'Estat de conservació' i 'Àmbit', que és la variable objectiu. Moltes de les variables descartades no eren rellevants o vàlides per un model de Machine Learning, com ara: links, bibliografia, descripcions llargues, noms d'autors de fitxa, coordenades, etc.

3.2 Exploració

Es va realitzar una exploració inicial de les dades utilitzant diferents tècniques i gràfics de count plot, ja que la majoria de les variables eren categòriques. Es van identificar dades poc equilibrades amb valors dominants. Per exemple, es va observar que hi havia molts casos de patrimoni immoble del

segle XX, situats a Sitges, de protecció inexistent i propietat privada, amb ús residencial o sense ús, i en bon estat de conservació.

3.3 Neteja de dades i tractament de valors faltants

Es van identificar i tractar els duplicats i els valors nulls i NaN, la majoria dels quals eren espais en blanc. Per exemple, la columna 'Any' contenia molts valors faltants i es va eliminar. La columna 'Estil' també tenia molts valors faltants i altres valors que aglutinaven diversos estils, per la qual cosa també es va eliminar.

També es va modificar algunes variables agrupant categories similars, com és el cas de la columna 'Centúria'. Es van ajuntar els segles per edats com 'Prehistòria', 'Antiga', 'Mitjana, 'Moderna', 'Contemporània' i Sense data"(ja que hi havia elements patrimonials dels quals no se'n coneixia la data de creació).

Altres dades es van omplir utilitzant la categoria més comuna o investigant el tipus de patrimoni. Per exemple, es va realitzar una cerca per a les barraques d'Olivella i es va actualitzar la informació indicant que no tenen ús actual. Encara que no hi havia dates disponibles, es va investigar i es va situar aquest tipus de patrimoni entre els segles XVII i XX.

3.4 Preprocessament

Es va realitzar la codificació de les variables categòriques, convertint les variables ordinals com 'Estat' (bo, regular i dolent) i 'Èpoques' (de més antiga a més nova) en valors numèrics (1, 2, 3, etc). Les altres variables categòriques es van convertir en columnes binàries utilitzant la tècnica de 'dummification' o 'one-hot encoding', incrementant el nombre de columnes de 9 a 48.

3.5 Divisió de les dades

Les dades es van dividir en conjunts de dades d'entrenament i de prova. El conjunt de dades d'entrenament contenia les variables independents (X) excepte el target (Àmbit) i el títol de cada element.

3.6 Aplicació del model

Es van aplicar diversos models, com ara SVC (Support Vector Classification), Regressió Logística multinomial, Random Forest i Gradient Boosting. Per abordar el desequilibri de classes, es va realitzar un balancejament de les dades utilitzant les tècniques d'undersampling i oversampling amb l'algoritme "auto". Es va utilitzar un pipeline que integrava els models seleccionats juntament amb les tècniques de balancejament, ajust de paràmetres amb GridSearch i selecció dels millors models. Posteriorment, es va avaluar el rendiment dels models amb dades noves. La decisió d'equilibrar les dades va ser fruit de veure que sense fer-ho les mètriques d'avalució donaven molt bons resultats en tots els models (entre el 0.9 i l'1) però no feien bones prediccions de les categories minoritaries. Aplicant-ho va baixar una mica la puntuació però van millorar les prediccions de les categories desequilibrades.

3.7 Avaluació

Es va realitzar una avaluació dels models utilitzant diverses mètriques, com ara l'exactitud (accuracy), precisió (precision), exhaustivitat (recall), puntuació F1, matriu de confusió, puntuacions de validació creuada (cross-validation scores) i mitjana de l'exactitud de la validació creuada (mean CV accuracy). Es va crear una taula per comparar les etiquetes reals amb les predites i es va observar que les mètriques i els resultats de la taula mostraven un bon rendiment dels models.

4 Resultats

4.1 Model: Support Vector Classifier (SVC)

- Accuracy: 0.8556701030927835
- Precisió: 0.6839809523809524

- Recall: 0.7246394833236938
- F1 Score: 0.6950315172867543

El model SVC va obtenir una accuracy del 0.85, una precisió del 0.68, un recall del 0.72 i un F1 Score del 0.69. Tot i que aquest model té un rendiment acceptable, presenta una puntuació en les 4 mètriques inferior als altres models analitzats. Això indica que el model SVC té dificultats per classificar correctament les categories del patrimoni cultural.

- Puntuació mitjana de validació creuada: 0.8302441108915122

La puntuació mitjana de validació creuada per al model SVC és del 0.83. Aquesta puntuació indica el rendiment mitjà del model en diferents subdivisions del conjunt de dades utilitzades durant la validació creuada.

Pel que fa a la matriu de confussió:

$$\begin{bmatrix} 3 & 3 & 0 & 1 & 0 \\ 0 & 19 & 1 & 0 & 10 \\ 3 & 16 & 247 & 1 & 13 \\ 1 & 1 & 0 & 14 & 3 \\ 0 & 1 & 2 & 0 & 49 \end{bmatrix}$$

Mostra les dificultats de classificació del model en les diferents classes. S'observen confusions entre les classes 0 i 1 (immoble i immaterial), així com entre les classes 1 i 4(immaterial i documental). La classe 2 (patrimoni natual) presenta la majoria de les mostres classificades correctament, però també hi ha confusions amb les classes 0, 1, 3 i 4. La classe 3 (patrimoni moble) té algunes mostres confoses amb les classes 0, 1 i 4. La classe 4 (patrimoni documental) té una alta precisió, tot i que hi ha algunes confusions amb la classe 1 i 2. En resum, el model té dificultats específiques en la classificació de certes classes, mentre que altres tenen un millor rendiment.

4.2 Model: Logistic Regression

- Accuracy: 0.9484536082474226
- Precisió: 0.8645102389236623
- Recall: 0.8978484673221516
- F1 Score: 0.8779664668783267

El model de Regressió Logística ha obtingut un rendiment significativament millor, amb una accuracy del 0.94, una precisió del 0.86, un recall del 0.89 i un F1 Score del 0.87. Aquest model és capaç de distingir amb èxit les diferents categories del patrimoni cultural.

- Puntuació mitjana de validació creuada: 0.9504754019550237

El model de Regressió Logística ha obtingut una puntuació mitjana de validació creuada del 0.95. Això indica un rendiment generalment alt i consistent del model en diferents subdivisions del conjunt de dades.

$$\begin{bmatrix} 5 & 1 & 0 & 1 & 0 \\ 0 & 30 & 0 & 0 & 0 \\ 0 & 8 & 268 & 0 & 4 \\ 1 & 0 & 0 & 17 & 1 \\ 1 & 0 & 3 & 0 & 48 \end{bmatrix}$$

Quant a la matriu de confussió, mostra que les classes 0, 1, 3 i 4 (patrimoni immoble, patrimoni immaterial, patrimoni moble i patrimoni documental) han estat classificades de manera força precisa, amb molt pocs errors. No obstant això, hi ha algunes confusions a la classificació de la classe 2 (patrimoni natural).

4.3 Model: Random Forest

- Accuracy: 0.9329896907216495
- Precisió: 0.8098965848965849
- Recall: 0.8938741083477926
- F1 Score: 0.8431329637073255

El model Random Forest ha obtingut una accuracy del 0.93, una precisió del 0.80, un recall del 0.89 i un F1 Score del 0.84. Aquest model té un rendiment satisfactori, però la seva precisió és lleugerament inferior als altres models.

- Puntuació mitjana de validació creuada: 0.9530286843651515

El model Random Forest ha obtingut una puntuació mitjana de validació creuada del 0.95. Això indica un bon rendiment general del model i una certa consistència en diferents subdivisions del conjunt de dades.

$$\begin{bmatrix} 5 & 1 & 0 & 1 & 0 \\ 0 & 29 & 1 & 0 & 0 \\ 4 & 12 & 261 & 0 & 3 \\ 1 & 0 & 0 & 17 & 1 \\ 0 & 0 & 2 & 0 & 50 \end{bmatrix}$$

Pel que fa la matriu de confussió, la majoria de les classes són classificades correctament, encara que hi ha força confusió entre les classes 1 i 2 (patrimoni immaterial i patrimoni natural). La classe 4 (patrimoni documental) és la millor classificada, amb quasi totes les mostres correctament etiquetades.

4.4 Model: Gradient Boosting

- Accuracy: 0.9458762886597938
- Precisió: 0.9176010634138713
- Recall: 0.9790476190476192
- F1 Score: 0.9390720419289174

El model Gradient Boosting ha mostrat un rendiment excel·lent, amb una accuracy del 0.94, una precisió del 0.91, un recall del 0.97 i un F1 Score del 0.93. Aquest model és capaç de classificar amb gran precisió les categories del patrimoni cultural.

- Puntuació mitjana de validació creuada: 0.9860637786443032

El model Gradient Boosting ha obtingut una puntuació mitjana de validació creuada excepcionalment alta del 0.98. Aquesta puntuació indica un rendiment molt robust i consistent en diferents subdivisions del conjunt de dades.

$$\begin{bmatrix} 7 & 0 & 0 & 0 & 0 \\ 0 & 29 & 1 & 0 & 0 \\ 0 & 20 & 260 & 0 & 0 \\ 0 & 0 & 0 & 19 & 0 \\ 0 & 0 & 0 & 0 & 52 \end{bmatrix}$$

Per últim, aquesta matriu de confussió, mostra que les classes van ser classificades correctament sense gairebé errades, menys la classe 2 (patrimoni natural) que sí que es confon bastant amb la classe 1(patrimoni immaterial). En general indica una alta precisió i un bon rendiment del model. En general, els models de Regressió Logística, Random Forest i Gradient Boosting mostren un rendiment sòlid en comparació del model SVC. Aquests models aconsegueixen una alta precisió i un equilibri entre recall i F1 Score. El que millors puntuacions obté en tots els sentits és el Gradient Boosting.

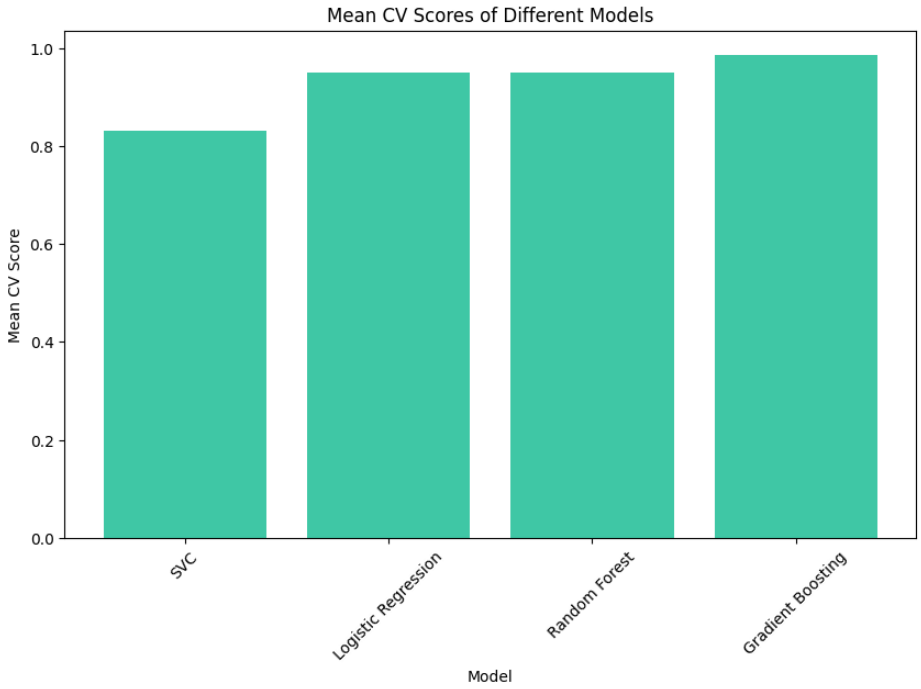

Figura 1: puntuacions de la mitjana de la cross validaton

5 Conclussions

El desenvolupament d'un classificador de patrimoni cultural utilitzant tècniques de Machine Learning ha demostrat ser prometedor. Els resultats obtinguts mostren que és possible classificar les diferents categories de patrimoni amb una precisió raonable.

La transformació de les variables categòriques a numèriques i el balancejament de les dades han estat factors clau per millorar el rendiment dels models. Tot i així, encara hi ha reptes a l'hora de classificar algunes categories minoritàries, com ara el patrimoni immaterial, documental i natural, que requereixen més investigació i millora en els models.

Els resultats podrien millorar amb més registres i variables, així com amb dades més precises. S'ha de tenir en compte que hi havia una gran categoria amb la descripció 'Sense data' a la columna 'Centúria' i també hi havia mancances de dates i estils a la columna 'Any' i 'Estil'. Malgrat això, les dades presentaven una correlació significativa, tal com es pot observar al mapa de calor del quadern adjunt. A partir del seu anàlisi, es poden extreure les següents observacions: la tipologia d'obra civil, com arquitectura i estructura, està estretament relacionada amb l'ús que se li dóna. Algunes d'aquestes construccions, tot i ser de titularitat privada, són accessibles al públic. És comú que els elements destinats a ús residencial tinguin protecció legal, cosa que és coherent ja que molts edificis són de titularitat privada i tenen un ús residencial. En molts casos, els elements patrimonials situats a Sitges gaudeixen de protecció legal. Les zones d'interès solen tenir un ús recreatiu actual i són de titularitat pública, mentre que els elements de titularitat pública acostumen a tenir un ús social. Les tradicions orals sovint tenen un ús simbòlic, mentre que els elements urbans com escultures i monuments sovint es destinen a fins ornamentals. L'antiguitat d'un element sol estar relacionada amb el seu estat de conservació, sent més probable que els elements més antics es trobin en un estat més deteriorat. També és possible que en algunes ocasions Olivella, Sant Pere de Ribes i Sitges mostrin un estat de conservació més precari.

En futurs treballs, es podria explorar l'ús de tècniques avançades de Machine Learning, com ara xarxes neuronals o algoritmes de Deep Learning, per millorar encara més la classificació del patrimoni cultural. També es podrien considerar altres variables o característiques rellevants per millorar la capacitat de discriminació dels models.

6 Referències

Patrimoni Cultural de la Diputació de Barcelona. (s.d.). Sitges. Recuperat el 12 de maig de 2023, de *https://patrimonicultural.diba.cat/municipi/sitges*

FreeCodeCamp. (s.d.). Limpieza de Datos en Pandas Explicado con Ejemplos. Recuperat el 14 de maig de 2023, de *https://www.freecodecamp.org/espanol/news/limpieza-de-datos-en-pandas-explicado-con-ejemplos/*

Medium. (2017, març 27). Aprendizaje Supervisado: Introducción a la Clasificación y Principales Algoritmos. Recuperat el 16 de maig de 2023, de

Analytics Vidhya. (2020, juliol 23). 10 Técnicas para Abordar el Desbalanceo de Clases en el Aprendizaje Automático. Recuperat el 18 de maig de 2023, de *https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/*

Analytics Vidhya. (2020, gener 29). Build Your First Machine Learning Pipeline Using Scikit-Learn. Recuperat el 18 de maig de 2023, de *https://www.analyticsvidhya.com/blog/2020/01/build-your-first-machine-learning-pipeline-using-scikit-learn/ https://medium.com/datos-y-ciencia/aprendizaje-supervisado-introducci*

Analytics Vidhya. (2020, juliol 23). 10 Técnicas para Abordar el Desbalanceo de Clases en el Aprendizaje Automático. Recuperat el 18 de maig de 2023, de https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/

Analytics Vidhya. (2020, gener 29). Build Your First Machine Learning Pipeline Using Scikit-Learn. Recuperat el 18 de maig de 2023, de https://www.analyticsvidhya.com/blog/2020/01/build-your-first-machine-learning-pipeline-using-scikit-learn/