
Proyecto Final Aprendizaje Automático

Lucía Herraiz Cano
Aprendizaje Automático
Universidad Pontificia Comillas
Abril 2025
202300465@alu.comillas.edu

Abstract

En este proyecto se desarrolla un análisis predictivo sobre el rendimiento académico de estudiantes de secundaria en dos institutos de Madrid durante el año 2005. A partir de un conjunto de datos, se construyen y comparan dos modelos para predecir la nota final del curso (T3). También se lleva a cabo un estudio de los datos y de las variables más influyentes en el desarrollo del estudiante. Este estudio busca no solo obtener predicciones precisas, sino también generar conocimiento accionable para mejorar el rendimiento estudiantil desde una perspectiva integral.

1 Exploratory data analysis

En esta sección se analiza la estructura y distribución del conjunto de datos, identificando patrones, valores atípicos y relaciones relevantes entre variables. También se detalla el proceso de limpieza y preparación necesario para el modelado posterior.

1.1 Limpieza de datos

El proceso de limpieza de datos viene recogido en la función `data_cleaning_pipeline` e incluye el manejo de outliers, la imputación de valores faltantes, la corrección de errores, la codificación de las variables categóricas y la estandarización de los datos.

Sólo 5 variables tenían valores faltantes, y con el objetivo de perder la menor cantidad de datos, todos los valores se han imputado siguiendo distintas estrategias. *AlcSem*, *Relfam* y *TiempoEstudio* han sido imputadas por la moda al ser variables categóricas, y tener menos de un 3% de valores faltantes. *Medu* y *Pedu*, al tener un porcentaje más relevante de valores faltantes, tener una correlación alta (0.653), y al ser variables con mucho peso, como se verá posteriormente, han sido imputadas con un regresor (IterativeImputer) que emplea el resto de datos para predecir sus valores de manera más robusta.

Respecto a los outliers, la variable más llamativa fue *faltas*. Se estableció un valor máximo de 150, correspondiente al número de días lectivos del calendario escolar estándar, considerando cualquier valor superior como erróneo. No se eliminaron registros, ya que *faltas* mostró ser una variable altamente relevante.

Se han corregido los datos de la columna *razon* al tener claves distintas para el mismo valor ("otras", "otros").

Finalmente, tras estudiar el balance de las clases, se ha visto que *EstPadres*, *EstSup* y *apoyo* están claramente desbalanceadas (80%-20%). Sin embargo, dado que ninguna es una variable muy relevante, y dado que los resultados reflejan bien el balance que se suele dar en la población real, se ha decidido mantener los datos y tener cuidado con los modelos que le den relevancia a estas variables.

1.2 Análisis no supervisado

Con el objetivo de comprender mejor la relevancia y relación de las variables, se han implementado técnicas de aprendizaje no supervisado, que se pueden consultar en los ficheros *Exploratory_Data_Analysis* y *Model_Testing*.

En primer lugar, se ha implementado **Principal Component Analysis (PCA)** con el objetivo de ver si la reducción de la dimensionalidad aporta beneficios y para ver los componentes principales. La varianza explicada con 2 y 3 componentes principales es menor al 30%, sin embargo, el estudio de sus loadings aporta información relevante. Para el 1º modelo, como cabía esperar, las variables más relevantes son *T1*, *T2* y *suspensos*, pero es más interesante estudiar el 2º modelo, puesto que al eliminar estas variables, las sustituyen *AlcSem*, *AlcFin*, *SalAm*, *TiempoLib* y otras, que indican que después de las **notas**, la **vida social del estudiante** es uno de los factores más relevantes para su desempeño. Además de estos, un tercer grupo de variables de peso que aparecen en los PC son *Medu*, *Pedu*, *Relfam* y empleando **Recursive Feature Elimination (RFE)** se obtiene también *Mtrab_docencia* y *Ptrab_docencia*, lo que indica que el **entorno familiar** es otro de los factores más influyentes en las notas.

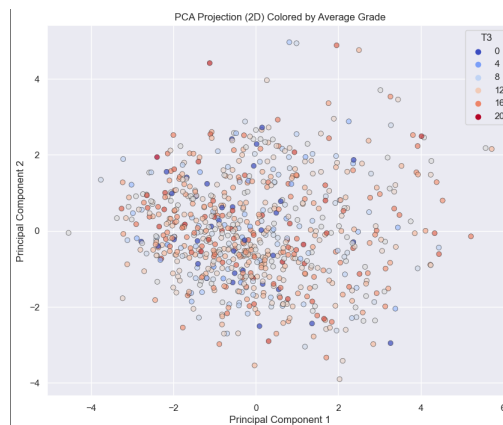


Figure 1: *

(a) PCA con 2 componentes principales

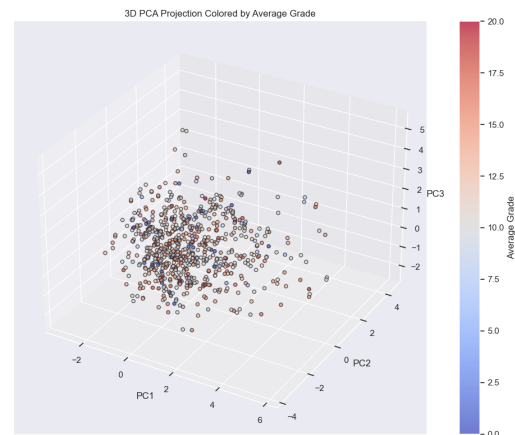


Figure 2: *

(b) PCA con 3 componentes principales

Figure 3: Visualización de los datos proyectados sobre 2 y 3 componentes principales mediante PCA.

Sin embargo, PCA requiere 16 componentes para explicar el 80% de la varianza. Al combinarlo con modelos como SVR o regresión lineal, su rendimiento disminuyó, por lo que se descartó su uso más allá de la exploración inicial. Este comportamiento podría deberse a la naturaleza lineal de PCA, incapaz de capturar relaciones no lineales en los datos. Por ello, se probaron técnicas no lineales como **ISOMAP** y **Kernel-PCA**, que operan en espacios transformados mediante distancias geodésicas o kernels y que pueden capturar relaciones más complejas entre los datos. No obstante, su combinación con modelos predictivos redujo el rendimiento entre un 20-30%. Podemos concluir que la reducción de dimensionalidad implica pérdida de información relevante para la predicción. Las variables no son fácilmente separables.

Asimismo, se aplicaron técnicas de **Clustering** a los datos (combinadas con PCA). Utilizando la métrica del *Elbow Method* y de la *Silueta*, el número óptimo de clusters se situó entre 2 y 3, en línea con los grupos de variables relevantes identificados anteriormente. No obstante, los valores obtenidos tras aplicar K-Means (*Silueta*: 0.233; Índice de Davies-Bouldin: 1.341) indican una calidad media en la segmentación, reafirmando la dificultad al separar los datos. Experimentalmente, se probó a añadir una nueva variable indicando la pertenencia a los clusters de los datos para entrenar a un regresor lineal, pero esto no modificó su *performance*, por lo que finalmente, usando la filosofía de la *Navaja de Occam*, se descartó el uso activo del clustering en los modelos.

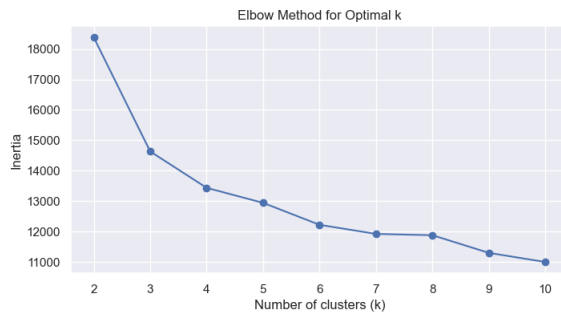


Figure 4: *
(a) Elbow Method

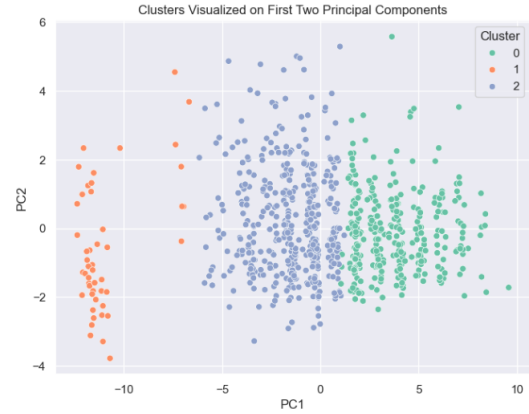


Figure 5: *
(b) Clusters

Figure 6: Análisis por clustering

1.3 Análisis adicional

A modo de comprobación del análisis anterior, y con el objetivo de descubrir nuevos patrones, se han empleado las características de varios modelos para ver la relevancia de las variables.

Se ha estudiado la representación conjunta de cada par de variables, empleando como código de color las distintas clases para identificar posibles relaciones¹, pero sólo se ha encontrado un patrón claro con $T1$ y $T2$.

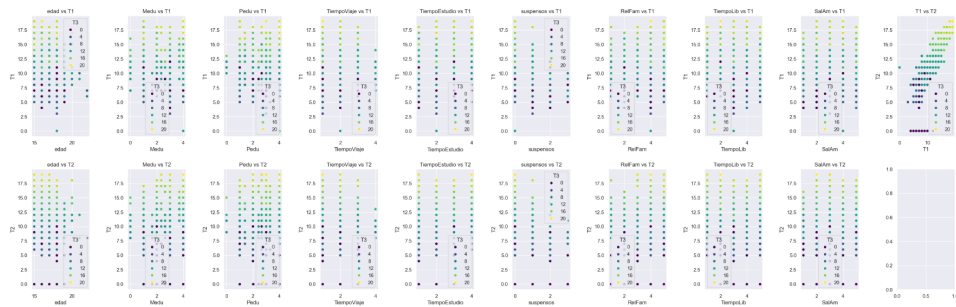


Figure 7: Scatter Plot con T1 y T2

Uno de los algoritmos con mejor desempeño en ambos modelos, como se verá más adelante, es **Random Forest**. Vemos que las variables que considera más importantes coinciden con las obtenidas con PCA en ambos casos.

A pesar de que el modelo de **regresión** logística no tiene el mejor rendimiento, al ser un clasificador, es interesante emplearlo, junto con regularización Lasso, para ver la evolución de los pesos de las variables para cada una de las clases, lo que aporta una mayor profundidad en el análisis. El estudio se ha realizado sin tener en cuente $T1$ y $T2$ y podemos ver que para las notas más bajas las variables relacionadas con la vida social y la familia son las más importantes, mientras que para las notas más altas, influyen variables más generales.

2 Implementación y comparación de los dos modelos predictivos

En esta sección se presentan los algoritmos con mayor rendimiento para ambos enfoques, junto con los criterios que justifican la elección de los modelos finales.

¹Ver la matriz completa en el fichero *Exploratory_Data_Analysis*

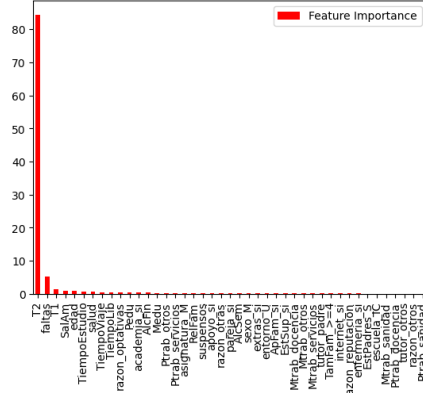


Figure 8: Feature importance en Random Forest

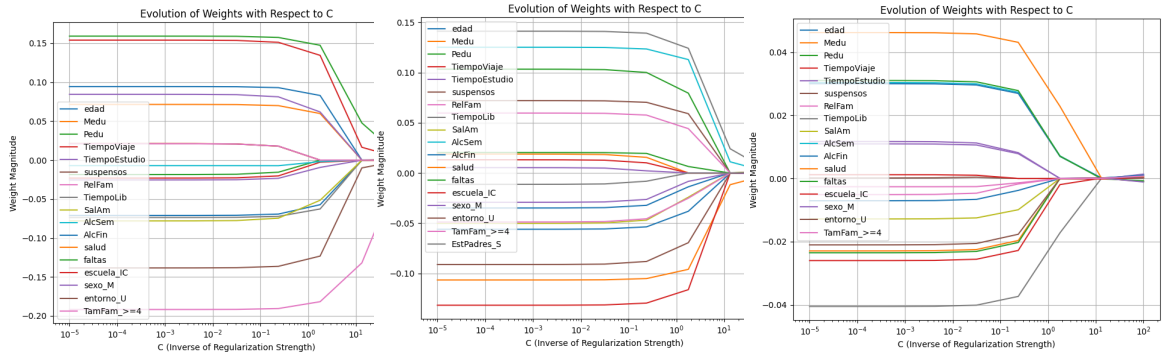


Figure 9: *
(a) Clase 1

Figure 10: *
(b) Clase 10

Figure 11: *
(c) Clase 20

Figure 12: Feature Importance para tres clases

2.1 Modelo 1: Enfoque Predictivo con la Información Completa del Dataset

Tras analizar múltiples modelos, se concluyó que los regresores superan consistentemente a los clasificadores, con una mejora media del 20% en las métricas. Por ello, se descartó en una fase inicial continuar con clasificadores y se priorizó profundizar en modelos de regresión.

Los modelos de clasificación probados junto con su *Accuracy* media fueron **Logistic Regression** (0.28), **Logistic Regression con Kernel-PCA** (0.57), **Support Vector Classifier (SVC)** (0.31), **Random Forest Classifier** (0.46) y **Bagging Classifier** (0.45).²

La Tabla 1 muestra los mejores modelos de regresión. Para garantizar la consistencia de las métricas, los valores presentados corresponden a los promedios obtenidos tras entrenar 300 modelos por cada técnica, utilizando datos mezclados en cada iteración.

Otros modelos de regresión probados, junto con sus valores R^2 medios fueron **Bagging Regressor** (0.80), **Support Vector Regressor (SVR) con Kernel-PCA** (0.80), **Linear Regression con ISOMAP** (0.61) y **Linear Regression con Kernel-PCA** (0.80).

2.2 Modelo 2: Enfoque Predictivo sin las variables T1 y T2

Al eliminar las variables de más peso, los modelos reducen a la mitad su poder predictivo.

²La implementación de estos modelos y otras métricas se pueden consultar en *Model1_Testing*

Table 1: Modelos de regresión

Comparativa entre modelos				
Nombre	R ² Score Test	MAE	MSE	R ² Score Train
Linear Regression	0.8301 \pm 0.033	0.2513	0.1693	0.853
LR con relaciones	0.8338 \pm 0.0319	0.2496	0.1682	0.855
Ensemble de LR	0.8332 \pm 0.0339	0.2491	0.1669	0.852
Stack	0.8332 \pm 0.0339	0.2513	0.1693	0.853
Bagging	0.830 \pm 0.033	0.2513	0.1693	0.853
Random Forest	0.830 \pm 0.033	0.2513	0.1693	0.853
SVR	0.830 \pm 0.033	0.2513	0.1693	0.853

3 Conclusiones accionables

El análisis de los datos nos indica que influyen tres grupos principales de variables a la hora de predecir el desempeño de los estudiantes, las **notas previas**, la **vida social** del estudiante, y su **entorno familiar**. Con esta información, los centros educativos pueden enfocar sus medidas en estas áreas. Sensibilizar al alumnado sobre la importancia de equilibrar vida social y estudio puede favorecer su rendimiento. Asimismo, promover políticas de conciliación familiar podría tener un impacto positivo en sus resultados académicos.

The only supported style file for NeurIPS 2023 is `neurips_2023.sty`, rewritten for L^AT_EX 2_ε. **Previous style files for L^AT_EX 2.09, Microsoft Word, and RTF are no longer supported!**

The L^AT_EX style file contains three optional arguments: `final`, which creates a camera-ready copy, `preprint`, which creates a preprint for submission to, e.g., arXiv, and `nonatbib`, which will not load the `natbib` package for you in case of package clash.

Preprint option If you wish to post a preprint of your work online, e.g., on arXiv, using the NeurIPS style, please use the `preprint` option. This will create a nonanonymized version of your work with the text “Preprint. Work in progress.” in the footer. This version may be distributed as you see fit, as long as you do not say which conference it was submitted to. Please **do not** use the `final` option, which should **only** be used for papers accepted to NeurIPS.

At submission time, please omit the `final` and `preprint` options. This will anonymize your submission and add line numbers to aid review. Please *do not* refer to these line numbers in your paper as they will be removed during generation of camera-ready copies.

The file `neurips_2023.tex` may be used as a “shell” for writing your paper. All you have to do is replace the author, title, abstract, and text of the paper with your own.

4 General formatting instructions

The text must be confined within a rectangle 5.5 inches (33 picas) wide and 9 inches (54 picas) long. The left margin is 1.5 inch (9 picas). Use 10 point type with a vertical spacing (leading) of 11 points. Times New Roman is the preferred typeface throughout, and will be selected for you by default. Paragraphs are separated by 1/2 line space (5.5 points), with no indentation.

The paper title should be 17 point, initial caps/lower case, bold, centered between two horizontal rules. The top rule should be 4 points thick and the bottom rule should be 1 point thick. Allow 1/4 inch space above and below the title to rules. All pages should start at 1 inch (6 picas) from the top of the page.

For the final version, authors’ names are set in boldface, and each name is centered above the corresponding address. The lead author’s name is to be listed first (left-most), and the co-authors’ names (if different address) are set to follow. If there is only one co-author, list both author and co-author side by side.

Please pay special attention to the instructions in Section regarding figures, tables, acknowledgments, and references.

5 Headings: first level

All headings should be lower case (except for first word and proper nouns), flush left, and bold.

First-level headings should be in 12-point type.

5.1 Headings: second level

Second-level headings should be in 10-point type.

5.1.1 Headings: third level

Third-level headings should be in 10-point type.

Paragraphs There is also a `\paragraph` command available, which sets the heading in bold, flush left, and inline with the text, with the heading followed by 1 em of space.

6 Citations, figures, tables, references

These instructions apply to everyone.

6.1 Citations within the text

The `natbib` package will be loaded for you by default. Citations may be author/year or numeric, as long as you maintain internal consistency. As to the format of the references themselves, any style is acceptable as long as it is used consistently.

The documentation for `natbib` may be found at

<http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf>

Of note is the command `\citet`, which produces citations appropriate for use in inline text. For example,

```
\citet{hasselmo} investigated\dots
```

produces

Hasselmo, et al. (1995) investigated...

If you wish to load the `natbib` package with options, you may add the following before loading the `neurips_2023` package:

```
\PassOptionsToPackage{options}{natbib}
```

If `natbib` clashes with another package you load, you can add the optional argument `nonatbib` when loading the style file:

```
\usepackage[nonatbib]{neurips_2023}
```

As submission is double blind, refer to your own published work in the third person. That is, use “In the previous work of Jones et al. [4],” not “In our previous work [4].” If you cite your other papers that are not widely available (e.g., a journal paper under review), use anonymous author names in the citation, e.g., an author of the form “A. Anonymous” and include a copy of the anonymized paper in the supplementary material.

6.2 Footnotes

Footnotes should be used sparingly. If you do require a footnote, indicate footnotes with a number³ in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).

³Sample of the first footnote.

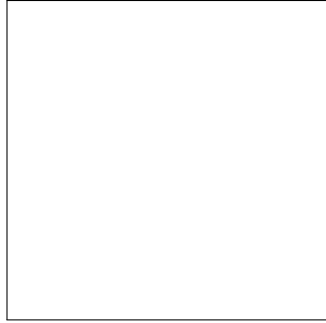


Figure 13: Sample figure caption.

Note that footnotes are properly typeset *after* punctuation marks.⁴

6.3 Figures

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction. The figure number and caption always appear after the figure. Place one line space before the figure caption and one line space after the figure. The figure caption should be lower case (except for first word and proper nouns); figures are numbered consecutively.

You may use color figures. However, it is best for the figure captions and the paper body to be legible if the paper is printed in either black/white or in color.

6.4 Tables

All tables must be centered, neat, clean and legible. The table number and

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

Note that publication-quality tables *do not contain vertical rules*. We strongly suggest the use of the `booktabs` package, which allows for typesetting high-quality, professional tables:

<https://www.ctan.org/pkg/booktabs>

6.5 Math

6.6 Final instructions

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle the text should fit into, and do not change font sizes (except perhaps in the **References** section; see below). Please note that pages should be numbered.

7 Preparing PDF files

Please prepare submission files with paper size “US Letter,” and not, for example, “A4.”

Fonts were the main cause of problems in the past years. Your PDF file must only contain Type 1 or Embedded TrueType fonts. Here are a few instructions to achieve this.

- You should directly generate PDF files using `pdflatex`.
- You can check which fonts a PDF file uses. In Acrobat Reader, select the menu Files>Document Properties>Fonts and select Show All Fonts. You can also use the program `pdf fonts` which comes with `xpdf` and is available out-of-the-box on most Linux machines.

⁴As in this example.

- xfig "patterned" shapes are implemented with bitmap fonts. Use "solid" shapes instead.
- The `\bbold` package almost always uses bitmap fonts. You should use the equivalent AMS Fonts:

```
\usepackage{amsfonts}
```

followed by, e.g., `\mathbb{R}`, `\mathbb{N}`, or `\mathbb{C}` for \mathbb{R} , \mathbb{N} or \mathbb{C} . You can also use the following workaround for reals, natural and complex:

```
\newcommand{\RR}{I\!\!R} %real numbers
\newcommand{\Nat}{I\!\!N} %natural numbers
\newcommand{\CC}{I\!\!C} %complex numbers
```

Note that `amsfonts` is automatically loaded by the `amssymb` package.

If your file contains type 3 fonts or non embedded TrueType fonts, we will ask you to fix it.

7.1 Margins in L^AT_EX

Most of the margin problems come from figures positioned by hand using `\special` or other commands. We suggest using the command `\includegraphics` from the `graphicx` package. Always specify the figure width as a multiple of the line width as in the example below:

```
\usepackage[pdftex]{graphicx} ...
\includegraphics[width=0.8\linewidth]{myfile.pdf}
```

See Section 4.4 in the `graphics` bundle documentation (<http://mirrors.ctan.org/macros/latex/required/graphics/grfguide.pdf>)

A number of width problems arise when L^AT_EX cannot properly hyphenate a line. Please give LaTeX hyphenation hints using the `\-` command when necessary.

Acknowledgments and Disclosure of Funding

Use unnumbered first level headings for the acknowledgments. All acknowledgments go at the end of the paper before the list of references. Moreover, you are required to declare funding (financial activities supporting the submitted work) and competing interests (related financial activities outside the submitted work). More information about this disclosure can be found at: <https://neurips.cc/Conferences/2023/PaperInformation/FundingDisclosure>.

Do **not** include this section in the anonymized submission, only in the final paper. You can use the `ack` environment provided in the style file to automatically hide this section in the anonymized submission.

8 Supplementary Material

Authors may wish to optionally include extra information (complete proofs, additional experiments and plots) in the appendix. All such materials should be part of the supplemental material (submitted separately) and should NOT be included in the main submission.

References

References follow the acknowledgments in the camera-ready paper. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to `small` (9 point) when listing the references. Note that the Reference section does not count towards the page limit.

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

- [2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural Simulation System*. New York: TELOS/Springer-Verlag.
- [3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.