

Propuesta del proyecto

Dataset

<https://www.kaggle.com/larsen0966/student-performance-data-set?select=student-por.csv>

Descripción del Dataset

Esta data se obtiene de colegios de secundaria de Portugal, los atributos del dataset incluyen información de los estudiantes, características demográficas y sociales. En el dataset se proporciona el desempeño de dos materias (matemática y lenguaje, que es este caso es portugués) en tres distintas épocas, el primer periodo del año, el segundo periodo del año y el último periodo del año, estas calificaciones se clasifican en 5 niveles.

Descripción del proyecto

Consiste en crear un modelo de regresión para poder predecir la nota que obtendrá el alumno en el último periodo del año.

Utilidad del modelo

Puede ser de gran ayuda para prestarle más atención a ciertos estudiantes que pueden llegar a perder el año o bajar su rendimiento académico.

También puede ser de utilidad para extraer hallazgos de qué características le afectan a los alumnos en su rendimiento académico

Variables (33)

Categorías binarias (13)

1. School: A qué escuela pertenece el estudiante 'GP' - Gabriel Pereira o 'MS' - Mousinho da Silveira
2. Sex: El género del estudiante
3. Address: Si el estudiante vive en una zona urbana o rural
4. Famsize: El tamaño de la familia, si es menor o igual a tres se cataloga como familia pequeña o si es mayor a 3 se cataloga como familia grande
5. Pstatus: El estado de cohabitación de sus padres, si viven juntos o si están separados
6. Schoolsup: Si tienen educación extracurricular
7. Famsup: Si la familia tiene un apoyo educativo
8. Paid: si pagan por clases extras de la materia
9. Actividades: si el alumno tiene actividades extracurriculares
10. Nursery: si asistieron a la guardería (prekinder) de bebés
11. Higher: si el alumno desea recibir educación superior
12. Internet: si el alumno tiene acceso a internet en su casa
13. Romantic: Si el alumno tiene alguna relación romántica

Categóricas multiclase (15):

1. Medu: Variable ordinal, educación de la madre, de 0 a 4, siendo 0 sin educación superior
2. Fedu: Variable ordinal, educación del padre, de 0 a 4, siendo 0 sin educación superior
3. Mjob: variable nominal, trabajo de la madre
4. Fjob: variable nominal, trabajo del padre
5. Reason: variable nominal, razón por la cual escogieron esa escuela
6. Guardian: quién es el tutor del estudiante, mamá, papá y otro
7. Traveltime: variable ordinal, cuanto tiempo le toma al alumno llegar de su casa a la escuela, ya está categorizada
8. Study time: variable ordinal, cuanto tiempo estudia el alumno a la semana, ya está categorizada
9. Failures: variable ordinal, cuantas clases ha perdido el alumno, de 0 a 4 o más
10. Famrel: variable ordinal, que tan buena relación tienen con su familia, de 0 a 5 siendo 5 excelente
11. Freetime: variable ordinal, tiempo libre del alumno después del colegio, de 1 a 5, siendo 5 mucho tiempo
12. Goout: variable ordinal, qué tanto sale con sus amigos, de 1 a 5 siendo 5 muchas veces
13. Dalc: variable ordinal, consumo de alcohol en horario laboral de 1 a 5 siendo 5 muchas veces
14. Walc: variable ordinal, consumo de alcohol entre semana, de 1 a 5 siendo 5 muchas veces
15. Health: variable ordinal, estado actual de su salud, de 1 a 5 siendo 5 excelente

Numéricas (5):

1. Age: Edad del alumno
2. Absences: Ausencias en el año
3. G1: nota en el primer periodo
4. G2: nota en el segundo periodo
5. G3: nota en el tercer periodo

Registros

Cuenta con 650 registros, sería bueno que tuviera por lo menos 1000 registros para poder entrenar bien el modelo y poderle tener más confianza a la hora de predecir, aun que considero que la data recolectada es de muy buena calidad por lo cual creo que podremos generar un buen modelo