Práctica Final:

Influencia de Diversos Factores Sociales y Escolares en la Calificación de Matemáticas de un Estudiante

Alejandro Cárdenas Barranco

Álvaro Rodríguez Gallardo

Juan Manuel Rodríguez Gómez

Estadística Multivariante (Grupo B)

Curso 2023 – 2024

UNIVERSIDAD
DE GRANADA

Índice

1. Introducción

El **objetivo principal** de este estudio es utilizar técnicas de **aprendizaje automático** para desarrollar un modelo predictivo **preciso**. Este modelo buscará establecer una relación entre las diferentes variables sociodemográficas y escolares proporcionadas en el dataset (como sexo, edad, tiempo de estudio, apoyo educativo recibido, entre otros) y la nota final obtenida por los estudiantes en Matemáticas.

El uso de técnicas de **aprendizaje automático**, como modelos de clasificación, permitirá construir un modelo robusto que pueda predecir con **precisión** las calificaciones de los estudiantes en función de las variables proporcionadas en el dataset. Esto puede ser de gran utilidad para profesores, administradores escolares y responsables de políticas educativas para mejorar el rendimiento estudiantil y brindar un apoyo más personalizado a los alumnos en sus estudios de Matemáticas.

El propósito de este análisis es **identificar las variables predictoras más relevantes** y determinar cómo influyen en el rendimiento académico en Matemáticas, así como hacer una **reducción de la dimensión**. Al comprender mejor estas relaciones, se busca no solo predecir las calificaciones futuras de los estudiantes, sino también obtener información útil para tomar decisiones educativas, identificar áreas de mejora y diseñar estrategias de apoyo más efectivas para los estudiantes.

2. Materiales y métodos

2.1. Materiales

Para realizar este estudio se ha tomado una base de datos. Esta contiene información sobre 31 indicadores medidos en 395 estudiantes. Estos indicadores son los siguientes:

- **school**: Centro escolar del estudiante
- **sex**: Sexo del estudiante
- **age**: Edad del estudiante
- **address**: Tipo de dirección de vivienda del estudiante
- **famsize**: Número de miembros de la familia del estudiante
- **Pstatus**: Estado de convivencia de los padres del estudiante

- **Medu**: Educación de la madre del estudiante
- **Fedu**: Educación del padre del estudiante
- **Mjob**: Trabajo de la madre del estudiante
- **Fjob**: Trabajo del padre del estudiante
- **reason**: Motivo del estudiante por el cual eligió su centro escolar
- **guardian**: Tutor del estudiante
- **traveltime**: Tiempo de viaje de casa del estudiante al centro escolar
- **studytime**: Tiempo de estudio semanal del estudiante
- **failures**: Número de materias suspendidas en el pasado
- **schoolsup**: El estudiante recibe apoyo educativo adicional
- **famsup**: El estudiante recibe apoyo educativo familiar
- **paid**: El estudiante recibe clases extras pagadas de la asignatura Matemáticas
- **activities**: El estudiante está apuntado en actividades extracurriculares
- **nursery**: El estudiante asistió de pequeño a la guardería
- **higher**: El estudiante quiere cursar estudios superiores
- **internet**: El estudiante tiene acceso a Internet en casa
- **romantic**: El estudiante se encuentra en una relación de pareja
- **famrel**: Calidad de las relaciones familiares del estudiante
- **freetime**: Cantidad de tiempo libre que posee el estudiante después del colegio
- **goout**: Frecuencia con la que el estudiante sale a la calle con sus amigos
- **Dalc**: Cantidad de consumo de alcohol del estudiante en jornada laboral
- **Walc**: Cantidad de consumo de alcohol del estudiante en el fin de semana
- **health**: Estado de salud actual del estudiante
- **absences**: Número de veces que el estudiante ha faltado al colegio
- **G3**: Nota final del estudiante en la asignatura Matemáticas

Se ha codificado el dataset, de forma que **todas las variables sean de tipo numérico** para poder trabajar con ellas más fácilmente.

Además, se muestran tablas donde aparece una visión general de los estadísticos de cada variable:

	Mínimo	Primer Cuartil	Mediana	Tercer Cuartil	Máximo	Rango
school	0.0000	0.0000	0.0000	0.0000	1.0000	1.0000
sex	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000
age	15.0	16.0	17.0	18.0	22.0	7.0000
address	0.0000	0.0000	0.0000	0.0000	1.0000	1.0000
famsize	0.0000	0.0000	1.0000	1.0000	1.0000	1.0000
Pstatus	0.0000	0.0000	0.0000	0.0000	1.0000	1.0000
Mjob	0.0000	2.0000	2.499	4.0000	4.0000	4.0000
reason	0.0000	0.0000	1.0000	2.0000	3.0000	3.0000
traveltime	1.0000	1.0000	1.0000	2.0000	4.0000	3.0000
studytime	1.0000	2.0000	2.0000	2.0230	4.0000	3.0000
failures	0.0000	0.0000	0.0000	0.0000	3.0000	3.0000
schoolsup	0.0000	1.0000	1.0000	1.0000	1.0000	1.0000
famsup	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000
paid	0.0000	0.0000	1.0000	1.0000	1.0000	1.0000
activities	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000
nursery	0.0000	0.0000	0.0000	0.0000	1.0000	1.0000
higher	0.0000	0.0000	0.0000	0.0000	1.0000	1.0000
internet	0.0000	0.0000	0.0000	0.0000	1.0000	1.0000
romantic	0.0000	0.0000	1.0000	1.0000	1.0000	1.0000
famrel	1.0000	4.0000	4.0000	5.0000	5.0000	4.0000
freetime	1.0000	3.0000	3.0000	4.0000	5.0000	4.0000
goout	1.0000	2.0000	3.0000	4.0000	5.0000	4.0000
Dalc	1.0000	1.0000	1.0000	2.0000	5.0000	4.0000
Walc	1.0000	1.0000	2.0000	3.0000	5.0000	4.0000
health	1.0000	3.0000	4.0000	5.0000	5.0000	4.0000
absences	0.0000	0.0000	4.0000	8.0000	75.000	75.0000
G3	0.0000	8.0000	11.0	14.0	20.0	20.0000

Tabla 1 Estadísticas de posición de las variables

	Media	Varianza	Desv. típica	Co. variación	Co. asimetría	Co. curtosis
school	0.1165	0.1031	0.3212	2.7579	2.3823	3.6848
sex	0.4734	0.2499	0.4999	1.0559	0.1061	-1.9938
age	16.7	1.6283	1.2760	0.0764	0.4627	-0.0314
address	0.2228	0.1736	0.4166	1.8702	1.3273	-0.2388
famsize	0.7114	0.2058	0.4537	0.6377	-0.9295	-1.1389
Pstatus	0.1038	0.0933	0.3054	2.9421	2.5882	4.7108
Mjob	2.499	1.9499	1.3964	0.5588	-0.4715	-0.9881
reason	1.273	0.9352	0.9671	0.7594	0.0383	-1.1256
traveltime	1.448	0.4865	0.6975	0.4817	1.5948	2.2727
studytime	2.023	0.6086	0.7802	0.3856	0.6884	0.4379
failures	0.3342	0.553	0.7437	2.2253	2.3689	4.8864
schoolsup	0.8709	0.1127	0.3358	0.3855	-2.2037	2.8636
famsup	0.3873	0.2379	0.4878	1.2592	0.4608	-1.7922
paid	0.5418	0.2489	0.4989	0.9208	-0.1670	-1.9771
activities	0.4911	0.2506	0.5006	1.0192	0.0353	-2.0038
nursery	0.2051	0.1634	0.4043	1.9714	1.4555	0.1187
higher	0.05063	0.0482	0.2195	4.3356	4.0836	14.7133
internet	0.1671	0.1395	0.3735	2.2355	1.7780	1.1643
romantic	0.6658	0.2231	0.4723	0.7093	-0.7004	-1.5132
famrel	3.944	0.8034	0.8967	0.2273	-0.9447	1.0895
freetime	3.235	0.9977	0.9999	0.3087	-0.1621	-0.3267
goout	3.109	1.2393	1.1133	0.3581	0.1156	-0.7869
Dalc	1.481	0.7934	0.8907	0.6014	2.1742	4.6454
Walc	2.291	1.6587	1.2879	0.5621	0.6073	-0.8072
health	3.554	1.9329	1.3903	0.3911	-0.4909	-1.0265
absences	5.709	64.0495	8.0031	1.4019	3.6437	21.3065
G3	10.42	20.9896	4.5814	0.4399	-0.7271	0.3661

Tabla 2 Estadísticas de dispersión de las variables

2.2. Métodos Estadísticos

Primero, se ha realizado un **análisis exploratorio** previo de los datos para detectar **valores pérdidos** y **valores atípicos (outliers)**. Se han tomado distintas decisiones para abordar estos valores anómalos.

- Para variables con **más de un 5 %** de valores pérdidos se ha analizado el patrón aleatorio de los mismos estudiando la **homogeneidad** según grupos con otras variables sin datos perdidos. Como las variables eran cuantitativas, se ha utilizado un **test de Student**.
- En el caso de las variables con **outliers**, se han detectado utilizando gráficos *boxplot* y se ha decidido sustituirlos por la **media** (al ser variables cuantitativas).

Después, se ha realizado un análisis **descriptivo numérico clásico** presentando medidas estadísticas de posición, dispersión y forma. Estas medidas han proporcionado una visión más detallada de la distribución de los datos, permitiendo una **comprensión** más profunda de su estructura y comportamiento.

En tercer lugar, se han aplicado diversas técnicas de análisis multivariante comprobando los supuestos para cada una de ellas:

- **Comprobación de la correlación entre datos**: Se ha realizado una **evaluación exhaustiva** de la correlación entre los datos, tanto a nivel poblacional como a nivel muestral. Para confirmar si las correlaciones eran significativamente distintas de cero a nivel poblacional se ha aplicado el **contraste de esfericidad de Bartlett**. A nivel muestral, se ha examinado la estructutra de las correlaciones utilizando la **matriz de correlaciones y la matriz policrítica**, así como otras representaciones gráficas. Estas representaciones nos han proporcionado una visión detallada de las relaciones entre las variables.
- **Comprobación de la normalidad univariante**: Se ha realizado una exploración gráfica mediante **histogramas y gráficos qqplots**. Estos gráficos nos han proporcionado una perspectiva inicial sobre si las variables unidimensionales seguían una distribución normal. No obstante, la conclusión final sobre la normalidad se obtuvo utilizando el test de **Shapiro-Willks**.
- **Comprobación de la normalidad multivariante**: Se han utilizado test de hipotésis como el de **Royston** y el de **Henze-Zirkler**.

3. Resultados

Primero, hemos tenido cuidado con las variables codificadas que eran **categóricas** originalmente y que tienen más de dos niveles de respuesta porque quizá no interese incluirlas. En nuestro caso, estas son **Medu**, **Fedu**, **Mjob**, **Fjob**, **reason** y **guardian**. Para decidir si las incluímos o no, deberíamos preguntar al diseñador del dataset si esas variables son **relevantes** para el problema bajo estudio. Como no podemos contactar con él, asumimos que lo son. Así, una forma de ver si pueden ser útiles es hacer un **análisis descriptivo** de ellas y ver si los niveles están compensados (más o menos hay el mismo número de individuos en cada uno). Si lo están es una variable que podría aportar **información**, pero si hay algún nivel muy bajo o si todo se acumula en un nivel, habría que plantearse descartar la variable. Tras hacer este análisis hemos decidido eliminar las variables: **Medu**, **Fedu**, **Fjob** y **guardian**.

Luego, se estudian los **valores perdidos** en el dataset. Se observa que la variable **Mjob** tiene menos de un 5 %, por lo que estos se pueden sustituir directamente por la **media** de los valores conocidos. Sin embargo, la variable continua **studytime** tiene más de un 5 % de valores perdidos, así que se debe estudiar la **aleatoriedad** de los mismos mediante un test de Student, dando como resultado que **no se puede rechazar la hipótesis nula** de homogeneidad. En consecuencia, se concluye que el patrón es **aleatorio**. Por tanto, se ha decidido **sustituir** los valores perdidos de esta variable por la media de los valores conocidos.

También se detecta la presencia de **outliers**, y se toma la decisión de **sustituirlos por la media** de sus variables.

Tras el tratamiento de **outliers**, hay algunas variables que tienen el mismo valor para todos los registros, haciendo que su **varianza** sea cero, lo cual nos da problemas, por ejemplo, el test de **normalidad** univariante de Shapiro-Willks. En concreto, estas variables son **school**, **address**, **Pstatus**, **failures**, **schoolsup**, **nursery**, **higher** e **internet**. Por ello, como tenemos un gran número de variables, se ha decidido **eliminarlas**. Como resultado, trabajamos en lo que sigue con 19 variables (recordemos que inicialmente había 31 variables).

Pasamos a estudiar la **independencia** entre variables con el test de Bartlett, dando como resultado que no se puede rechazar la hipótesis de independencia entre variables. Por tanto, no tiene sentido plantearse una **reducción de la dimensión** del conjunto de variables mediante un **Análisis de Componentes Principales** o un **Análisis Factorial**.

Además, se ha concluido mediante el test de Saphiro-Willks que no se puede afirmar que haya **normalidad univariante** para la mayoría de las variables condicionadas a cada modalidad de la variable cualitativa estudiada en el Análisis Discriminante.

Se han usado también distintos tests para comprobar la **normalidad multivariante**, dando como resultado que tampoco se puede afirmar que la haya. Debido a la ausencia de estas hipótesis de normalidad, se ha decidido no tener en cuenta los resultados del Análisis Discriminante Lineal, pero sí los del Análisis Discriminante Cuadrático, por ser **más robusto** a esta ausencia.

Para el Análisis Discriminante se ha definido una variable respuesta categórica a partir de la **calificación del estudiante** almacenada en la variable G3. Esta variable categórica toma el valor aprobado si el valor de G3 es mayor o igual a 10.0 y suspenso si es menor a 10.0. Por los motivos expuestos en el párrafo anterior, vamos a centrarnos exclusivamente en los resultados del Análisis Discriminante Cuadrático, para el cual se ha obtenido un porcentaje de error del 31.65 % (es decir, tenemos un porcentaje de acierto del 68.35 %).

A continuación, se muestran los tres modelos de clasificación cuadráticos con dos predictores que tienen un porcentaje de error más bajo:

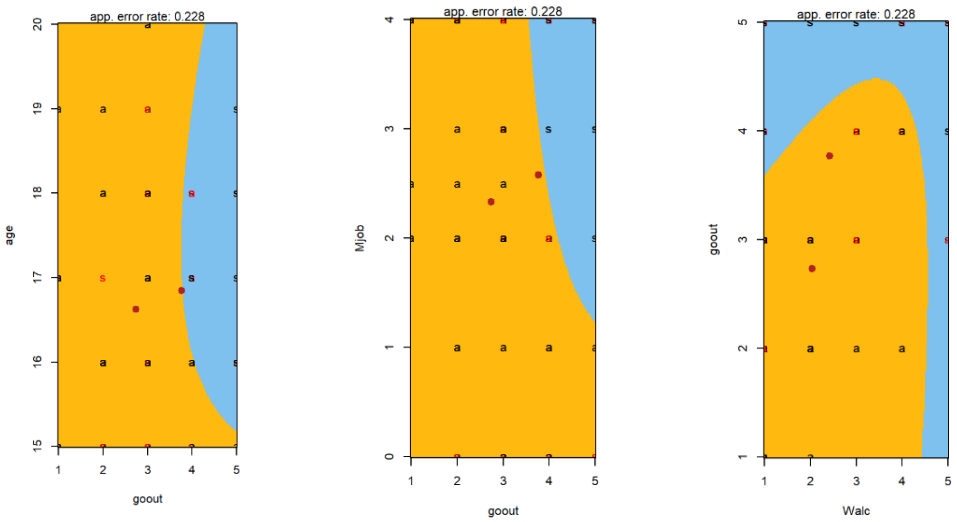

Figura 1: Modelos de clasificación cuadráticos con dos predictores que tienen un porcentaje de error más bajo

Observamos que estos tres modelos con dos predictores tiene un porcentaje de error **más bajo que el modelo en el que usamos todas las variables** del dataset como predictores (excluyendo obviamente la variable G3, usada para definir la variable respuesta). Por tanto, nos interesará utilizar **menos predictores** con el fin de conseguir el menor porcentaje de error posible, además de un **menor coste** en recogida y computación de los datos.

4. Discusión

El estudio se enfocó en datos correspondientes a 31 variables, los cuales fueron evaluados en 395 estudiantes diferentes. El propósito principal de la investigación consistió en **implementar una reducción de la dimensionalidad y determinar las variables predictoras más relevantes** para determinar si un estudiante aprueba o suspende.

Tras aplicar el test correspondiente, no pudimos descartar la independencia de las variables. Por tanto, **no realizamos una reducción de la dimensión** mediante el procedimiento de Análisis de Componentes Principales o Análisis Factorial.

Finalmente se ha querido **clasificar** los datos en función de la calificación final del estudiante en la asignatura de Matemáticas, teniendo en cuenta el resto de variables del dataset como variables predictoras y con el objetivo de, dados los datos de un nuevo estudiante, poder determinar si aprueba o suspende. Entonces, mediante un Análisis Discriminante Cuadrático se ha obtenido un **método de clasificación** para nuevas observaciones.

Cabe destacar que, en base a los resultados obtenidos, la **variable goout** (frecuencia con la que el estudiante sale a la calle con sus amigos) es muy **influyente** a la hora de determinar la calificación final del estudiante.

5. Conclusión

En este trabajo se ha estudiado una serie de datos relacionados con el ámbito social y escolar de diferentes estudiantes. Se ha obtenido un **método de clasificación** de la calificación final del estudiante en la asignatura de Matemáticas en función de dichos datos, lo cual puede ayudar a mejorar las políticas educativas de los centros escolares con el fin de **incrementar el rendimiento académico** y brindar un apoyo más personalizado a los alumnos en sus estudios de Matemáticas en función de la **situación** de cada estudiante.

Como limitaciones existentes, muchos supuestos necesarios para aplicar algunas técnicas **no se dan**, y aunque se ha podido decir poco, **no se pueden desprender más conclusiones** de los datos mediante las técnicas propuestas. Sin embargo, se considera que usando otras técnicas más robustas como **regresión logística** se podrían obtener mejores resultados.