一、 摘要

為了瞭解家長背景、考試準備等對學生的表現是否有影響，進而選擇了「Students Performance in Exams」這個題目來做分析。

二、 介紹(研究背景及研究目的)

在求學階段，每位同學都來自不同的家庭背景，各有各的讀書方法，有人天資聰穎、有人則是後天努力，但也有人埋怨因家庭經濟狀況而沒有受到好的教育，在這樣的狀況下，想藉由本次的分析探討成績的高低是否與個人性別、父母的學歷和考前有無準備之相關性。

三、 資料集介紹(含資料特徵)及資料集來源

「Students Performance in Exams」的資料集來自 kaggle，它是一個數據建模和數據分析的競賽平台。本資料集共有 8 個欄位，分別是性別、種族/族裔、父母的教育水準、午餐、考試準備課程、數學分數、閱讀分數和寫作分數等，而下圖為各欄位的資料特徵。

[4]:
```
col=["gender", "race/ethnicity","parental level of education", "lunch", "test preparation course"]
for item in col:
    print(item.upper(),":")
    print(df[item].value_counts(),"\n")
```

```
GENDER :
female    518
male      482
Name: gender, dtype: int64

RACE/ETHNICITY :
group C    319
group D    262
group B    190
group E    140
group A     89
Name: race/ethnicity, dtype: int64

PARENTAL LEVEL OF EDUCATION :
some college          226
associate's degree    222
high school           196
some high school      179
bachelor's degree     118
master's degree        59
Name: parental level of education, dtype: int64

LUNCH :
standard        645
free/reduced    355
Name: lunch, dtype: int64

TEST PREPARATION COURSE :
none         642
completed    358
Name: test preparation course, dtype: int64
```

四、 資料預處理

針對資料完整性去過濾不符合規則的資料，在此可得知資料集並無缺失。

In [194]:
```
#missing data
total = df.isnull().sum().sort_values(ascending=False)
percent = (df.isnull().sum()/df.isnull().count()).sort_values(ascending=False)
missing_data = pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])
missing_data.head(20)
```

Out[194]:

	Total	Percent
writing score	0	0.0
reading score	0	0.0
math score	0	0.0
test preparation course	0	0.0
lunch	0	0.0
parental level of education	0	0.0
race/ethnicity	0	0.0
gender	0	0.0

五、 機器學習或深度學習方法(使用何種方法)

1. Decision tree

[11]:
```
from sklearn.tree import DecisionTreeClassifier
from sklearn import metrics

tree = DecisionTreeClassifier(criterion='gini',max_depth=5)
tree.fit(X_train, y_train)
print(metrics.classification_report(y_test, tree.predict(X_test)))
```

```
             precision    recall  f1-score   support

     female       0.79      0.80      0.79        98
       male       0.80      0.79      0.80       102

avg / total       0.80      0.80      0.80       200
```

2. LogisticRegression

[12]:
```
from sklearn.linear_model import LogisticRegression
from sklearn import metrics

lr = LogisticRegression()
lr.fit(X_train, y_train)
print(metrics.classification_report(y_test, lr.predict(X_test)))
```

```
             precision    recall  f1-score   support

     female       0.77      0.94      0.85        98
       male       0.93      0.74      0.82       102

avg / total       0.85      0.83      0.83       200
```

3. KNN

```
[13]: from sklearn.neighbors import KNeighborsClassifier

knn = KNeighborsClassifier(n_neighbors=2)
knn.fit(X_train, y_train)
print(metrics.classification_report(y_test, knn.predict(X_test)))
```

	precision	recall	f1-score	support
female	0.74	0.93	0.82	98
male	0.91	0.69	0.78	102
avg / total	0.83	0.81	0.80	200

4. SVM

```
In [209]: from sklearn.svm import SVC

svc = SVC(C=1.0, kernel="rbf")
svc.fit(X_train, y_train)
print(metrics.classification_report(y_test, svc.predict(X_test)))
```

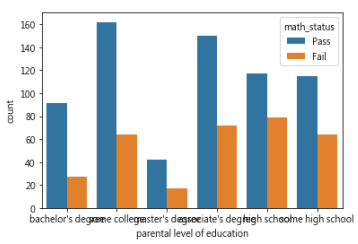
	precision	recall	f1-score	support
female	0.92	0.88	0.90	109
male	0.86	0.91	0.89	91
avg / total	0.90	0.90	0.90	200

六、 研究結果及討論(含模型評估與改善)

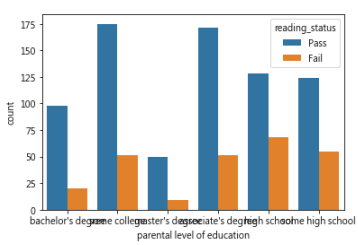
比較四個方法的分析結果後，我們可以發現使用SVM方法得到最佳的結果，因為我們可以推斷出這些特徵是線性可分離的。

將成績進一步處理成有無及格兩類，接著與父母親的學歷相互對應後發現學生的考試成績是否及格與父母親的學歷並沒有很直接的關係。

```
#數學成績
sns.countplot(x='parental level of education', data = df, hue=df['math_status'])
plt.show()
```

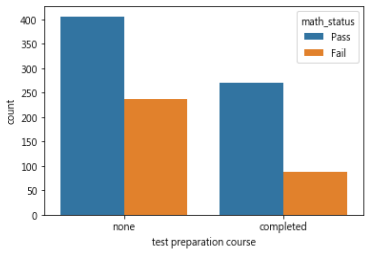



```
#閱讀成績
sns.countplot(x='parental level of education', data = df, hue=df['reading_status'])
plt.show()
```

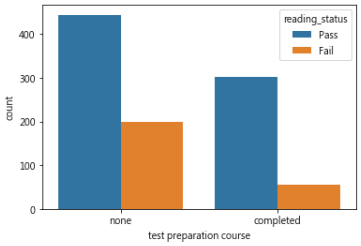


下圖為考前有無事先做準備對數學成績與閱讀成績的影響，從圖表得知不管有無準備的及格率都很高，但有做準備的人很明顯的不及格率相對低很多。

```
#數學成績
sns.countplot(x='test preparation course', data = df, hue=df['math_status'])
plt.show()
```




```
#閱讀成績
sns.countplot(x='test preparation course', data = df, hue=df['reading_status'])
plt.show()
```



結論

由上述的方法分析後得知，學生的成績與父母的學歷並沒有很直接的關係，有無事前做準備相對重要許多!而透過建立模型根據三項成績來預測性別，藉由不同的分析方法找出最佳結果， 讓我更加了解到資料集的相關特徵。

參考文獻

https://www.kaggle.com/spscientist/students-performance-in-exams

https://www.kaggle.com/aerodinamicc/eda-regression-based-on-knn