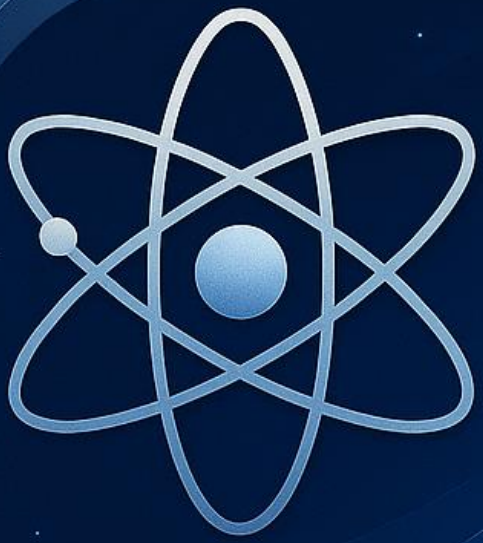

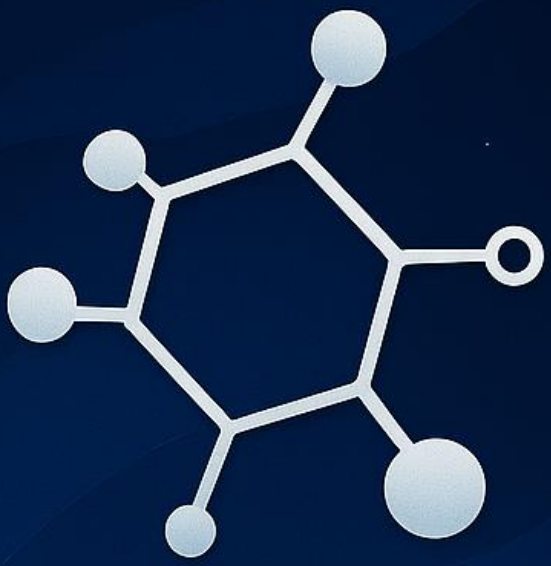

حل ورقة عمل

Alaa Saood

ورقة العمل الثانية ضمن المسار

يتضمّن ملف Startups.csv_50 بيانات خمسين شركة ناشئة؛ وتشمل: إنفاق البحث والتطوير (R&D Spend)، الإنفاق الإداري (Administration)، الإنفاق التسويقي (Marketing Spend)، الولاية (State)، وأخيراً الأرباح (Profit) خلال سنة مالية واحدة. الهدف هو بناء نموذج انحدار خطي متعدّد (MLR) للتنبؤ بالأرباح، وتحديد:

1): أيّ شركة حقّقت أعلى ربح؟

2): ما العامل الأكثر تأثيراً في الربح؟

3): ما مدى دقّة النموذج؟

الحل:

1): يوضح الجدول التالي حلول الأسئلة السابقة:

رقم السؤال	الإجابة	شرح مختصر
1	الشركة ذات السجلّ الأوّل بقيمة ربح ≈ 192 262 $.	-
2	إنفاق البحث والتطوير (R&D Spend)	يملك أعلى معامل مطلق (+0.77) وأدنى قيمة p (أقل من 0.001) في تحليل OLS.
3	$R^2 \approx 0.899$، RMSE ≈ 9 000 $	يفسّر النموذج 93.5 ٪ من التباين؛ الخطأ الجذري منخفض نسبياً

2): الكود البرمجي النهائي:

الكود البرمجي النهائي لهذا النموذج متوفر هنا.

خطوات العمل:

1): المعالجة المسبقة للبيانات (Data Pre-processing):

- تحميل البيانات من الملف Startups.csv_50 الذي يحتوي على بيانات 50 شركة ناشئة.
- تحويل العمود النصي State إلى أعمدة رقمية باستخدام تقنية one-hot encoding، وذلك لأن خوارزميات الانحدار لا تتعامل مباشرة مع البيانات النصية.
- فصل الأعمدة: استخدمنا الأعمدة R&D Spend, Administration, Marketing Spend, و State كمدخلات (Features)، واعتبرنا Profit هو المتغيّر الهدف (Target).
- تطبيع البيانات باستخدام StandardScaler لتقليل التفاوت بين القيم العددية، مما يُحسّن تدريب النموذج ويجعل تفسير المعاملات أكثر وضوحاً.
- تقسيم البيانات إلى مجموعة تدريب (80٪) واختبار (20٪) باستخدام train_test_split لضمان التقييم العادل لأداء النموذج.

2): تدريب نموذج الانحدار الخطي المتعدّد (Fitting the MLR Model):

- بعد تحضير البيانات، قمنا بتدريب النموذج باستخدام خوارزمية الانحدار الخطي المتعدد (Multiple Linear Regression) من مكتبة scikit-learn.
- هذا النوع من النماذج يحاول إنشاء علاقة رياضية بين الربح (Profit) وكل من المدخلات الأربعة.
- الصيغة العامة التي يبنيها النموذج:

$$\begin{aligned} \boldsymbol{Profit} \approx\ & \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \times (\boldsymbol{R\&D\ Spend}) \\ & + \boldsymbol{\beta}_2 \times (\boldsymbol{Administration}) \\ & + \boldsymbol{\beta}_3 \times (\boldsymbol{Marketing\ Spend}) \\ & + \boldsymbol{\beta}_{4,5} \times (\boldsymbol{State\ variables}) \end{aligned}$$

3): التنبؤ على بيانات الاختبار (Predicting the Result):

- بعد تدريب النموذج، قمنا باستخدامه لتوقّع الأرباح (Profit) على بيانات الاختبار التي لم يرها أثناء التدريب.
- تم ذلك باستخدام الدالة:

$$y_{pred} = model.predict(X\_test)$$

4): تقييم النموذج (Evaluate Your Model):

استخدمنا مؤشّرين رئيسيين لتقييم أداء النموذج:

- R^2 (معامل التحديد):

يشير إلى نسبة التباين في الأرباح التي يمكن للنموذج تفسيرها.

في حالتنا كانت القيمة $R^2 \approx 0.899$، أي أن النموذج يفسّر حوالي 89.9٪ من التغيّرات في الأرباح.

- RMSE (جذر متوسط مربعات الخطأ):

يقيس مقدار الانحراف بين القيم الحقيقية والمتوقعة.

في حالتنا كانت RMSE ≈ 9,056 دولار، وهي قيمة منخفضة نسبياً تدل على دقة جيدة.

- كما قمنا بتحليل إحصائي أكثر تفصيلاً باستخدام مكتبة statsmodels، ووجدنا أن إنفاق البحث والتطوير (R&D Spend) هو المتغيّر الأكثر تأثيراً والأكثر دلالة إحصائية ($p\text{-value} < 0.001$)، في حين أن بقية المتغيرات لم تكن مؤثرة بشكل واضح.
