Master in Computer Science (MCS)

Machine Learning Techniques

MLT

Tutor: Dr: Ubai Sandouk

Project by:

Id&class	Name
Ahmad_250641 / C1	Ahmad Alwareh
Lava_255656 / C1	Lava Mirkhan
Mahmoud_248123 / C1	Mahmoud Shourbaji

1. البيانات المستخدمة والهدف من النماذج

البيانات التي تم استخدامها في التدريب تحتوي على مجموعة من الميزات المستقلة مع "الراتب الشهري" كهدف تنبؤي. الهدف من هذه النماذج هو بناء نماذج انحدار قادرة على التنبؤ بالراتب الشهري بناءً على الميزات الأخرى، وقد تم استخدام ثلاثة نماذج مختلفة: الانحدار الخطي، شجرة القرار، وغابة عشوائية

2. أداء النماذج: تحليل "معدل الخطأ التربيع (MSE) وR^2

- الانحدار الخطي Linear regression:

حصل الانحدار الخطي على MSE عالي جدا، بلغ حوالي 9 تريليون في كلا الجولاتين (9.066 تريليون و9.066 تريليون على التوالي)، مما يشير إلى أن النموذج يتوقع فروقات كبيرة مقارنةً بالقيم الحقيقية للراتب الشهري. معامل التحديد R^2 كان حوالي 0.67، مما يعني أن حوالي 67٪ من التباين في البيانات يمكن تفسيره من قبل هذا النموذج، بينما يبقى جزء كبير من التباين غير مفسر.

- شجرة القرار decision tree :

أظهرت شجرة القرار أداءً أفضل بكثير مع MSE بلغ حوالي 1.3 تريليون في الجولة الأولى و1.33 تريليون في الجولة الثانية، مما يدل على دقة أعلى من الانحدار الخطي. معامل التحديد R^2 كان حوالي 0.95 في كلتا الجولات، مما يشير إلى أن شجرة القرار كانت قادرة على تفسير حوالي 95٪ من التباين في البيانات.

- الغابة العشوائية Random Forest:

كان أداء الغابة العشوائية هو الأفضل بين النماذج الثلاثة، حيث بلغ MSE حوالي 773 مليار في الجولة الأولى و715 مليار في الجولة الثانية. كما بلغ معامل التحديد R^2 حوالي 0.97، مما يعني أن النموذج كان قادرًا على تفسير حوالي 97٪ من التباين. يعكس هذا الأداء الممتاز قدرة الغابة العشوائية على التعامل مع البيانات بشكل أكثر تعقيدًا مقارنةً بالنماذج الأخرى.

3. التحقق المتقاطع (Cross-validation)

تحليل التحقق المتقاطع يعطينا فكرة عن استقرار ودقة النماذج على عينات مختلفة من البيانات:

- الانحدار الخطي Linear regression:

نتائج التحقق المتقاطع للانحدار الخطي كانت سلبية حيث كانت النتائج تتراوح بين -8.5 تريليون و-9.8 تريليون، مما يعكس تقلبات كبيرة وضعف في أداء النموذج عبر العينات المختلفة. هذا يشير إلى أن الانحدار الخطي قد لا يكون الخيار الأمثل مع البيانات الحالية، خاصةً إذا كانت تحتوي على تفاعلات معقدة بين الميزات.

- شجرة القرار decision tree :

أظهرت شجرة القرار تحسنًا ملحوظًا في نتائج التحقق المتقاطع مقارنةً بالانحدار الخطي، حيث تراوحت القيم بين -664 مليار و-1.73 تريليون في الجولة الأولى، وبين -596 مليار و-1.59 تريليون في الجولة الثانية. هذا الأداء المتقلب يعكس أن شجرة القرار قد تتأثر بالتغيرات الطفيفة في البيانات.

- الغابة العشوائية Random Forest:

حققت الغابة العشوائية أدنى قيم للـ MSEفي التحقق المتقاطع، مما يدل على استقرار عالٍ ودقة عالية في التوقعات. تراوحت القيم بين -476 مليار و-995 مليار في الجولة الأولى، وبين -489 مليار و-976 مليار في الجولة الثانية. يعكس هذا أن الغابة العشوائية نموذج أكثر استقرارًا.

4. **الاستنتاجات والتوصيات**

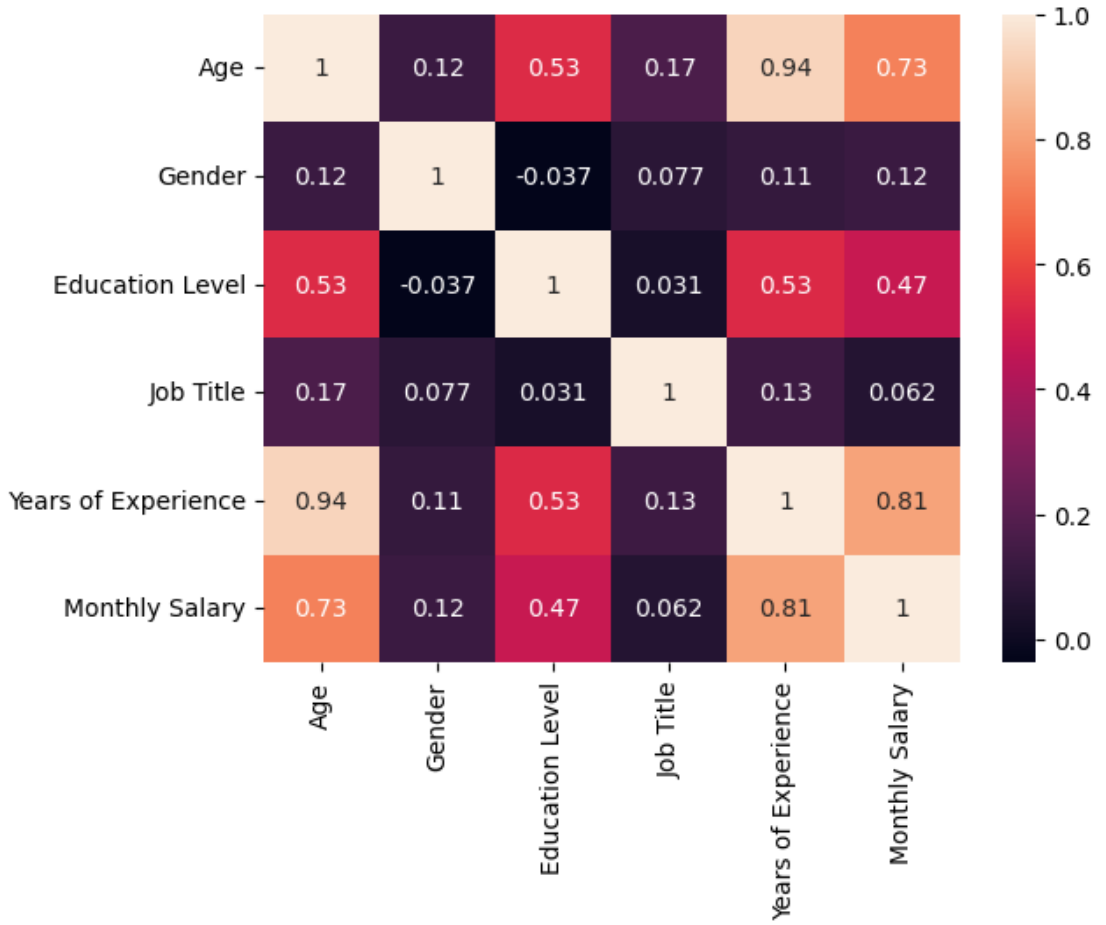
تظهر النتائج أن نموذج الغابة العشوائية يوفر دقة واستقرارًا أعلى مقارنةً بالنماذج الأخرى، ويرجع ذلك إلى قدرتها على الاستفادة من التكرار والدمج لتقليل التباين في التوقعات.

5. **أمثلة ضمن الحل:**

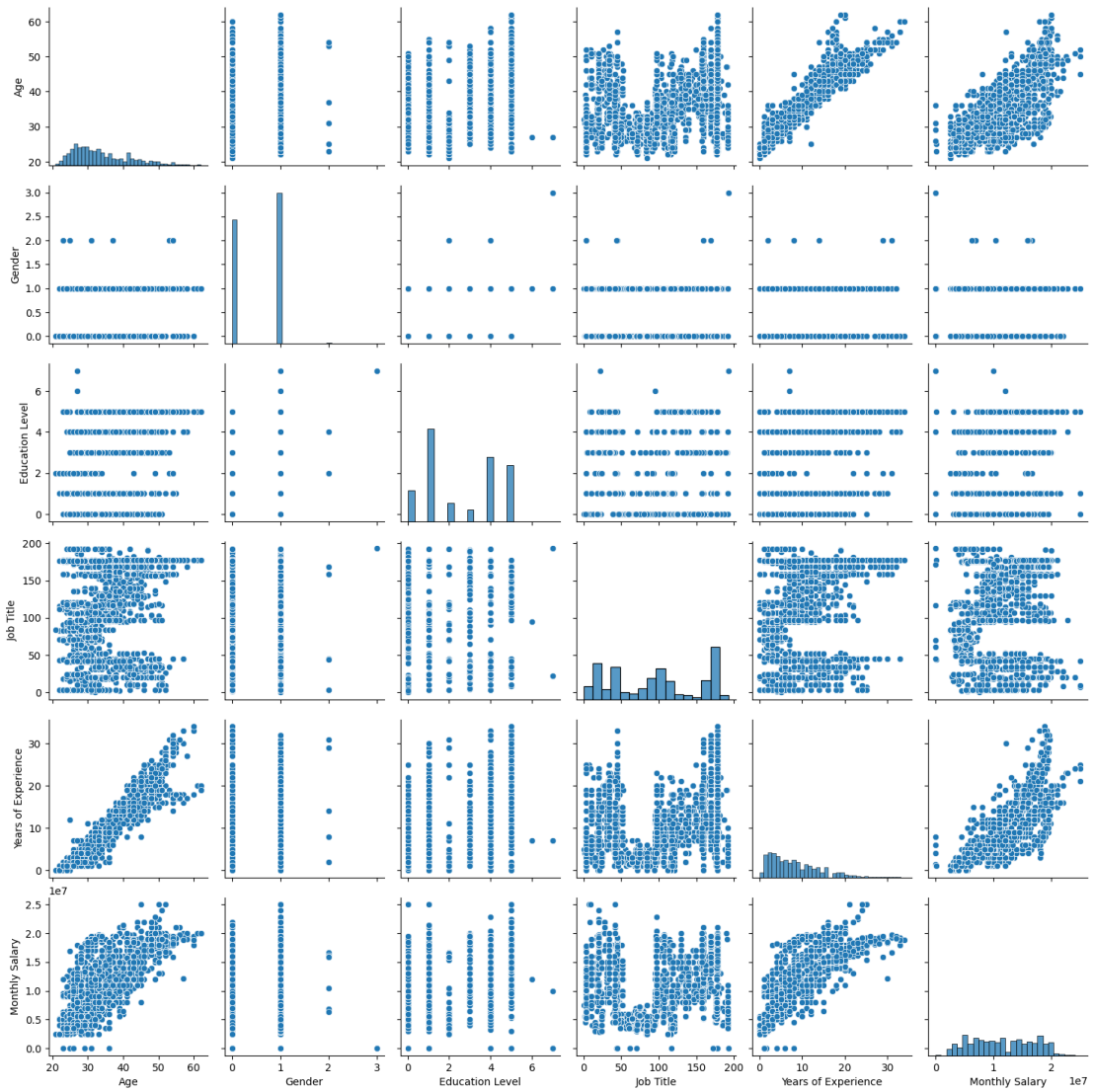
الكود المطلوب لعرض 10 وظائف ذات أعلى دخل:

```
top_10_jobs = data.groupby('Job Title')['Monthly Salary'].mean().nlargest(10)
print(top_10_jobs)
```

مثال عن ايجاد صلة الربط والعلاقات بين البيانات correlations



استخدام تقنية box plot لعرض الروابط بين البيانات المدخلة مثل العمر و الطبيعة العمل والراتب



ملاحظات:

- التقدم في العمر مرتبط بزيادة الراتب بشكل خطي
- عدد سنوات الخبرة تعلب دور في الراتب الشهري
- حيث ازدياد عدد سنوات الخبرة ملحوظ بزيادة الراتب الشهري
- بشكل طبيعي تزداد سنوات الخبرة مع تقدم العمر
- نلاحظ ارتباط بين المسمى الوظيفي مع التقدم في العمر