Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Computación

Máquinas de aprendizaje

Reporte: Ejemplo de Naive Bayes (otro)

Docente: Abraham Sánchez López

Alumno	Matrícula
Taisen Romero Bañuelos	202055209

Naive Bayes - Ejemplo

Como repaso, se utilizó Naive Bayes para predecir a cuál de los tres programas (general, academic, vocation) debería pertenecer un estudiante, basándose en sus puntuaciones de las asignaturas (read, write, math, science, socst).

Parte del análisis exploratorio es ver la distribución de los datos con la función summary(). Aquí me percaté que no fue necesario un proceso de normalización ya que los datos son cercanos entre ellos, lo cual tiene sentido ya que se trata de puntajes de asignaturas escolares.

```
      read            write           math
 Min.   :28.00   Min.   :31.00   Min.   :33.00
 1st Qu.:44.00   1st Qu.:45.75   1st Qu.:45.00
 Median :50.00   Median :54.00   Median :52.00
 Mean   :52.23   Mean   :52.77   Mean   :52.65
 3rd Qu.:60.00   3rd Qu.:60.00   3rd Qu.:59.00
 Max.   :76.00   Max.   :67.00   Max.   :75.00
```

Ahora, respecto a los resultados de las predicciones obtuvimos una precisión no tan buena (del 56%). Esto me lleva a dos hipótesis.

1. La precisión relativamente baja se debe a que sólo contamos con 200 observaciones y por lo tanto, con más datos podríamos mejorar los resultados.
2. La distribución de datos para entrenamiento y prueba debería ser distinta debido a esa falta de datos (90/10 fue una proporción sugerida para manejar pocos datos).
3. Otro tipo de modelo sería más efectivo en esta tarea.

Sobre mi primera hipótesis no hay mucho que hacer, pero para verificar la veracidad de la segunda cambié las proporciones y obtuve los siguientes resultados:

```
Confusion Matrix for Training Data
> print(trainTable)
          trainPred
           general academic vocation
  general        7       16       18
  academic       7       69       19
  vocation       8        6       31
```

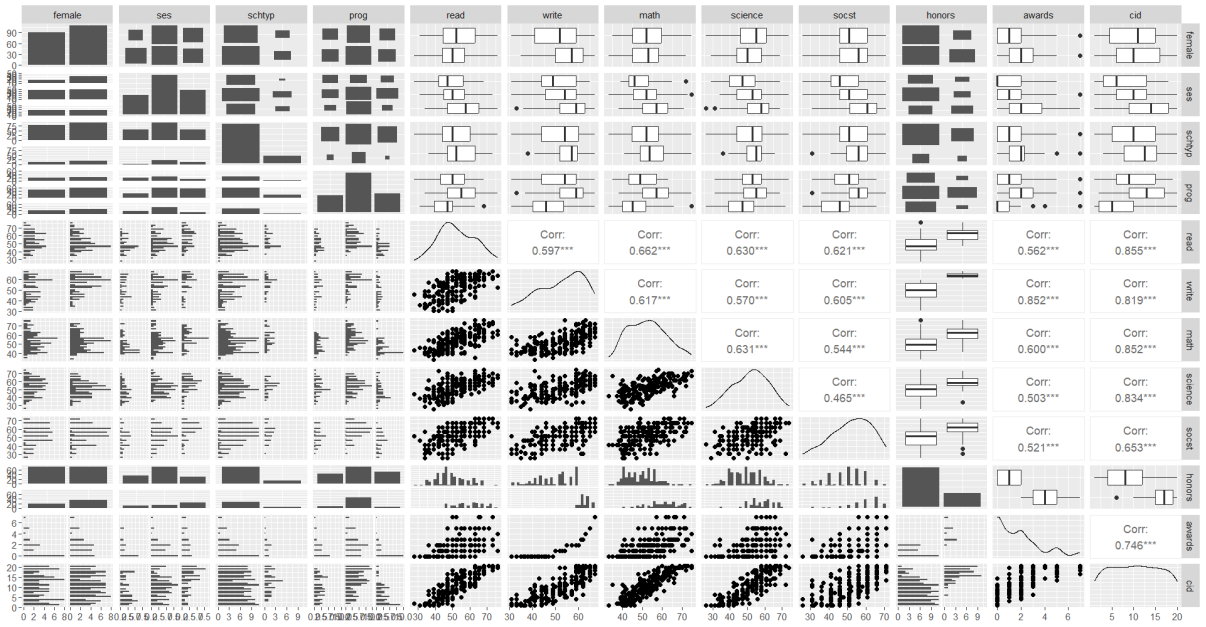
```
Confusion Matrix for Test Data
> print(testTable)
          testPred
           general academic vocation
  general        1        1        2
  academic       2        6        2
  vocation       1        1        3
```

```
Accuracy
> print(round(cbind(trainAccuracy
     trainAccuracy testAccuracy
[1,]         0.591        0.526
```

Los resultados son un poco peores que los que se obtuvieron con 70/30, así que podemos descartar esta opción.

Y sobre la tercera hipótesis, mi razón para formularla es que los errores en la matriz de confusión podrían reflejar que las clases tienen atributos similares, lo que dificultará la tarea de diferenciarlas, especialmente con las correlaciones moderadas que existen entre las variables de los grupos. Veamos por qué menciono esto de la correlación entre variables.

Matriz de correlación (general)



1. Correlation matrix within general

```
[1] "Correlation matrix for General:"
> print(cor_general)
            read      write      math   science      socst
read    1.0000000 0.4739121 0.3945974 0.6586988 0.5418732
write   0.4739121 1.0000000 0.3586417 0.5629392 0.6505204
math    0.3945974 0.3586417 1.0000000 0.5752819 0.3787115
science 0.6586988 0.5629392 0.5752819 1.0000000 0.4222026
socst   0.5418732 0.6505204 0.3787115 0.4222026 1.0000000
```

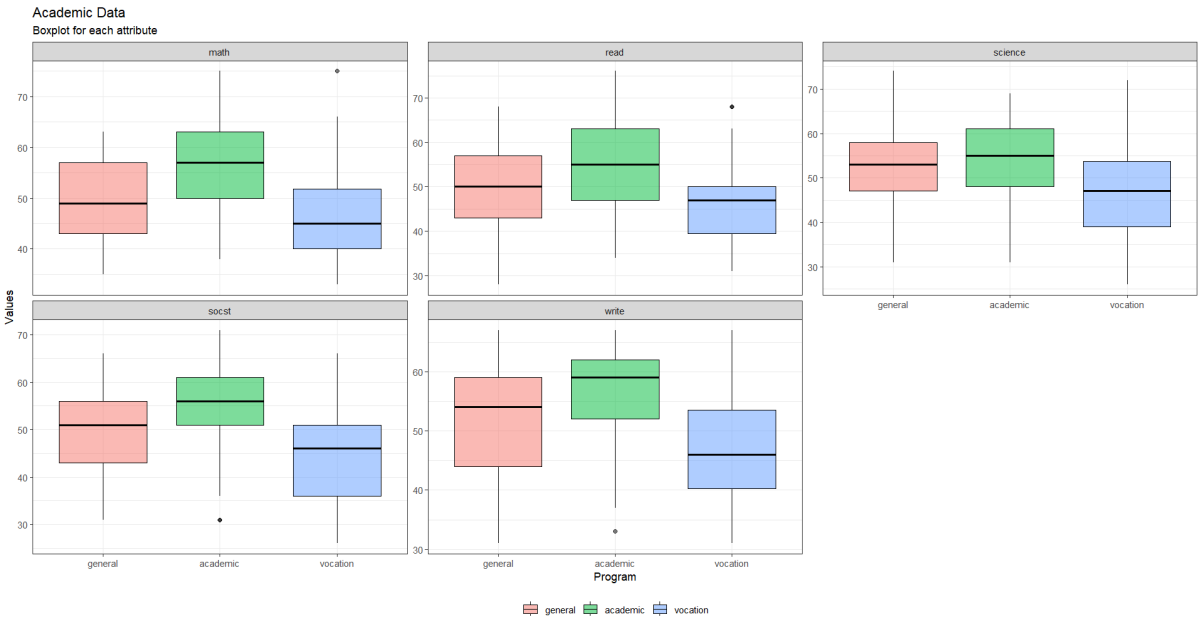
2. Correlation matrix within academic

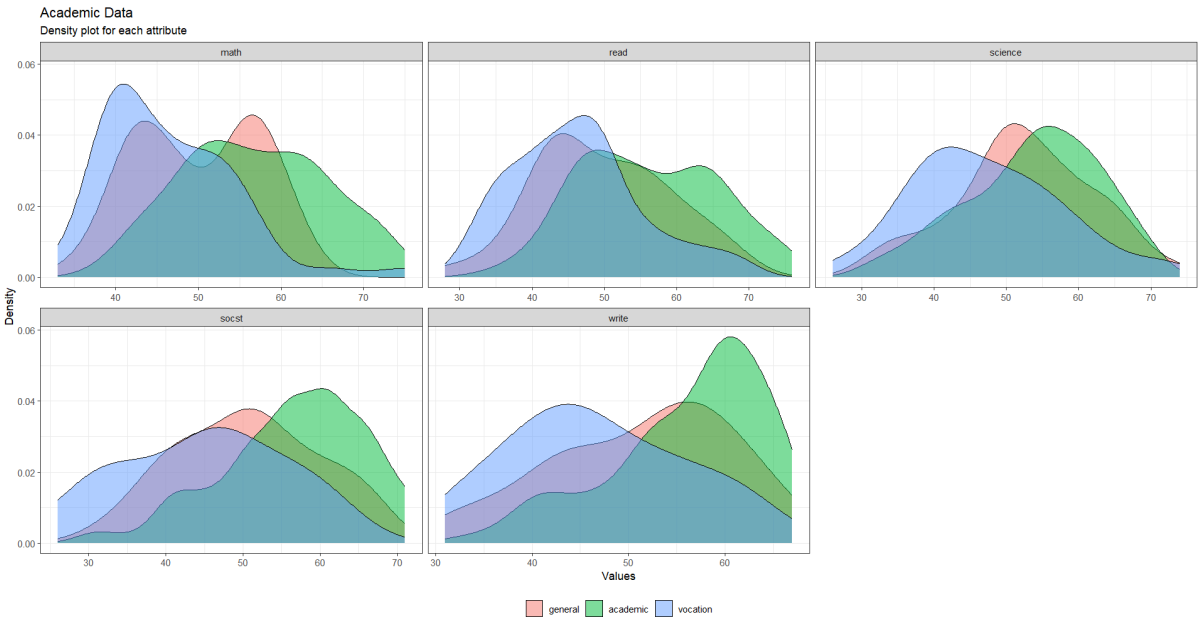
```
[1] "Correlation matrix for Academic:"
> print(cor_acad)
            read      write      math   science      socst
read    1.0000000 0.5608413 0.6917634 0.6250391 0.5851566
write   0.5608413 1.0000000 0.6130255 0.5128848 0.4538175
math    0.6917634 0.6130255 1.0000000 0.6410174 0.4591657
science 0.6250391 0.5128848 0.6410174 1.0000000 0.4383806
socst   0.5851566 0.4538175 0.4591657 0.4383806 1.0000000
```

3. Correlation matrix within vocational

```
[1] "Correlation matrix for Vocation:"
> print(cor_voc)
             read     write      math   science     socst
read    1.0000000 0.4615702 0.4570520 0.5132068 0.4325037
write   0.4615702 1.0000000 0.5090928 0.5225355 0.4926333
math    0.4570520 0.5090928 1.0000000 0.5706508 0.3769207
science 0.5132068 0.5225355 0.5706508 1.0000000 0.3348232
socst   0.4325037 0.4926333 0.3769207 0.3348232 1.0000000
```

4. Boxplots y gráficos de densidad





Bueno, luego de tanto gráfico podemos notar que la clase "General" tiene correlaciones entre las materias moderadas, sobre todo entre **read** y **science**, y entre **write** y **socst**. Para la clase "Academic" las correlaciones son incluso más fuertes que en General, hay una relación significativa entre **read**, **math**, **science**, y **socst**. Y para la clase "Vocation" las correlaciones son más débiles pero todavía presentes. Además, según los diagramas de dispersión y los boxplot podemos notar que la clase "academic" tiende a tener una distribución de puntuaciones más alta y con menos dispersión, lo que sugiere que los estudiantes tienden a obtener mejores calificaciones en comparación con los grupos "general" y "vocation".

Ok, pero, ¿esto qué tiene que ver con que Bayes puede no ser la mejor opción para nuestro problema?, pues al descubrir las fuertes correlaciones entre variables también descubrimos que el supuesto de independencia de Bayes se echa a perder. Por ejemplo, si un estudiante tiene un buen rendimiento en una asignatura, es probable que también tenga un buen rendimiento en otras, lo que rompe la suposición de independencia que Naive Bayes hace entre las variables. Además, en los gráficos de distribución notamos que hay datos sesgados, lo que también puede empeorar la capacidad predictiva de Bayes.

La distribución sesgada puede arreglarse con técnicas para manejar los datos, pero la dependencia entre variables es mejor tratarla con un modelo que no haga ese supuesto de independencia, como árboles de decisión.

Como extra, dejaré este histograma para ver qué tan cierta es su correlación.