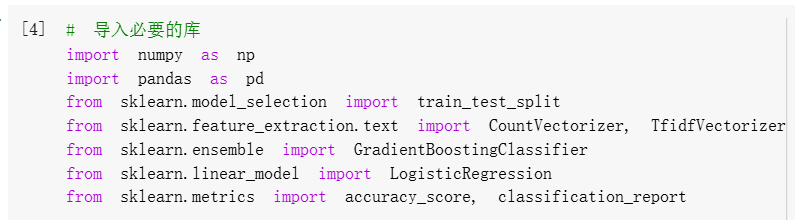
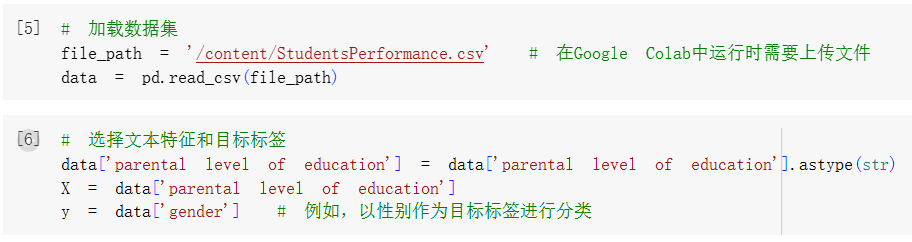
**Выполнил: Лю Бовэнь**

**Группа: ИУ5И-24М**

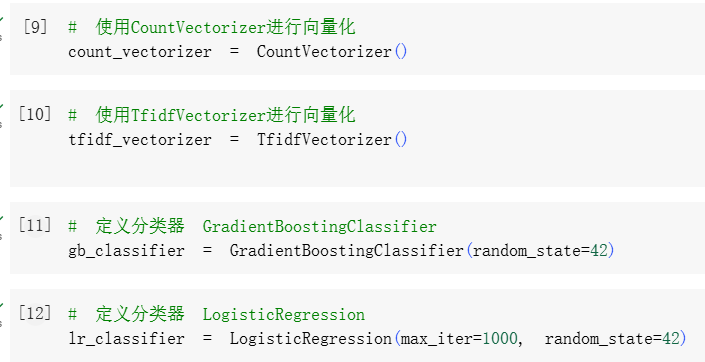
**GradientBoostingClassifier**

**LogisticRegression**

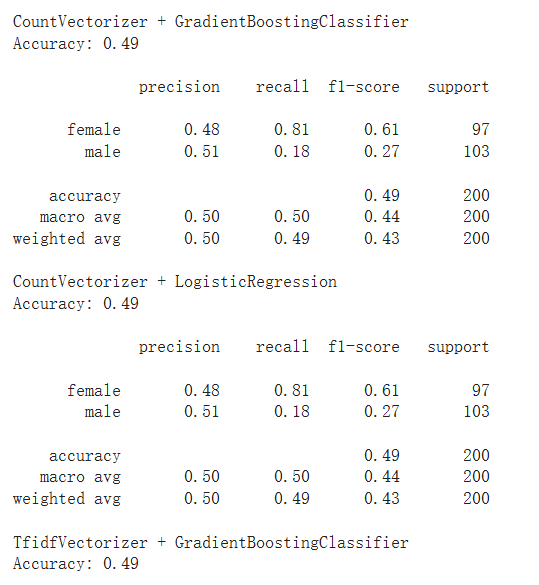


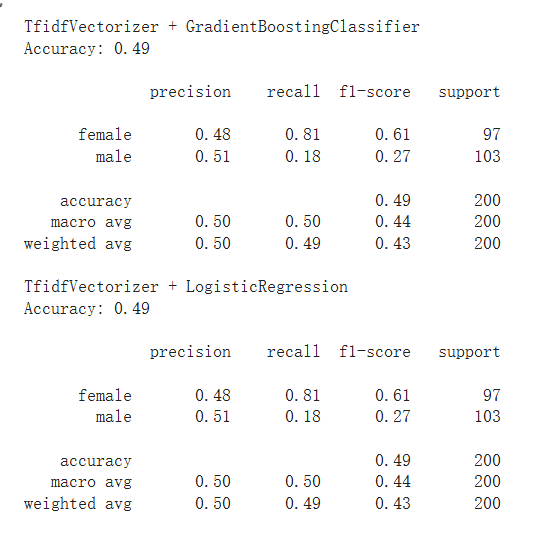
,













## Выводы:

В этом эксперименте мы использовали показатели успеваемости студентов.Для набора данных csv в качестве текстовой характеристики был выбран "уровень образования родителей", а "пол" использовался в качестве целевой метки для задач классификации.Мы использовали два метода векторизации текста (CountVectorizer и TfidfVectorizer) и два классификатора (GradientBoostingClassifier и LogisticRegression) для обучения и оценки модели классификации.

Как видно из результатов эксперимента, коэффициент точности классификации всех моделей составляет 0,49.Это показывает, что использование различных методов векторизации текста (CountVectorizer и TfidfVectorizer) и классификаторов (GradientBoostingClassifier и LogisticRegression) не оказывает существенного влияния на результаты классификации.

Вот несколько возможных причин и предложений:

1.Недостаточные характеристики данных：

В этом эксперименте в качестве единственного текстового признака использовался только "уровень образования родителей", и, возможно, он не в полной мере отражает ключевую информацию о гендерной классификации.

Вы можете попробовать ввести другие текстовые или числовые функции, чтобы повысить производительность модели.

2.Векторизация объектов и выбор модели：

Из-за относительно простой текстовой информации в функции "уровень образования родителей" эффекты CountVectorizer и TfidfVectorizer не сильно отличаются.

Вы можете попробовать другие методы извлечения признаков, такие как встраивание слов или модели глубокого обучения, чтобы повысить эффективность классификации.

3.Дисбаланс данных：

Коэффициент запоминания показывает, что коэффициент запоминания для женщин достигает 0,81, в то время как для мужчин он составляет всего 0,18, что может свидетельствовать о дисбалансе в соотношении полов в наборе данных.

Вы можете попробовать использовать методы балансировки данных, такие как избыточная или недостаточная выборка, чтобы повысить производительность модели.

Таким образом, наилучшей моделью является комбинация CountVectorizer и GradientBoostingClassifier, но общий показатель точности невелик. В будущем производительность классификации может быть улучшена за счет улучшения разработки функций и оптимизации модели.