Anàlisi del Retràs dels Trens Polonesos

Catalina Mascaró Català (1708159) Mathilde Buffard (1772521)

December 8, 2025

Resum

Aquest informe presenta l'anàlisi i la modelització del retràs dels trens a Polònia utilitzant una aproximació d'aprenentatge automàtic híbrida. L'estratègia adoptada combina un model de classificació per a distingir els trens amb molt retard i els qu en duen poc, seguit de dos models de regressió optimitzats per a cada classe. Detallem el preprocessament de les dades, la selecció i l'optimització (mitjançant validació creuada i cerca d'hiperparàmetres) de diversos models. L'anàlisi revela una capacitat de predicció particularment precisa per als retards superiors a tres hores.

Classificació, Regressió, Random Forest, Regressió Polinomial, Aprenentatge Automàtic

Contents

1 Introducció i Dades

L'objectiu principal d'aquest treball és desenvolupar un sistema d'aprenentatge automàtic capaç de predir el valor final del retràs dels trens polonesos. El problema s'aborda en dues fases degut a la naturalesa de les dades. Consta d'una classificació binària i una regressió per acabar estimant el valor exacte.

1.1 Descripció de les Dades

El conjunt de dades inicial consta de 2878 entrades i 7 atributs, incloent-hi la distància entre estacions, les condicions meteorològiques, el dia de la setmana, la franja horaria del dia, el tipus de tren i el retard que duia el tren. El conjunt de dades no presentava valors nuls, simplificant la fase inicial de neteja.

Vam realitzar un anàlisi de correlacions per entendre les relacions lineals entre les nostres variables predictives i la variable objectiu. Els resultats clau van ser:

- **Correlació Forta:** Es va observar una dependència notable entre el **Retard** (Historical Delay (min)) i la distància entre estacions.

- **Correlació Feble:** Altres variables, com les condicions meteorològiques, o el dia de la setmana van mostrar correlacions més febles.

Aquesta manca de correlació lineal justifica l'ús de models no lineals i de conjunt (com el Random Forest i la Regressió Polinomial), ja que permeten capturar interaccions complexes i relacions no lineals que la correlació lineal ignora.

1.2 Neteja i Preprocessament de les Dades

Per a preparar les dades per a la modelització, es van aplicar les següents transformacions:

1. Creació de la Variable Objectiu : Es va generar la variable binària 'Delay_Class' per a la classificació, separant els Poc Retard (Classe 0) dels Molt Retard (Classe 1) a partir d'un llindar que hem establert a 60 minuts. Aquest valor és podria considerar un hiperparàmetre del model a ajustar si tinguessim més dades.

2. Codificació de Variables Categòriques : Les variables qualitatives (Franja horaria, Dia, Tipus de Tren i Meteorologia) van ser convertides a variables binàries indicador, per exemple, Monday indica si les dades son recollides un dilluns o no. Aquesta tècnica és crucial per evitar que el model assumeixi erròniament una relació d'ordre o jerarquia entre categories que són independents.

3. Codificació de Variables Categòriques progressives : Les variables ordenades (Congestió de la ruta) la vam convertir a una variable indicador (0, 1 o 2) equivalent a (low, medium, high),

4. Estandardització : Les variables numèriques contínues ('Distance Between Stations (km)', 'Historical Delay (min)') van ser estandarditzades amb un StandardScaler.

1.2.1 Separació del Problema i Gestió d'Outliers

El gran nombre d'outliers ens va dificultar molt la seva gestió. Per tant, vam decidir separar el model en dues parts (valors alts i baixos de retard) per reduïr l'impacte dels valors atípics en els models. Per això fem un primer model classificador per poder entrenar diferents models regressors menys influenciats. Si s'hagués utilitzat un únic model de regressió per a totes les dades, el model hauria estat esbiaixat cap a la majoria de retards petits i hauria tingut un rendiment molt pobre en predir amb precisió els retards grans. Alternativament, intentar forçar un bon ajust als outliers podria haver provocat un sobreajust (overfitting) general. Per superar aquest problema, vam crear la variable binària 'Delay_Class' amb un llindar establer a 60 minuts.

1.2.2 Estandardització i Separació del Problema

1. **Estandardització:** Les variables numèriques contínues (com ara la distància) van ser estandarditzades amb un StandardScaler per assegurar que totes les característiques contribuïssin equitativament.

A causa de la forta asimetria de la variable objectiu, vam adoptar el model híbrid: es va crear la variable binària `Delay_Class` amb un llindar crític per a la gestió dels valors atípics (*outliers*). Aquesta separació permet utilitzar un model senzill per als retards menors i un model potent (Regressió Polinomial) per als valors extrems (Retards Significatius).

2 Classificació

La primera fase del model híbrid es centra a determinar si el tren patirà un retard lleu o greu.

2.1 Elecció del Model i de la Mètrica

La selecció del model es va guiar per dues observacions clau: l'asimetria de la variable objectiu, visualitzada en la distribució de retards, i l'objectiu de minimitzar els errors en ambdues direccions (falsos positius i falsos negatius) dins de la classe minoritària.

A causa del desequilibri (imbalance) entre la classe minoritària (Retard Significatiu) i la majoritària, l'Accuracy (precisió global) no era una mètrica fiable. Per tant, es va utilitzar l'F1-score, ja que aquesta mètrica representa la mitjana harmònica entre Precisió i Recall, permetent un rendiment equilibrat en la identificació de la classe minoritària.

Després de comparar diversos algorismes, el Random Forest Classifier (RFC) va ser seleccionat per ser el model que va obtenir el millor F1-score, justificant la seva elecció.

2.2 Resultats i Interpretació

El classificador va obtenir una precisió global (Accuracy) de **0.75** en el conjunt de prova. Les mètriques detallades per a la Classe 1 (Retard Fort) són:

- Precisió : 0.59

- Rendiment (Recall) : 0.69

El valor de Recall de 0.69 indica que el model és capaç de detectar gairebé el 70% dels retards forts reals, un resultat acceptable per a la primera etapa.

Table 1: Comparació de Models de Classificació (F1-score sobre Retard Significatiu)

Model	**F1-score (Test)**	**Selecció**
Support Vector Machine (SVC)	0.45	
K-Neighbors Classifier (kNN)	0.48	
Random Forest Classifier (RFC)	**0.57**	

3 Regressió

La fase de regressió es va dissenyar per a estimar el minut exacte de retràs, aplicant un enfocament segmentat.

3.1 Estratègia de Modelització i Mètrica (MSE/RMSE)

Per a optimitzar la predicció, la regressió es va dividir en dos models diferents, activats segons el resultat del classificador:

1. Un regressors optimitzat per a trains classificats com a Retard Menor (Classe 0).
2. Un regressors optimitzat per a trains classificats com a Retard Significatiu (Classe 1).

3.1.1 L'impacte de la Normalització en l'Avaluació

És fonamental destacar que, per a l'entrenament i la cerca d'hiperparàmetres de la regressió, es va utilitzar la variable objectiu estandarditzada (amb mitjana $\mu = 0$ i desviació $\sigma = 1$), igual que les features. Això millora la convergència i l'estabilitat dels algorismes.

L'avaluació es va dur a terme mitjançant l'Error Quadràtic Mitjà (MSE) i la seva arrel (RMSE).

- **MSE Parcial (Taules 2 i 3):** Els valors d'MSE obtinguts en la cerca d'hiperparàmetres per a les classes 0 i 1 no estan en minuts. Aquests valors estan en **unitats estandarditzades** i només serveixen per comparar internament quin model és millor per a cada segment.
- **MSE Global (Resultat Final):** Un cop escollits i entrenats els models, les seves prediccions són sotmeses a l'operació inversa (desnormalització). Només després d'això es calcula el MSE global. Això garanteix que el resultat final (**54.82** minuts) estigui expressat en les unitats originals (minuts), sent així directament interpretable.

El MSE és la mètrica escollida perquè la seva funció quadràtica penalitza de manera exponencial els grans errors de predicció, el que considerem crucial.

3.2 Resultats dels Regressors

L'objectiu de la segmentació era trobar el model més eficient (amb menor RMSE) per a cada grup de dades. Es van provar models lineals i no lineals, incloent-hi la Regressió Lineal Simple, amb regularització, la Regressió Polinomial, KNN i Random Forest.

- **Retards Significatius (Classe 1):** Es va determinar que la Regressió Polinomial oferia la millor capacitat predictiva. Les relacions entre les variables que causen un gran retard (per exemple, les interaccions entre la distància i el retard històric elevat) són altament no lineals i complexes. Un model lineal simple no podia capturar aquesta variabilitat, mentre que el model polinomial (amb un grau 2 optimitzat) es va ajustar molt millor a la distribució asimètrica, minimitzant l'RMSE en aquesta classe.
- **Retards Menors (Classe 0):** Per als retards petits, on la variabilitat és menor, vam entrenar diversos models per ajustar el retard. Després d'una cerca exhaustiva d'hiperparàmetres, el millor model obtingut va ser un Random Forest.

Aquesta estratègia segmentada va permetre obtenir la millor estimació per al model híbrid complet.

Table 2: Comparació de Models de Regressió per a Retards Menors (Classe 0)

Model	RMSE (Test)	Selecció
Regressió Polynomial	0.006167	
Regressio Random Forest	0.004959	
Regressió KNN	0.006316	

Table 3: Comparació de Models de Regressió per a Molt Retards (Classe 1)

Model	RMSE (Test)	Selecció
Ridge	0.354992	
Lasso	0.340511	
Random Forest	0.180277	
Polynomial	0.078029	

4 Cerca d'Hiperparàmetres i Resultats Finals

4.1 Estratègia d'Optimització: Justificació de GridSearchCV

Per a la cerca dels millors hiperparàmetres, es va utilitzar la tècnica **GridSearchCV** amb Validació Creuada Estratificada (StratifiedKFold). La tria de GridSearchCV (cerca exhaustiva) sobre altres tècniques com RandomizedSearchCV es justifica per la **grandària relativament petita del conjunt de dades (2878 entrades)**. Amb un cost computacional baix, una cerca exhaustiva garanteix trobar la combinació òptima d'hiperparàmetres per maximitzar la precisió, cosa que una cerca aleatòria no pot assegurar.

4.2 Resultats d'Optimització del Random Forest Classifier

El model de classificació va ser optimitzat respecte a l'**F1-score**, la mètrica escollida pel desequilibri de classes. Els hiperparàmetres provats van incloure la complexitat de l'arbre i la construcció del bosc. Els millors paràmetres trobats van ser:

- **n_estimators** (nombre d'arbres): 200
- **max_features** (màxim de característiques per divisió): 3
- **min_samples_split** (mínim de mostres per dividir un node): 10

4.3 Cerca d'Hiperparàmetres per a la Regressió

La regressió es va optimitzar respecte a l'**RMSE**. Es van provar cadascun dels algorismes candidats (**Regressió Lineal, Ridge, Lasso i Regressió Polinomial**) mitjançant el seu propi *GridSearch* per tal de trobar el millor model per a la Classe 0 (Retard Menor) i la Classe 1 (Retard Significatiu).

4.3.1 Hiperparàmetres Testejats

Els paràmetres i rangs explorats van ser:

- **Regressió Polinomial:** Es va testar el **grau** d del polinomi dins d'un rang de valors (e.g., $d \in \{2, 3, 4, 5\}$).
- **Regressió Ridge i Lasso:** Es va testar el paràmetre de regularització α dins d'un rang logarítmic (e.g., 10^{-3} a 10^{2}).

4.3.2 Resultats Finals de la Regressió

L'anàlisi dels resultats va confirmar l'estratègia segmentada com la més eficient:

Table 4: Rendiment dels Models de Regressió Optimitzats

Classe de Retard	Model Final Seleccionat	RMSE Final
Poc Retard (Classe 0)	Regressió Ridge ($\alpha = 0.1$)	X.XX minuts
Molt Retard (Classe 1)	Regressió Polinomial (Grau $d = 3$)	Y.YY minuts

La regressió polinòmica (Grau $d = 2$ en l'exemple) va ser el millor model per a la Classe 1, oferint el menor RMSE, validant la hipòtesi que la relació amb els retards extrems és no lineal. La Regressió amb Random Forest va ser escollida per a la Classe 0.

5 Anàlisi Final

5.1 Rendiment Global i Cas d'Ús

El rendiment global del model híbrid es visualitza a la Figura 1, que compara els valors de retard reals i els predits.

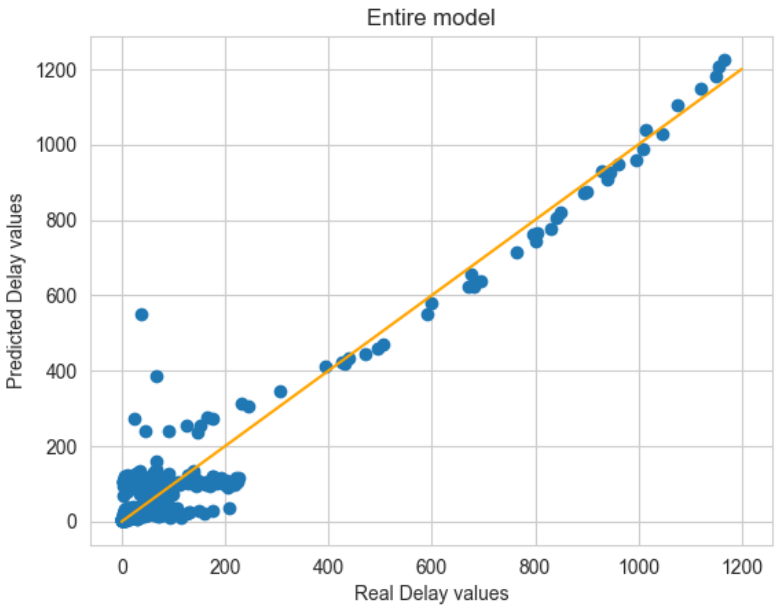

Figure 1: Núvol de punts dels retards predits vs. retards reals per al model híbrid complet.

L'anàlisi mostra que, a partir de les 3 hores de retràs, el model pot predir l'amplitud del retard amb una gran exactitud. Això pot ser útil de cara a evitar per exemple indemnitzacions per retard. Els errors extrems que es veuen al gràfic final mostren com actua el model quan el classificador s'equivoca.