MLT تقانات تعلم الآلة

S25

MLT

أستاذ المقرر

د. عصام سلمان

إعداد الطالب	الرقم الجامعي	الصف
مها ضعون	336128	C2
أحمد علي	328914	C3
هبة علي	332195	C3
رزان علي	336063	C3
محمد صايمه	330903	C2

رابط المشروع

https://github.com/mahadaoon6-crypto/MCS_MLT_HW1_S25_maha_336128_C2_mohamad_330903_C3_razan_336063_ahmad_328914_hiba_332195

مشروع بناء نظام كامل للحصول على أفضل وأقصر رحلة للعميل

Customer Journey

مقدمة:

يمرّ أي نشاط مبيعات بسلسلة من التفاعلات بين الشركة والعميل المحتمل، تبدأ من أول اتصال—سواء كان مكالمة هاتفية، بريدًا إلكترونيًا، أو تواصلًا مباشرًا—وتمتد عبر مجموعة من الخطوات التي تهدف إلى تحويل هذا التواصل الأولي إلى فرصة مبيعات حقيقية. تُعرَف الشركات أو الجهات التي يتم التواصل معها باسم **الحسابات(Accounts)** ، وتُسجَّل كل خطوة أو تفاعل معها ضمن ما يُعرف بـ **الأنشطة.(Activities)**

ومع تراكم هذه الأنشطة، تتشكل **فرصة مبيعات (Opportunity)** يمكن أن تنتهي إمّا بإغلاق ناجح وتحقيق صفقة، أو بالفشل وخسارة العميل. التحدي الأساسي لقسم المبيعات هو فهم:

- ما الأنشطة الأكثر تأثيرًا على نجاح الفرص؟
- وما المسارات (Paths) التي تزيد احتمال إغلاق الصفقة؟
- وكيف يمكن التنبؤ مبكرًا بفرص النجاح أو الفشل بناءً على الأنشطة المنفّذة؟

يهدف هذا المشروع إلى **بناء نموذج تنبؤي** قادر على تقييم فرص المبيعات بشكل مستمر، بحيث يتم تحديث تقدير احتمال النجاح مع كل نشاط جديد يتم تسجيله. إضافة إلى ذلك، يسعى المشروع إلى تحديد **أفضل خمس أنشطة** يمكن لموظف المبيعات تنفيذها لتعزيز فرص النجاح، وكذلك **أفضل خمسة مسارات** يمكن اتباعها للوصول إلى نتيجة إيجابية.

بهذا النهج، يوفر المشروع أداة تحليلية تساعد الشركات على تحسين استراتيجيات المبيعات، رفع كفاءة الفريق، واتخاذ قرارات مبنية على البيانات بدلًا من الحدس.

المتطلبات التقنية الأساسية للمشروع:

1. Jupyter Notebook:

دفتر Jupyter هو تطبيق ويب مفتوح المصدر يُستخدم لإنشاء ومشاركة المستندات الحاسوبية التفاعلية. يُعد أداة أساسية لعلماء البيانات، حيث يسمح بتشغيل التعليمات البرمجية، وعرض النتائج فورًا، وتعديلها بسهولة. يوفر JupyterLab واجهة محسنة وقائمة على الويب للعمل مع دفاتر Jupyter والبيانات، ويعتبر ترقية لواجهة Jupyter Notebook التقليدية. لتنزيله، عادةً ما يتم تثبيته عبر مدير حزم مثل Anaconda أو pip.

2. GitHub:

هو منصة استضافة سحابية لمشاريع تطوير البرمجيات، تعتمد على نظام التحكم في الإصدارات **Git**، وتتيح للمطورين العمل معًا بشكل تعاوني، وتتبع التغييرات في الكود، وإدارة المشاريع، ومشاركة الأكواد (سواء كانت مفتوحة المصدر أو خاصة)، وتوفر أدوات لتتبع الأخطاء وإدارة المهام، وتعتبر أكبر مجتمع للمطورين في العالم .

النقاط الرئيسية:

- **نظام تحكم في الإصدارات :**يستخدم Git لتسجيل تاريخ التعديلات على الكود، مما يسمح بالرجوع للإصدارات السابقة أو العمل على فروع منفصلة.
- **منصة تعاونية :**يجمع المطورين، ويتيح لهم مشاركة الأكواد ومراجعتها ودمجها بسهولة، حتى في المشاريع الكبيرة.
- **استضافة الأكواد :**يوفر مساحات لتخزين المستودعات (Repositories) وجعلها عامة أو خاصة، مع إمكانية الوصول إليها من أي مكان.
- **ميزات إضافية :**يتضمن أدوات لإدارة الأخطاء(Bug Tracking) ، وإدارة المهام(Task Management) ، و Wikiلكل مشروع.
- **مجتمع ضخم :**يضم ملايين المطورين حول العالم، وهو أساسي للمشاريع مفتوحة المصدر .

Gitو - GitHub ما الفرق؟

- **Git:**أداة سطر أوامر تعمل محليًا لتتبع التغييرات في الكود، وهو مجاني ومفتوح المصدر.
- **GitHub:**منصة على الإنترنت تعمل فوق Git لتوفير بيئة سحابية للمشاركة والتعاون وإدارة المستودعات بشكل مرئي وسهل .

باختصار، GitHubهو "المكان" الذي يتم فيه تخزين ومشاركة الكود الذي تتم إدارته بواسطة "الأداةGit "، لتمكين التطوير الجماعي بشكل فعال.

3. المكتبات الأساسية المستخدمة في المشروع

1. مكتبة pandas:

المكتبة الأساسية لمعالجة البيانات في المشروع، وهي مكتبة برمجية مفتوحة المصدر بلغة بايثون، تُستخدم بشكل أساسي لتحليل ومعالجة البيانات بكفاءة وسهولة، وتوفر هياكل بيانات قوية مثل DataFrame (إطار البيانات) للتعامل مع البيانات الجدولية والسلاسل الزمنية، وهي أداة لا غنى عنها في مجال علم البيانات والتحليل الإحصائي لمرونتها وقوتها وسرعتها. تُستخدم في المشروع في:

- قراءة ملفات CSV وExcel

- تنظيف البيانات
- إنشاء الجداول المشتقة مثل df_next .
- تجهيز صفوف التنبؤ للنموذج

2. مكتبة ydata_profiling (ProfileReport) :

وهي أداة قوية ومفتوحة المصدر تُستخدم لإجراء تحليل استكشافي سريع وشامل للبيانات (EDA)، حيث تقوم تلقائيًا بإنشاء تقارير تفاعلية مفصلة بصيغة HTML، تحتوي على ملخصات إحصائية، رسوم بيانية، تحليل لأنواع البيانات، اكتشاف مشاكل جودة البيانات (مثل القيم المفقودة أو المتطرفة)، وتفاعلات بين المتغيرات، كل ذلك في سطر واحد من الكود تقريبًا، مما يسرع فهم البيانات بشكل كبير .

تُستخدم في مرحلة الاستكشاف الأولي لفهم:

- جودة البيانات
- القيم المفقودة
- التوزيعات
- العلاقات بين المتغيرات

3. مكتبة numpy (Numerical Python):

مكتبة أساسية مفتوحة المصدر في بايثون للحوسبة العددية، توفر هياكل بيانات فعالة للمصفوفات (Arrays) متعددة الأبعاد، وتتيح مجموعة واسعة من الدوال الرياضية عالية الأداء للتعامل معها، مما يجعلها حجر الزاوية في مجالات علوم البيانات، التعلم الآلي، والحوسبة العلمية، وتدعم عمليات الجبر الخطي والإحصاء وغيرها بكفاءة عالية .

تُستخدم في المشروع في:

- إنشاء الأوزان الاحتمالية
- عمليات الاختيار العشوائي الموزون
- دعم النموذج أثناء التنبؤ

مكتبة Matplotlib (matplotlib.pyplot) :

مكتبة الرسم البياني الأساسية في بايثون، وهي مكتبة بايثون مفتوحة المصدر وقوية لإنشاء رسوم بيانية وتصورات بيانات احترافية، ثابتة ومتحركة وتفاعلية، تدعم أنواعاً عديدة مثل المخططات الخطية والشريطية والدائرية والمبعثرة، وتُستخدم في المشروع في إنشاء الرسوم التوضيحية أثناء تحليل البيانات. .

4. مكتبة scikit-learn (sklearn) :

مكتبة scikit-learn من مكتبات تعليم الالة بلغة البايثون، وتحتوي على العديد من الخوارزميات والطرق المستخدمة في مجال تعليم الالة مثل التصنيف Classification، العنقدة Clustering والانحدار Regression بالإضافة لاستخدامها في مرحلة تجهيز البيانات وتقييم النماذج.

وهي المكتبة الأساسية لبناء نموذج شجرة القرار.

5. مكتبة imblearn :

تُستخدم لمعالجة عدم توازن البيانات ، وفي بعض التجارب أثناء تنظيف البيانات و إعادة توزيع العينات.

6. مكتبة pyjanitor :

مكتبة تساعد في تبسيط عمليات تنظيف البيانات.تُستخدم عادة في:

- توحيد أسماء الأعمدة
- إزالة القيم المفقودة
- تحسين جودة البيانات

7. مكتبة joblib:

تُستخدم لحفظ وتحميل النماذج والكائنات الكبيرة بكفاءة عالية.

يُعتمد عليها في المشروع لحفظ النموذج الذكي وجميع مكوّناته داخل ملف واحد.

8. مكتبة json:

تُستخدم للتعامل مع الملفات بصيغةJSON . تُستخدم لقراءة أو حفظ ملفات مرجعية مثل ملف أسماء الأعمدة.

هذه المكتبات تُشكّل الأساس التقني للمشروع، بجميع مراحله:

- **تنظيف البيانات**
- **تحليلها واستكشافها**
- **بناء نموذج التنبؤ**
- **حفظ النموذج**
- **تشغيل الواجهة النهائية**

إعداد بيئة التشغيل:

1. تثبيت : Jupyter Notebook

دفتر Jupyter هو تطبيق ويب مفتوح المصدر يُستخدم لإنشاء ومشاركة المستندات الحاسوبية التفاعلية. يُعد أداة أساسية لعلماء البيانات، حيث يسمح بتشغيل التعليمات البرمجية، وعرض النتائج فورًا، وتعديلها بسهولة. يوفر JupyterLab واجهة محسنة وقائمة على الويب للعمل مع دفاتر Jupyter والبيانات، ويعتبر ترقية لواجهة Jupyter Notebook التقليدية. لتنزيله، عادةً ما يتم تثبيته عبر مدير حزم مثل Anaconda أو pip.

إعداد بيئة التشغيل:

1. تثبيت Jupyter Notebook :
2. تثبيت المكتبات الأساسية اللازمة
3. تحميل البيانات Dataset import ومعاينة ملف Excel:

يتم تحديد مسار ملف البيانات الرئيسي data_all.xlsx باستخدام مكتبة pathlib لضمان التعامل الصحيح مع المسارات داخل المشروع. بعد ذلك يتم التحقق من وجود الملف داخل مجلد المشروع، ثم قراءة محتواه باستخدام pandas.يشمل ذلك:

- استعراض أسماء الأوراق داخل ملف Excel
- قراءة الورقة الأولى باعتبارها تحتوي على البيانات الأساسية
- تحويل عمود التاريخ activity_date إلى نوع بيانات تاريخي
- عرض أول الصفوف للتأكد من سلامة القراءة
- عرض معلومات الأعمدة وأنواع البيانات

يتيح هذا الإجراء التأكد من أن البيانات تم تحميلها بشكل صحيح قبل البدء بعمليات التنظيف والتحليل اللاحقة.

```
Current working directory: C:\Users\acc\Customer_Journey_Project
File exists: True
Available sheets: ['data', 'Data_dictionary']
```

	account_id	SourceSystem	activity_date	who_id	opportunity_id	opportu
0	0010L00001hVmFhQAK	SFDC_US	2022-07-25	0030L00001vlbHLQAY	NaN	
1	0010L00001hVmFhQAK	SFDC_US	2023-02-08	0034X00002xZlQtQAK	NaN	
2	0010L00001hVmFhQAK	SFDC_US	2023-02-14	0030L00001vlbHLQAY	NaN	
3	0010L00001hVmFhQAK	SFDC_US	2023-02-20	0030L00001vlbHLQAY	NaN	
4	0010L00001hVmFhQAK	SFDC_US	2023-03-16	0034X00003GOUrFQAX	NaN	

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 259917 entries, 0 to 259916
Data columns (total 10 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   account_id         259917 non-null  object
 1   SourceSystem       259917 non-null  object
 2   activity_date      205516 non-null  datetime64[ns]
 3   who_id             238294 non-null  object
 4   opportunity_id     45849 non-null   object
 5   opportunity_stage  259901 non-null  object
 6   is_lead            259901 non-null  float64
 7   types              259901 non-null  object
 8   Country            259730 non-null  object
 9   solution           259901 non-null  object
dtypes: datetime64[ns](1), float64(1), object(8)
```

4. حفظ البيانات كملف CSV :

ملف CSV(**Comma-Separated Values**)، وهو صيغة بسيطة وخفيفة لتخزين البيانات على شكل جدول. يمتاز بأنه: ملف نصي، صغير الحجم، سهل القراءة من قبل الإنسان والبرامج، مدعوم من جميع لغات البرمجة وأدوات التحليل، وبما أن البيانات ستخضع لعمليات تنظيف متعددة وسيتم إنشاء نسخ مختلفة أثناء التطوير والنموذج سيحتاج بيانات خام وسريعة التحميل، لذلك تحويل البيانات إلى CSV يجعل سير العمل أكثراستقرارًا وسرعةوقابلية للتكرار وقابلية للمشاركة.

```
✅ تم حفظ الملف بنجاح: C:\Users\acc\Customer_Journey_Project\data_all.csv
 الحجم: 22836.44 KB
['account_id,SourceSystem,activity_date,who_id,opportunity_id,opportunity_stage,is_lead,types,Country,solution', '0010L00001hVmFhQAK,SFDC_US,2022-07-25,0030L00001vlbHLQAY,,no_opp,1.0,Follow Up,US,MRS', '0010L00001hVmFhQAK,SFDC_US,2023-02-08,0034X00002xZlQtQAK,,no_opp,1.0,Inbound Call,US,MRS']
```

المراحل الرئيسية للمشروع

أولاً: ✓ تنظيف البيانات (Clean the data):

1- فحص أولي للبيانات

1. عرض ملخص هيكلي للبيانات df.info() :

تم استخدام الدالة df.info() للحصول على ملخص هيكلي للبيانات، يشمل أنواع الأعمدة وعدد القيم غير المفقودة في كل عمود. يساعد هذا الفحص في تحديد الأعمدة التي تحتوي على نسب عالية من القيم المفقودة، والتأكد من أنواع البيانات قبل البدء بمرحلة التنظيف والمعالجة.

- يعطي صورة سريعة عن هيكل البيانات: عدد الصفوف، الأعمدة، الأنواع، والقيم المفقودة.
- هذا يساعد على معرفة الأعمدة التي تحتاج تحويل نوع أو معالجةNaN .

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 259917 entries, 0 to 259916
Data columns (total 10 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   account_id         259917 non-null  object
 1   SourceSystem       259917 non-null  object
 2   activity_date      205516 non-null  datetime64[ns]
 3   who_id             238294 non-null  object
 4   opportunity_id     45849 non-null   object
 5   opportunity_stage  259901 non-null  object
 6   is_lead            259901 non-null  float64
 7   types              259901 non-null  object
 8   Country            259730 non-null  object
 9   solution           259901 non-null  object
dtypes: datetime64[ns](1), float64(1), object(8)
memory usage: 19.8+ MB
```

2. الإحصاءات الوصفية للبيانات(Descriptive Statistics) :

تم استخدام الدالة df.describe () للحصول على الإحصاءات الوصفية للبيانات، بهدف فهم التوزيع العام للأعمدة، واكتشاف القيم المفقودة، وتحديد الأنماط الأولية قبل البدء بمرحلة التنظيف.

- تعطي ملخصًا إحصائيًا عن القيم نفسها: المتوسطات، التكرارات، القيم الشاذة المحتملة.
- ونبدأ برؤية التوزيع داخل الأعمدة النصية والتصنيفية والتاريخية.

نستخدم التعليمة : df.describe (include='all') للحصول على إحصاءات لجميع الأعمدة:

- الرقمية، النصية، الفئوية.

	account_id	SourceSystem	activity_date	who_id	opportunity_id	opportunity_stage	is_lead	types	Country	solution
count	259917	259917	205516	238294	45849	259901	259901.000000	259901	259730	259901
unique	13293	6	NaN	56849	8003	24	NaN	12	54	3
top	0012A00002SwdGoQAJ	SFDC_US	NaN	0030y00002JqdZyAAJ	0064X00001vC3zvQAC	no_opp	NaN	Email	US	MRS
freq	2503	128171	NaN	538	298	214052	NaN	105798	159831	185120
mean	NaN	NaN	2022-08-29 04:38:32.561552384	NaN	NaN	NaN	1.301165	NaN	NaN	NaN
min	NaN	NaN	1999-12-31 00:00:00	NaN	NaN	NaN	1.000000	NaN	NaN	NaN
25%	NaN	NaN	2021-07-27 00:00:00	NaN	NaN	NaN	1.000000	NaN	NaN	NaN
50%	NaN	NaN	2023-01-10 00:00:00	NaN	NaN	NaN	1.000000	NaN	NaN	NaN
75%	NaN	NaN	2024-06-20 00:00:00	NaN	NaN	NaN	2.000000	NaN	NaN	NaN
max	NaN	NaN	2026-12-31 00:00:00	NaN	NaN	NaN	2.000000	NaN	NaN	NaN
std	NaN	NaN	NaN	NaN	NaN	NaN	0.458765	NaN	NaN	NaN

- **تقييم مختصر لنتائج الفحص الأولي للبيانات(Initial Data Exploration)**

أظهر الفحص الأولي للبيانات عدّة ملاحظات مهمة تتعلق بجودة البيانات وتوزيعها، ويمكن تلخيصها كما يلي:

- البيانات كبيرة وغنية، لكنها تحتوي على نسب مفقودات مرتفعة في بعض الأعمدة الأساسية.
- هناك عدم توازن واضح في عمود **opportunity_stage**.
- بعض الأعمدة تحتاج تنظيفًا وترميزًا قبل إدخالها في النموذج.
- عمود التاريخ يتطلب معالجة وتحويل إلى ميزات مشتقة.
- الأعمدة المعرفية (IDs) غير مناسبة للنمذجة كما هي.

2- تنظيف البيانات الأساسية (مرحلة مبكرة):

1. حفظ نسخة أصلية من البيانات وإنشاء نسخة قابلة للتعديل:

قبل البدء بعمليات تنظيف البيانات ومعالجتها، تم إنشاء نسخة أصلية من البيانات باستخدام df.copy(deep=True) لضمان الاحتفاظ بالبيانات الخام دون أي تعديل. كما تم حفظ هذه النسخة خارجيًا بصيغة CSV تحت اسم data_original.csv بهدف الرجوع إليها عند الحاجة.

بعد ذلك، تم إنشاء نسخة ثانية (df_clean) مخصّصة لعمليات التنظيف والتحويل. وتم التحقق من تطابق النسختين باستخدام df.equals() للتأكد من أن نسخة المعالجة تبدأ من نفس البيانات الأصلية تمامًا. هذه الخطوة ضرورية لضمان العمل بشكل آمن ومنهجي، ولتسهيل تتبع التعديلات خلال مراحل المشروع.

```
تم حفظ النسخة الأصلية بنجاح: df_original.shape = (259917, 10)
المسار الحالي: C:\Users\acc\Customer_Journey_Project
معاينة سريعة للنسخة الأصلية (head):
          account_id SourceSystem activity_date             who_id  \
0  0010L00001hVmFhQAK      SFDC_US    2022-07-25  0030L00001vlbHLQAY
1  0010L00001hVmFhQAK      SFDC_US    2023-02-08  0034X00002xZlQtQAK
2  0010L00001hVmFhQAK      SFDC_US    2023-02-14  0030L00001vlbHLQAY
3  0010L00001hVmFhQAK      SFDC_US    2023-02-20  0030L00001vlbHLQAY
4  0010L00001hVmFhQAK      SFDC_US    2023-03-16  0034X00003GOUrFQAX

  opportunity_id opportunity_stage  is_lead         types Country solution
0            NaN            no_opp      1.0     Follow Up      US      MRS
1            NaN            no_opp      1.0  Inbound Call      US      MRS
2            NaN            no_opp      1.0  Inbound Call      US      MRS
3            NaN            no_opp      2.0  Inbound Call      US      MRS
4            NaN            no_opp      1.0  Inbound Call      US      MRS
تم حفظ نسخة خارجية باسم data_original.csv
تم إنشاء df_clean للمعالجة: df_clean.shape = (259917, 10)
معاينة سريعة لـ df_clean (head):
          account_id SourceSystem activity_date             who_id  \
0  0010L00001hVmFhQAK      SFDC_US    2022-07-25  0030L00001vlbHLQAY
1  0010L00001hVmFhQAK      SFDC_US    2023-02-08  0034X00002xZlQtQAK
2  0010L00001hVmFhQAK      SFDC_US    2023-02-14  0030L00001vlbHLQAY
3  0010L00001hVmFhQAK      SFDC_US    2023-02-20  0030L00001vlbHLQAY
4  0010L00001hVmFhQAK      SFDC_US    2023-03-16  0034X00003GOUrFQAX

  opportunity_id opportunity_stage  is_lead         types Country solution
0            NaN            no_opp      1.0     Follow Up      US      MRS
1            NaN            no_opp      1.0  Inbound Call      US      MRS
2            NaN            no_opp      1.0  Inbound Call      US      MRS
3            NaN            no_opp      2.0  Inbound Call      US      MRS
4            NaN            no_opp      1.0  Inbound Call      US      MRS
هل df_original و df_clean متطابقتان تماماً؟ نعم
```

2. تنظيف أسماء الأعمدة:

تم استخدام مكتبة **pyjanitor** لتنظيف أسماء الأعمدة بشكل تلقائي عبر الدالة clean_names() . تقوم هذه الدالة بتوحيد صيغة أسماء الأعمدة من خلال تحويلها إلى أحرف صغيرة، واستبدال المسافات والرموز غير المناسبة بشرطة سفلية، مما يسهل التعامل معها خلال مراحل التحليل والتنظيف اللاحقة.

بعد عملية التنظيف، تم حفظ ملف مرجعي (column_names_reference.json) يحتوي على أسماء الأعمدة قبل وبعد التعديل، وذلك بهدف التوثيق وضمان إمكانية الرجوع للأسماء الأصلية عند الحاجة. كما تم التأكد من أن عدد الصفوف لم يتغير بعد عملية التنظيف، مما يؤكد أن التعديل اقتصر فقط على أسماء الأعمدة دون المساس بالبيانات.

```
أسماء الأعمدة بعد التنظيف (معاينة):
Index(['account_id', 'sourcesystem', 'activity_date', 'who_id',
       'opportunity_id', 'opportunity_stage', 'is_lead', 'types', 'country',
       'solution'],
      dtype='object')
✓ تم تنظيف أسماء الأعمدة وحفظ المرجع في column_names_reference.json
عدد الصفوف الأصلي: 259917
بعد جميع خطوات التنظيف: 259917
```

3. إزالة التكرارات من البيانات:

تم في هذه الخطوة استخدام الدالة drop_duplicates() لإزالة الصفوف المكررة داخل مجموعة البيانات. تساعد هذه العملية في تحسين جودة البيانات ومنع تكرار الأنشطة أو السجلات التي قد تؤثر على التحليل أو على أداء النموذج التنبؤي لاحقًا.

قبل الإزالة، تم تسجيل عدد الصفوف الأصلية، ثم حساب عدد الصفوف بعد الإزالة لمعرفة عدد السجلات المكررة التي تم حذفها. كما تمت معاينة عينة عشوائية من البيانات بعد التنظيف للتأكد من سلامة السجلات المتبقية.

هذه الخطوة ضرورية لضمان أن النموذج سيُدرَّب على بيانات نظيفة وغير متحيزة بسبب التكرار.

```
=== إزالة التكرارات من البيانات ===
📊 عدد الصفوف قبل الإزالة: 259917
✓ تم إزالة 45460 صف مكرر

=== معاينة سريعة بعد إزالة التكرارات (5 صفوف عشوائية) ===
                account_id sourcesystem activity_date             who_id opportunity_id opport
unity_stage  is_lead           types country solution
97408   0016g00000P5B9ZAAV  SFDC_GLOBAL    2020-06-10                NaN            NaN
no_opp       1.0  2nd Appointment      UK       MRS
237958  001b000003HAVIuAAP     SFDC_ROW    2025-03-17  0030X00002whWByQAM            NaN
no_opp       1.0           Email      FR       MRS
126968  001E000000ITdN6IAL      SFDC_US    2023-03-08  0034X00003GLqqHQAT            NaN
no_opp       1.0     Inbound Call      US  Digital
177599  001E0000017dB3nIAE      SFDC_US    2025-01-08  0030y00002RtcIRAAZ            NaN
no_opp       1.0           Email      US       MRS
236852  001b000003HATL0AAP     SFDC_ROW    2022-01-09  0036700003e4FMTAA2            NaN
no_opp       1.0           Email      FR       MRS
259917 :عدد الصفوف الأصلي
214457 :بعد جميع خطوات التنظيف
```

```
=== إزالة التكرارات من البيانات ===
📊 عدد الصفوف قبل الإزالة: 259917
✅ تم إزالة 45460 صف مكرر

=== معاينة سريعة بعد إزالة التكرارات (5 صفوف عشوائية) ===
                account_id sourcesystem activity_date             who_id      opportunity_
id opportunity_stage  is_lead         types country solution
74936   0014X00002J4UqRQAV  SFDC_GLOBAL    2022-02-02  0034X00002sdwOLQAY                N
aN           no_opp       1.0         Email      DE      MRS
208075  001Vw00000AoJsxIAF      SFDC_US    2024-11-12                NaN  006Vw000008km1pI
AA              Won       1.0       Meeting      US      MRS
107834  001E000000BS8MkIAL      SFDC_US    2022-11-09  0030L00001uGCIFQA4                N
aN           no_opp       2.0  Inbound Call      US      MRS
126476  001E000000ITcdVIAT      SFDC_US    2023-04-09  0030y00002RVMuuAAH  0064X00001y0d6WQ
AQ              Lost       1.0     Follow Up      US      MRS
113423  001E000000Cp9P1IAJ      SFDC_US    2024-10-08  0034X00003T1ZpAQAV                N
aN           no_opp       1.0  Inbound Call      US      MRS
259917 :عدد الصفوف الأصلي
214457 :بعد جميع خطوات التنظيف
```

4. معالجة القيم المفقودة:

- **تحليل القيم المفقودة**

تم في هذه الخطوة حساب عدد القيم المفقودة ونسبتها في كل عمود باستخدام دوال isnull() و mean()من مكتبة pandas. يهدف هذا التحليل إلى تحديد الأعمدة التي تحتوي على نسب عالية من القيم الفارغة، مما يساعد في اتخاذ قرارات مناسبة خلال مرحلة تنظيف البيانات، مثل الإزالة أو التعويض أو إعادة الترميز.

تم عرض النتائج في جدول مرتب تنازليًا حسب النسبة المئوية للقيم المفقودة، مما يوفر رؤية واضحة للأعمدة الأكثر تأثرًا. كما تم توثيق عدد الصفوف قبل وبعد عمليات التنظيف لضمان عدم حدوث تغييرات غير مقصودة في حجم البيانات.

```
=== نسب القيم الفارغة في كل عمود ===
                  count_nan  percent_nan
opportunity_id       170070        79.30
who_id                20992         9.79
activity_date         12171         5.68
country                 184         0.09
opportunity_stage        14         0.01
is_lead                  14         0.01
types                    14         0.01
solution                 14         0.01
sourcesystem              0         0.00
account_id                0         0.00
259917 :عدد الصفوف الأصلي
214457 :بعد جميع خطوات التنظيف
```

- **تحديد الأعمدة المرشحة للحذف بناءً على نسبة القيم المفقودة**

اعتمادًا على تحليل القيم المفقودة، تم تحديد الأعمدة التي تتجاوز نسبة 75% من القيم الفارغة، وذلك باستخدام جدول na_summary الذي تم إنشاؤه مسبقًا. الأعمدة ذات النسب العالية من القيم المفقودة غالبًا ما تكون غير مفيدة في التحليل أو النمذجة، وقد تؤثر سلبًا على جودة النموذج.

بعد تحديد هذه الأعمدة، تم حذفها من نسخة التنظيف df_clean، مع توثيق عدد الأعمدة المحذوفة والشكل الجديد للبيانات. هذه الخطوة تساعد في تقليل الضوضاء داخل البيانات، وتسهّل عمليات التنظيف والتحليل اللاحقة دون التأثير على عدد الصفوف.

```
الأعمدة المرشحة للحذف (أكثر من 75% قيم فارغة):
['opportunity_id']
✓ تم حذف 1 عمود. الشكل الجديد: (214457, 9)
عدد الصفوف الأصلي: 259917
بعد جميع خطوات التنظيف: 214457
```

- **معالجة القيم الفارغة في الأعمدة الأساسية**

تم في هذه المرحلة التعامل مع القيم المفقودة بطريقة منهجية تعتمد على أهمية كل عمود.

حيث تم إسقاط الصفوف التي تفتقد قيمًا في الأعمدة الحرجة مثل account_id, types, country, وsolution، نظرًا لأن هذه الأعمدة ضرورية لتعريف السجل ولا يمكن تعويضها بقيم افتراضية.

أما الأعمدة غير الحرجة، فقد تمت معالجتها بطرق مناسبة لطبيعة كل عمود:

- تعويض القيم المفقودة في who_id بقيمة "unknown"
- تعويض opportunity_stage بالقيمة الأكثر تكراراً "no_opp"
- تعويض القيم المفقودة في is_lead بالوسيط (median) للحفاظ على توزيع البيانات

تضمن هذه الخطوات الحفاظ على جودة البيانات وتقليل التشويش قبل الانتقال إلى مراحل التحليل والنمذجة.

```
=== معالجة القيم الفارغة في الأعمدة الأساسية ===
📉 تم إسقاط 184 صفوف بسبب فقدان قيم في الأعمدة الحرجة
✓ تم معالجة القيم الفارغة في الأعمدة الأساسية
🔎 معاينة سريعة:
              account_id sourcesystem activity_date             who_id  \
7495    0010L00001jaUBkQAM       SFDC_US    2023-11-13  0030y00002UxIn2AAF
251512  001b000003Ha1w7AAB      SFDC_ROW    2016-07-29             unknown
147352  001E000000RGfntIAD       SFDC_US    2013-02-04             unknown
215340  001b000003HA6ThAAL      SFDC_ROW    2020-07-28  003b000001hVVAHAA4
74785   0014X00002J4UI0QAN   SFDC_GLOBAL    2021-09-26  0034X00002zr7qdQAA

       opportunity_stage  is_lead      types country solution
```

```
7495             no_opp      1.0      Email      US      MRS
251512           no_opp      1.0  Follow Up      FR  Digital
147352           no_opp      1.0    Meeting      US      MRS
215340           no_opp      1.0      Email      FR      MRS
74785            no_opp      1.0      Email      DE      MRS
259917 :عدد الصفوف الأصلي
214273 :بعد جميع خطوات التنظيف
```

5. توحيد القيم النصية في الأعمدة الأساسية

تم في هذه المرحلة توحيد القيم النصية داخل الأعمدة الفئوية بهدف تحسين جودة البيانات وتقليل التباين غير الضروري.

بدأت العملية بتحويل جميع النصوص إلى أحرف صغيرة وإزالة الفراغات الزائدة لضمان اتساق القيم ومنع التكرارات الناتجة عن اختلاف الشكل فقط.

بعد ذلك، تم دمج القيم المتشابهة داخل الأعمدة الأساسية مثل opportunity_stage و types من خلال خرائط تحويل مخصّصة.

على سبيل المثال، تم دمج مراحل مثل negotiation و discovery و prospecting ضمن فئة موحدة هي "ongoing"، كما تم توحيد أنواع الأنشطة مثل outbound call و inbound call ضمن فئة "call".

يساعد هذا التوحيد في تقليل عدد الفئات(Cardinality Reduction) ، وتحسين قابلية البيانات للتحليل والنمذجة، وضمان اتساق القيم عبر السجلات المختلفة.

```
=== توحيد القيم النصية في الأعمدة الأساسية ===
✓ تم توحيد القيم النصية
 :معاينة سريعة بعد التوحيد
              account_id sourcesystem activity_date             who_id opportunity_stage  is
_lead  types      country solution
44298   0012A00002Swai4QAB     sfdc_cxm           NaT  0033k00003UrF72AAF               won
2.0  email  Switzerland  digital
226893  001b000003HAFwtAAH     sfdc_row    2025-01-12  0030X00002gRJe3QAG            no_opp
1.0   call           FR      mrs
259248  001b000003i3IkvAAE     sfdc_row    2022-07-07  0030X00002wLjWJQA0            no_opp
1.0  email           FR      mrs
131241  001E000000ITk3DIAT      sfdc_us    2022-04-05  0030y00002RVCM7AAP            no_opp
1.0   call           US  digital
223662  001b000003HADTaAAP     sfdc_row    2022-06-29  0030X00002Q1DdvQAF            no_opp
1.0  email           FR      mrs
259917 :عدد الصفوف الأصلي
214273 :بعد جميع خطوات التنظيف
```

6. عرض تقرير البيانات بعد التنظيف الأساسي المبكر باستخدام ProfileReport :

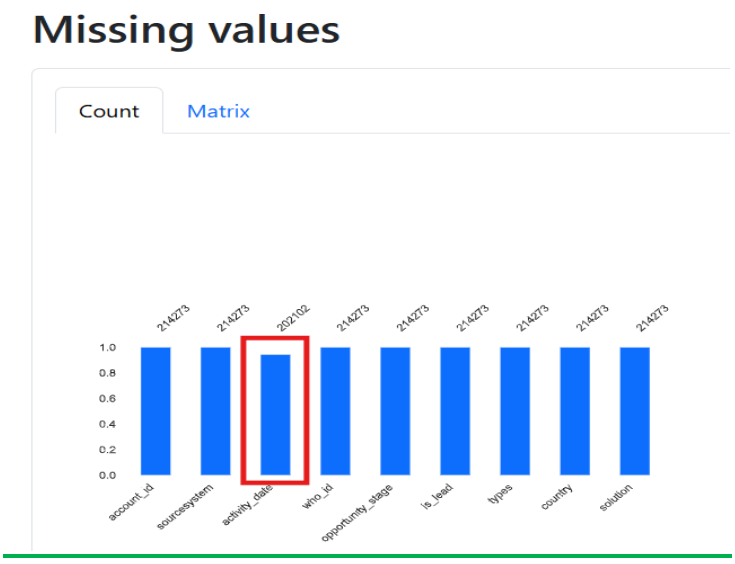
بعد الانتهاء من مرحلة التنظيف الأساسي، والتي شملت إزالة التكرارات، ومعالجة القيم المفقودة، وتوحيد القيم النصية، وحذف الأعمدة ذات المفقودات العالية، تم إنشاء تقرير شامل باستخدام مكتبة ydata_profiling بهدف تقييم جودة البيانات بعد هذه المرحلة.

يعرض التقرير توزيع القيم، ونسب المفقودات المتبقية، وأنواع الأعمدة، والقيم الشاذة، بالإضافة إلى العلاقات بين المتغيرات.

يساعد هذا التقرير في التأكد من نجاح عمليات التنظيف الأساسية، وتحديد ما إذا كانت هناك خطوات إضافية مطلوبة قبل الانتقال إلى مرحلة التحليل المتقدم أو النمذجة.

قدّم تقرير ProfileReport تحليلًا شاملًا لحالة البيانات بعد الانتهاء من مرحلة التنظيف الأساسي، وأظهر أن البيانات أصبحت أكثر استقرارًا وانخفاضًا في القيم المفقودة، حيث اقتصرت المفقودات على عمود التاريخ فقط. كما بيّن التقرير أن الأعمدة الفئوية أصبحت موحّدة القيم بعد المعالجة، مع وجود عدم توازن واضح في بعض الأعمدة مثل opportunity_stage و types و solution، حيث تهيمن فئات محددة مثل no_opp و email و mrs على معظم السجلات. هذا النوع من التوزيع يُعد طبيعيًا في بيانات الأنشطة لكنه مهم عند الانتقال إلى مرحلة النمذجة. كذلك أظهر التقرير عدم وجود قيم شاذة أو تواريخ غير صالحة، إضافة إلى انخفاض عدد الصفوف المكررة. بشكل عام، يعكس التقرير أن البيانات أصبحت جاهزة للانتقال إلى مرحلة التنظيف المتقدم وهندسة الميزات.

تم عرض التقرير داخل الـ Notebook ، كما تم حفظ نسخة HTML منه لتوثيق النتائج والرجوع إليها لاحقًا.



3- مرحلة التنظيف البيانات المتقدم:

1. تنظيف وتوحيد عمود البلد (country):

في مرحلة التنظيف المتقدم، تم تنفيذ عملية معيارية لعمود country بهدف توحيد أسماء الدول وتحسين جودة البيانات الجغرافية.

بدأت العملية بتنظيف النصوص وإزالة الفراغات، ثم استخدام مكتبة pycountry لمطابقة القيم مع قاعدة بيانات رسمية للدول.

تم إنشاء عمود جديد باسم يحتوي على الاسم الرسمي للدولة عند توفر تطابق، بينما تم الاحتفاظ بالقيم غير المطابقة دون حذف لضمان عدم فقدان البيانات.

يساعد هذا التوحيد في تقليل التباين غير الضروري في أسماء الدول، وتحسين دقة التحليل الجغرافي، وضمان جاهزية العمود لعمليات الترميز والنمذجة لاحقًا.

```
أسماء الدول المحصل عليها في country_standard (أو القيم الأصلية في غير المطابقة):
country_standard
United States         132211
France                 53957
UK                     13866
Germany                 6724
Canada                  1528
Switzerland              952
Belgium                  631
Netherlands              585
Italy                    569
Ireland                  418
Australia                326
Brazil                   323
Austria                  320
Czechia                  176
Guadeloupe               168
Spain                    132
Singapore                120
China                    120
[NULL]                   114
Korea, Republic of        96
Name: count, dtype: int64
```

- **حذف الصفوف التي تحمل "[NULL]" في عمودcountry :**

خلال مرحلة التنظيف المتقدم، تم التعامل مع القيمة "[NULL]" الموجودة في عمود country، والتي تُمثل قيمة غير صالحة ولا تشير إلى دولة حقيقية. وبما أنها لا يمكن توحيدها أو مطابقتها باستخدام قواعد بيانات الدول الرسمية، فقد تم حذف الصفوف التي تحتوي على هذه القيمة لضمان جودة البيانات الجغرافية.

يساعد هذا الإجراء في إزالة الضوضاء من البيانات، ويمنع إدخال قيم غير منطقية في عمليات الترميز أو النمذجة اللاحقة، كما يضمن أن عمود الدولة يحتوي فقط على قيم قابلة للاستخدام والتحليل.

بعد الحذف، تم طباعة عدد الصفوف المحذوفة والمتبقية للتأكد من تأثير العملية على حجم البيانات.

```
تم حذف 114 صفًا يحتوي على '[NULL]' في عمود country.
عدد الصفوف المتبقية بعد الحذف: 214159
```

2. التنظيف المتقدم لعمود التاريخ (activity_date):

في مرحلة التنظيف المتقدم، تم تطبيق خوارزمية مخصّصة لمعالجة عمود activity_date بهدف تحويل القيم النصية غير القياسية إلى تواريخ صالحة قدر الإمكان.

اعتمدت العملية على تنظيف السلاسل النصية من الرموز غير الضرورية، ثم تجربة مجموعة واسعة من صيغ التواريخ الشائعة، بما في ذلك الصيغ القياسية وصيغةISO . وفي حال تعذّر التحويل، يتم إرجاع القيمة كـ NaT لضمان الاتساق.

أظهر التقرير النهائي أن نسبة كبيرة من القيم تم تحويلها بنجاح إلى تواريخ صالحة، بينما تم تحديد القيم غير القابلة للمعالجة بشكل واضح.

يسهم هذا الإجراء في تحسين جودة البيانات الزمنية، ويضمن جاهزية العمود لعمليات التحليل الزمني وهندسة الميزات لاحقًا.

```
Total (df_clean): 214159
Valid (parsed date): 201988
Invalid (NaT): 12171
النسبة المئوية للصفوف الصالحة: 94.32%
```

- **حذف الصفوف ذات التواريخ غير الصالحة في عمود activity_date:**

بعد تطبيق خوارزمية التنظيف المتقدم على عمود activity_date، تبيّن أن جزءًا من القيم لا يمكن تحويله إلى تاريخ صالح حتى بعد محاولات الإصلاح المتعددة، وتم تصنيف هذه القيم كـ NaT.

ونظرًا لأهمية هذا العمود في التحليل الزمني وهندسة الميزات، ولأن القيم غير الصالحة تمثل نسبة صغيرة من البيانات، فقد تم حذف الصفوف التي تحتوي على تواريخ غير قابلة للاستخدام.

تضمن هذه الخطوة الحفاظ على جودة البيانات الزمنية، وتمنع إدخال ضوضاء أو قيم ناقصة في النماذج اللاحقة، مما يرفع من موثوقية التحليل ودقة النتائج.

```
تم حذف 12171 صفًا يحتوي على قيم تاريخ غير صالحة (NaT).
عدد الصفوف المتبقية بعد الحذف: 201988
```

3. الكشف عن القيم الشاذة في عمود is_lead ومعالجتها:

تم تطبيق منهجية Interquartile Range (IQR)للكشف عن القيم الشاذة في عمود is_lead، وذلك من خلال حساب الربع الأول (Q1) والربع الثالث (Q3) وتحديد الحدود المقبولة للقيم.

أظهر التحليل وجود قيم تقع خارج النطاق الطبيعي، وهي قيم غير منطقية بالنسبة لعمود يفترض أن يحتوي على فئات محدودة.

ولمعالجة هذه الحالات، تم استبدال القيم الشاذة بالوسيط، وهو خيار مناسب يحافظ على التوزيع الأصلي للعمود دون إدخال تحيز.

بعد المعالجة، أصبح العمود نظيفًا وخاليًا من القيم غير المتوقعة، مما يضمن جاهزيته لعمليات الترميز والنمذجة اللاحقة.

```
=== الكشف عن القيم الشاذة ومعالجتها في العمود is_lead ===
Q1: 1.0, Q3: 2.0, IQR: 1.0
Lower bound: -0.5, Upper bound: 3.5
عدد القيم الشاذة في is_lead: 0
✓ تم استبدال القيم الشاذة في is_lead بالوسيط: 1.0
📊 توزيع القيم بعد المعالجة:
is_lead
1.0    145965
2.0     56023
Name: count, dtype: int64
```

4. تنظيف وتوحيد عمودtypes :

في مرحلة التنظيف المتقدم، تم إجراء سلسلة من الخطوات لمعالجة عمود types بهدف تحسين جودة البيانات الفئوية وتقليل التشتت في القيم.

بدأت العملية بتوحيد صيغة النصوص عبر تحويلها إلى حروف صغيرة وإزالة الفراغات، ثم استبدال القيم غير الصالحة بقيم مفقودة حقيقية.

بعد ذلك، تم ملء القيم المفقودة باستخدام الفئة الأكثر تكرارًا لضمان الحفاظ على التوزيع الأصلي. كما تم توحيد القيم المتشابهة ودمج الفئات النادرة التي تقل نسبتها عن 1% ضمن فئة عامة باسم "other".

أسهمت هذه الخطوات في تقليل عدد الفئات، وتحسين اتساق العمود، وتجهيزه لعمليات الترميز والنمذجة اللاحقة بشكل أكثر فعالية

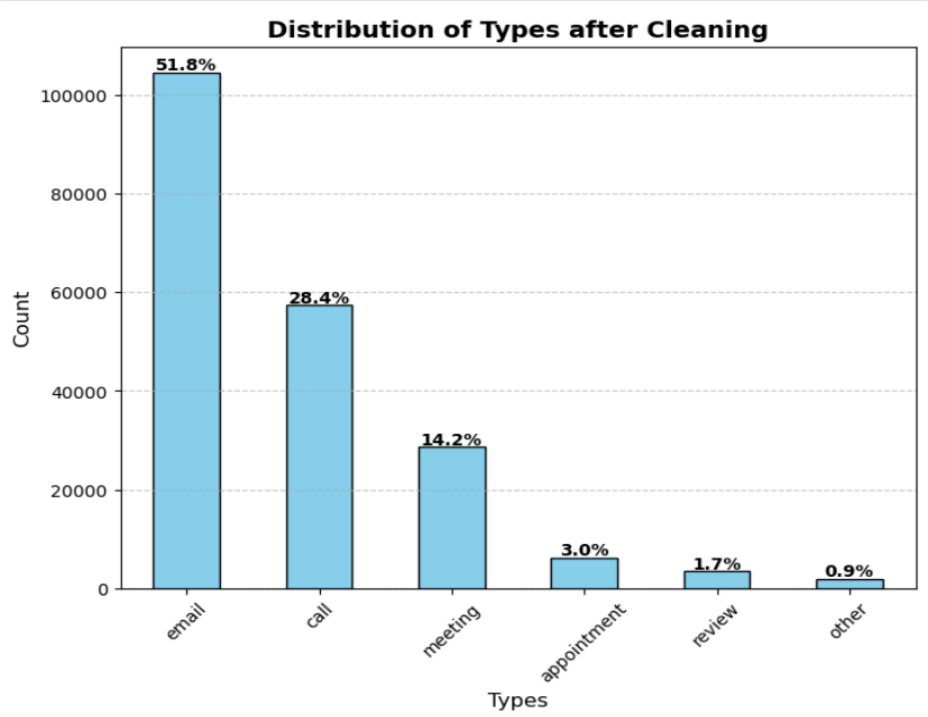
```
=== تنظيف عمود types ===
✓ القيمة الأكثر تكرارًا: email

التوزيع بعد التوحيد:
types
email          104562
call            57386
meeting         28618
appointment      6154
review           3397
demo             1773
discovery          98
Name: count, dtype: int64

التوزيع بعد دمج الفئات النادرة:
types
email          104562
call            57386
meeting         28618
appointment      6154
review           3397
other            1871
Name: count, dtype: int64

النسب المئوية لكل فئة:
types
email          51.77
call           28.41
meeting        14.17
appointment     3.05
review          1.68
other           0.93
Name: proportion, dtype: float64

قائمة الفئات الفريدة:
['appointment', 'call', 'email', 'meeting', 'other', 'review']
```

- **رسم مخطط شريطي لتوزيع فئات types بعد التنظيف:**

بعد الانتهاء من تنظيف وتوحيد عمود types، تم إنشاء مخطط شريطي يوضح التوزيع النهائي للفئات بعد دمج القيم المتشابهة وتجميع الفئات النادرة.

يعرض المخطط عدد كل فئة بالإضافة إلى نسبتها المئوية، مما يساعد على فهم مدى انتشار كل نوع من الأنشطة داخل البيانات.

يسهم هذا التمثيل البصري في تقييم جودة التنظيف، والتأكد من أن الفئات أصبحت أكثر اتساقًا وتوازنًا، كما يسهّل اتخاذ القرارات المتعلقة بالترميز والنمذجة لاحقًا.



5. التنظيف المتقدم للعمود solution:

تم إجراء تنظيف متقدم لعمود solution بهدف توحيد القيم الفئوية وضمان اتساقها قبل استخدامها في التحليل أو النمذجة.

بدأت العملية بتحويل النصوص إلى صيغة موحدة عبر إزالة الفراغات وتحويلها إلى حروف صغيرة، مما يمنع تكرار الفئات بسبب اختلاف الكتابة.

بعد ذلك، تم تطبيق خريطة تحويل (mapping) لتجميع القيم المختلفة ضمن ثلاث فئات رئيسية هي MRS, Digital, وPLS.

ونظرًا لأن هذا العمود كان مُعرّفًا كأحد الأعمدة الحرجة في مرحلة التنظيف المبكر، فقد تم حذف الصفوف التي تحتوي على قيم مفقودة فيه، مما يعني أن جميع القيم المتبقية كانت صالحة ولا تحتاج إلى تعبئة.

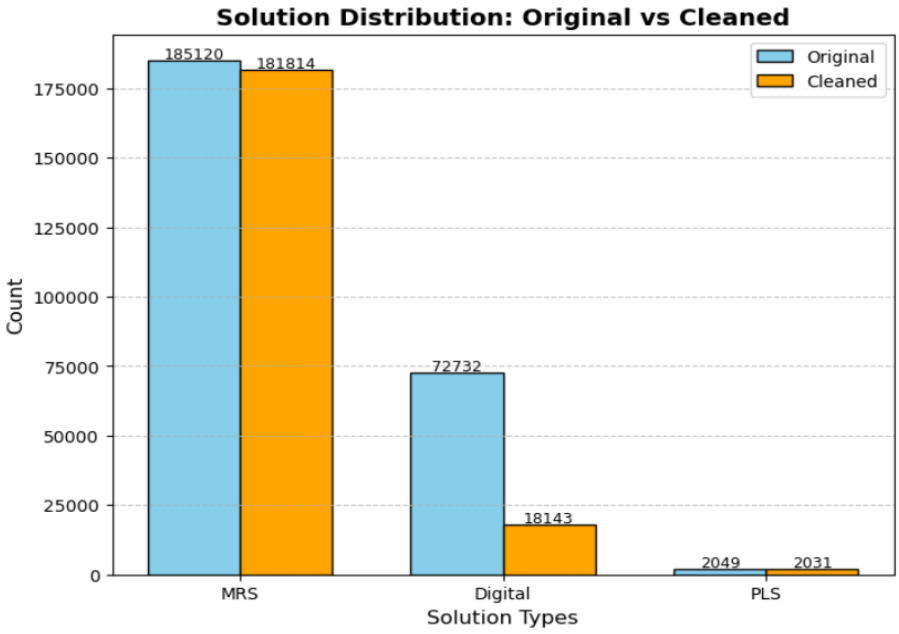
أسهمت هذه الخطوات في تقليل التشتت داخل العمود، وتوحيد الفئات، وضمان جاهزية البيانات لعمليات الترميز والنمذجة اللاحقة بشكل أكثر دقة واستقرار.

- **التمثيل البصري لتوزيع قيم solution قبل وبعد التنظيف:**

لتحليل تأثير عملية التنظيف على عمود solution، تم إنشاء مخطط أعمدة مزدوج يقارن بين التوزيع الأصلي للقيم والتوزيع بعد تطبيق خطوات التنظيف.

يعرض المخطط عدد كل فئة قبل التنظيف مقابل عددها بعده، مما يساعد على فهم كيفية توحيد القيم، وإزالة الصفوف غير الصالحة، وتقليل التشتت داخل العمود.

يسهم هذا التمثيل البصري في توضيح أثر التنظيف على جودة البيانات، ويؤكد أن الفئات النهائية أصبحت أكثر اتساقًا واستقرارًا، مما يجعل العمود جاهزًا للاستخدام في عمليات التحليل والنمذجة اللاحقة.



6. تنظيف متقدم للعمود sourcesystem :

تم إجراء تنظيف متقدم لعمود sourcesystem بهدف توحيد القيم الفئوية وضمان اتساقها.

بدأت العملية بتحويل النصوص إلى صيغة موحدة عبر إزالة الفراغات وتحويلها إلى حروف كبيرة، ثم تحديد قائمة بالقيم الصحيحة المعتمدة في النظام.

تم تصنيف أي قيمة غير متوقعة ضمن فئة عامة باسم "Other" لتقليل التشتت ومنع تضخم عدد الفئات.

وأخيرًا، تم تحويل العمود إلى نوع categoryلتحسين كفاءة التخزين والمعالجة.

أسهمت هذه الخطوات في تبسيط الفئات، وتحسين جودة البيانات، وتجهيز العمود لعمليات الترميز والنمذجة اللاحقة.

```
تقرير التنظيف:
- عدد الصفوف قبل التنظيف: 201988
- عدد الصفوف بعد التنظيف: 201988
- عدد الصفوف المحذوفة: 0

توزيع الفئات بعد التنظيف:
sourcesystem
SFDC_US           126013
SFDC_ROW           53447
SFDC_GLOBAL        22504
SFDC_BEANWORKS        14
Other                 10
Name: count, dtype: int64
```

7. تنظيف عمود opportunity_stage :

تم تنفيذ عملية تنظيف متقدمة لعمود *opportunity_stage* بهدف توحيد القيم الفئوية وتقليل التشتت داخل البيانات.

بدأت العملية بتحويل النصوص إلى صيغة موحدة عبر إزالة الفراغات وتحويلها إلى حروف صغيرة، ثم استبدال القيم غير الصالحة بقيم مفقودة حقيقية.

نظرًا لأهمية هذا العمود في التحليل، تم حذف الصفوف التي تحتوي على قيم مفقودة فيه لضمان جودة البيانات.

بعد ذلك، تم تحليل التوزيع النسبي للفئات وتحديد الفئات النادرة التي تقل نسبتها عن 1%، ثم دمجها ضمن فئة عامة باسم **"other"**لتقليل عدد الفئات وتحسين استقرار النماذج.

أخيرًا، تم تحويل العمود إلى نوع *category* لزيادة كفاءة التخزين والمعالجة.

```
تقرير التنظيف:
- عدد الصفوف قبل التنظيف: 201988
- عدد الصفوف بعد التنظيف: 201988
- عدد الصفوف المحذوفة: 0

توزيع الفئات بعد التنظيف:
opportunity_stage
no_opp     158851
won         18216
lost        14235
other        5813
ongoing      4873
Name: count, dtype: int64

النسب المئوية لكل فئة بعد التنظيف:
opportunity_stage
no_opp     78.64
won         9.02
lost        7.05
other       2.88
```

```
ongoing     2.41
Name: proportion, dtype: float64

📋 قائمة الفئات الفريدة في العمود opportunity_stage:
['lost', 'no_opp', 'ongoing', 'other', 'won']
```

أظهر التوزيع النهائي أن الفئة no_opp تمثل النسبة الأكبر من البيانات بنسبة 78.64%، تليها الفئات won (9.02%) و lost (7.05%)، بينما تشكل الفئتان other و ongoing نسبًا صغيرة نسبيًا (2.88% و 2.41%).

ويعكس هذا التوزيع الطبيعة الحقيقية للبيانات، حيث تمثل معظم السجلات فرصًا غير نشطة أو غير مؤهلة، بينما تشكل الحالات الرابحة والخاسرة جزءًا أصغر من البيانات.

أسهمت هذه الخطوات في إنتاج عمود نظيف، موحّد، وخالٍ من القيم غير الصالحة، مع تقليل عدد الفئات النادرة، مما يجعل العمود جاهزًا للاستخدام في عمليات التحليل والترميز والنمذجة اللاحقة.

8. تنظيف عمود is_lead :

تم تنظيف عمود is_lead من خلال تحويل جميع القيم إلى صيغة رقمية باستخدام دالة to_numeric، مما يسمح بالكشف عن أي قيم غير صالحة وتحيلها تلقائيًا إلى قيم مفقودة.

بعد ذلك، تم حذف الصفوف التي تحتوي على قيم غير قابلة للتحويل لضمان أن العمود يحتوي فقط على قيم رقمية صحيحة.

أسهمت هذه الخطوة في إزالة القيم الشاذة أو النصية، وضمان جاهزية العمود للاستخدام في التحليل الإحصائي أو النمذجة اللاحقة.

```
القيم الفريدة بعد التنظيف:
[1. 2.]
```

9. إنشاء تقرير البيانات بعد التنظيف المتقدم:

بعد الانتهاء من جميع خطوات التنظيف المتقدم، تم إنشاء تقرير شامل باستخدام مكتبة Pandas Profiling بهدف تقييم جودة البيانات بعد المعالجة.

يتضمن التقرير تحليلاً استكشافيًا موسعًا يشمل توزيع المتغيرات، القيم المفقودة، الارتباطات، القيم الشاذة، الأنماط الإحصائية، بالإضافة إلى ملخص شامل لكل عمود.

تم عرض التقرير داخل بيئةJupyter Notebook ، كما تم حفظ نسخة HTML تحمل تاريخ ووقت الإنشاء لضمان إمكانية الرجوع إليها لاحقًا أو مشاركتها ضمن التقرير النهائي.

- **ملخص مضمون تقرير البيانات بعد التنظيف المتقدم (Overview Analysis):**

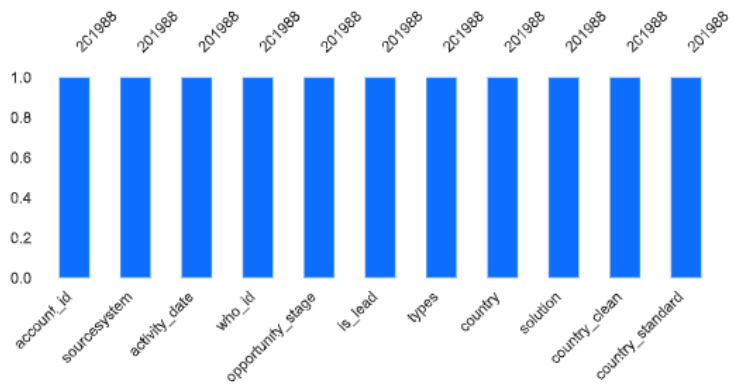
ملخص جودة البيانات بعد التنظيف والتحديات المتبقية

بعد تنفيذ سلسلة من خطوات التنظيف الشاملة—بما في ذلك توحيد أسماء الأعمدة، معالجة القيم المفقودة، إزالة الصفوف غير الصالحة، وتوحيد أسماء الدول—أظهر تقرير البيانات الناتج تحسنًا كبيرًا في جودة البيانات. فقد أصبحت جميع الأعمدة خالية من القيم المفقودة بنسبة 0%، وتم تقليل التكرار إلى حدٍّ محدود، كما أصبحت البنية العامة للبيانات أكثر اتساقًا واستعدادًا للنمذجة. ومع ذلك، كشف التقرير عن عدد من التحديات التي بقيت بعد التنظيف، أبرزها عدم التوازن الواضح في بعض الفئات مثل:

- عمود opportunity_stage الذي تهيمن عليه فئة no_opp
- عمود types الذي يتركز بشكل كبير حول email
- عمود country_standard المنحاز للسوق الأمريكي
- عمود solution الذي تهيمن عليه فئة MRS

كما أظهر التقرير نطاقًا زمنيًا واسعًا لعمود التاريخ، ووجود أعمدة معرفية ذات تباين عالٍ لا تصلح للاستخدام المباشر في النمذجة.

ولمعالجة هذه التحديات، تم لاحقًا تنفيذ مجموعة من الإجراءات، شملت: ترميز المتغيرات الفئوية، استبعاد الأعمدة المعرفية من التدريب، تحويل التاريخ إلى ميزات مشتقة، وبناء قاموس Top-4 لكل دولة وحل لمعالجة عدم التوازن وتحسين جودة التنبؤ. وقد ساهمت هذه الخطوات في تجهيز البيانات بشكل مثالي لبناء نموذج تنبؤي دقيق وفعّال.



10. المعالجة الأولية بعد التقرير

بعد تحليل البيانات باستخدام تقرير Pandas Profiling، تم تنفيذ مجموعة من الخطوات الإضافية لتحسين جودة البيانات وتهيئتها للنمذجة.

بدأت العملية بإزالة الصفوف المكررة لضمان عدم تكرار السجلات، ثم إعادة ترميز عمود is_lead إلى قيم ثنائية (1 و 0) بما يتوافق مع متطلبات النماذج التنبؤية.

كما تم حذف الأعمدة المساعدة المستخدمة في التحقق من صحة التواريخ، نظرًا لعدم الحاجة إليها في المراحل اللاحقة.

بعد ذلك، تم تحليل التوزيع الفئوي للأعمدة الأساسية وعرض النسب المئوية لكل فئة، بالإضافة إلى إنشاء جدول يوضح عدد القيم الفريدة في كل عمود فئوي.

نتائج المعالجة بعد التقرير : 📋 تقرير إزالة التكرارات:
- عدد الصفوف قبل الإزالة: 201988
- عدد الصفوف بعد الإزالة: 201284
- عدد الصفوف المكررة المحذوفة: 704

بعد الترميز: 📋is_lead[.0 .1] القيم الفريدة في العمود

الأعمدة المحذوفة: ['is_valid_date', 'activity_date_parsed']
عدد الأعمدة بعد الحذف: 11

📊 جداول التوزيع (عدد + نسبة مئوية) لكل عمود فئوي:

```
--- sourcesystem ---
                Count  Percentage (%)
sourcesystem
SFDC_US        125357           62.28
SFDC_ROW        53432           26.55
SFDC_GLOBAL     22471           11.16
SFDC_BEANWORKS     14            0.01
Other              10            0.00

--- opportunity_stage ---
                   Count  Percentage (%)
opportunity_stage
no_opp            158766           78.88
won                18082            8.98
lost               13785            6.85
other               5790            2.88
ongoing             4861            2.41

--- types ---
             Count  Percentage (%)
types
email       103991           51.66
call         57386           28.51
```

```
meeting       28512           14.17
appointment    6131            3.05
review         3395            1.69
other          1869            0.93

--- country ---
                    Count  Percentage (%)
country
US                 125382           62.29
FR                  53059           26.36
UK                  13098            6.51
DE                   5781            2.87
CA                   1021            0.51
BE                    588            0.29
CH                    560            0.28
IT                    412            0.20
IE                    353            0.18
AT                    296            0.15
NL                    290            0.14
Guadeloupe            168            0.08
Rî©union               83            0.04
LU                     71            0.04
Denmark                55            0.03
Martinique             40            0.02
Guyane Franî§aise      26            0.01
DK                      1            0.00

--- solution ---
           Count  Percentage (%)
solution
MRS       181205           90.02
Digital    18055            8.97
PLS         2024            1.01

📋 عدد القيم الفريدة في الأعمدة الفئوية:
              Column  Unique Values
0       sourcesystem              5
1  opportunity_stage              5
2              types              6
3            country             18
4           solution              3
```

11. معالجة عدم التوازن باستخدام خوارزمية SMOTENC :

تُظهر نتائج تحليل البيانات بعد التنظيف وجود عدم توازن واضح في المتغيرات الفئوية والهدفية، حيث تهيمن فئات معينة بشكل كبير على بقية الفئات. فعلى سبيل المثال، تمثل فئة no_opp في عمود opportunity_stage ما يقارب 79% من البيانات، بينما تشكل الفئات الأخرى نسبًا صغيرة جدًا. كما أن بعض الأعمدة الفئوية الأخرى تحتوي على فئات نادرة تقل نسبتها عن 1%، وهو ما يجعل النماذج التنبؤية تميل تلقائيًا إلى الفئات الأكثر شيوعًا وتتجاهل الفئات الأقل تمثيلًا.

في مثل هذه الحالات، يصبح استخدام SMOTENC ضروريًا لأنها مصممة خصيصًا للتعامل مع عدم توازن البيانات في وجود أعمدة فئوية(Categorical + Numerical) .

تعمل خوارزمية SMOTENC على توليد عينات اصطناعية للفئات الأقل تمثيلًا بطريقة تحافظ على طبيعة المتغيرات الفئوية دون تشويهها، وذلك عبر استخدام مسافات مناسبة (Hamming distance) بدلًا من المسافات العددية التقليدية.

يساعد هذا الأسلوب في تحسين قدرة النموذج على التمييز بين الفئات المختلفة، ويمنع انحيازه نحو الفئات المسيطرة، مما يؤدي إلى نموذج أكثر عدالة واستقرارًا ودقة في التنبؤ.

وبناءً على التوزيعات التي حصلنا عليها، فإن تطبيق SMOTENC يُعد خطوة أساسية قبل بناء أي نموذج تصنيفي، لأنه يعالج مشكلة التوازن الهيكلي في البيانات ويضمن أن النموذج يتعلم من جميع الفئات بشكل متوازن.

- **استخراج الفئات النادرة:**

يهدف هذا الكود إلى تحليل الأعمدة الفئوية في البيانات وتحديد الفئات التي تظهر بنسبة أقل من 1% داخل كل عمود.

تُعد هذه الخطوة مهمة قبل تطبيق خوارزميات إعادة التوازن مثل SMOTENC، لأنها تساعد في اتخاذ قرار بشأن كيفية التعامل مع الفئات النادرة، سواء بدمجها ضمن فئة "Other" أو استبعادها أو الإبقاء عليها كما هي.

يقوم الكود بالخطوات التالية:

- تحديد الأعمدة الفئوية الأساسية المراد تحليلها.
- حساب التوزيع النسبي لكل فئة داخل كل عمود.
- استخراج الفئات التي تقل نسبتها عن 1% وتصنيفها كفئات نادرة.
- تجميع النتائج في جدول منظم يحتوي على:
  - اسم العمود
  - اسم الفئة النادرة
  - نسبتها المئوية

يساعد هذا الجدول في فهم مدى انتشار الفئات الصغيرة داخل البيانات، مما يدعم اتخاذ قرارات مدروسة قبل مرحلة النمذجة أو إعادة التوازن.

في الأعمدة الفئوية (<1%) الفئات النادرة جدًا:

	Column	Category	Percentage
0	sourcesystem	SFDC_BEANWORKS	0.0070
1	sourcesystem	Other	0.0050
2	types	other	0.9285
3	country	CA	0.5072
4	country	BE	0.2921
5	country	CH	0.2782
6	country	IT	0.2047
7	country	IE	0.1754
8	country	AT	0.1471
9	country	NL	0.1441
10	country	Guadeloupe	0.0835
11	country	Rî©union	0.0412
12	country	LU	0.0353
13	country	Denmark	0.0273
14	country	Martinique	0.0199
15	country	Guyane Franî§aise	0.0129
16	country	DK	0.0005

- **دمج الفئات النادرة في فئة "Other":**

أظهر تحليل التوزيع الفئوي للبيانات وجود عدد من الفئات النادرة جدًا (أقل من 1%) في عدة أعمدة فئوية، مثل sourcesystem وtypes وخصوصًا عمود country الذي يحتوي على العديد من الدول ذات التمثيل الضعيف جدًا (أقل من 0.5%).

وجود مثل هذه الفئات الصغيرة يؤدي إلى مشكلات في النمذجة، خاصة عند استخدام خوارزميات تعتمد على التمثيل الإحصائي للفئات، أو عند تطبيق تقنيات إعادة التوازن مثل SMOTENC، حيث قد تتسبب الفئات النادرة في توليد عينات اصطناعية غير واقعية أو غير مستقرة.

لذلك تم تطبيق خطوة دمج الفئات النادرة ضمن فئة موحدة باسم "Other" باستخدام عتبة 1%، بحيث يتم الحفاظ على الفئات الرئيسية فقط، بينما تُجمع الفئات الصغيرة في فئة واحدة ذات حجم كافٍ.

يقوم الكود بحساب النسبة المئوية لكل فئة داخل كل عمود فئوي، ثم يحدد الفئات التي تقل نسبتها عن العتبة المحددة، ويستبدلها تلقائيًا بفئة "Other". كما يعرض عدد السجلات التي تم دمجها ونسبتها من إجمالي البيانات، مما يوفر شفافية كاملة حول تأثير هذه العملية.

تسهم هذه الخطوة في:

- تقليل التشتت داخل الأعمدة الفئوية
- تحسين استقرار النماذج التنبؤية
- منع تأثير الفئات الصغيرة على خوارزميات إعادة التوازن

- تبسيط التمثيل الفئوي قبل الترميز(Encoding)
- ضمان أن SMOTENC يعمل على فئات ذات حجم كافٍ لتوليد عينات اصطناعية ذات معنى

وبذلك تصبح البيانات أكثر اتساقًا واستعدادًا لمرحلة النمذجة، مع الحفاظ على البنية الفئوية دون فقدان المعلومات المهمة.

```
📋 العمود sourcesystem: ضمن (0.007%) سجل 14 تم دمج 'Other'
📋 العمود opportunity_stage: ضمن (0.0%) سجل 0 تم دمج 'Other'
📋 العمود types: ضمن (0.9285%) سجل 1869 تم دمج 'Other'
📋 العمود country: ضمن (1.9694%) سجل 3964 تم دمج 'Other'
📋 العمود solution: ضمن (0.0%) سجل 0 تم دمج 'Other'

✓ تم دمج جميع الفئات النادرة (<1%) في 'Other' كـ Categorical مع الحفاظ على الأعمدة
```

- **تحويل عمود التاريخ إلى ميزات مشتقة:**

تم تحويل عمود activity_dateإلى نوع تاريخي (datetime) لضمان إمكانية التعامل معه بشكل صحيح في التحليل والنمذجة.

خطوة بعد ذلك، تم استخراج ثلاث ميزات مشتقة من هذا العمود، وهي :السنة(year) ، الشهر(month) ، واليوم (day).

تساعد هذه الميزات في تحليل الأنماط الزمنية داخل البيانات، مثل التغيرات الموسمية أو الاتجاهات السنوية، كما تُعد أساسية في مرحلة Feature Engineeringلتحسين أداء النماذج التنبؤية.

أسهمت هذه العملية في إثراء البيانات بمتغيرات زمنية إضافية يمكن أن تعزز قدرة النموذج على فهم السلوك الزمني للأنشطة.

```
✓ تم استخراج الميزات المشتقة من activity_date: year, month, day
📋 الأعمدة الجديدة: ['activity_year', 'activity_month', 'activity_day']
```

- **تطبيق خوارزمية SMOTENC لإعادة توازن البيانات:**

نظرًا لوجود عدم توازن واضح في المتغير الهدف solution، حيث تهيمن فئة MRSعلى أكثر من 90% من البيانات، تم استخدام خوارزمية SMOTENCلإعادة توازن الفئات.

تم أولاً إنشاء نسخة احتياطية من البيانات لضمان إمكانية التراجع عند الحاجة، ثم فصل المتغير الهدف عن الميزات، وتحديد الأعمدة الفئوية المطلوبة للخوارزمية.

بعد تطبيق خوارزمية SMOTENC، أظهر التوزيع الجديد توازنًا كاملاً بين الفئات الثلاث (MRS، Digital، PLS)، مما يعزز قدرة النماذج اللاحقة على التعلّم من جميع الفئات دون انحياز للفئة الأكبر.

أسهمت هذه الخطوة في تحسين جودة البيانات وتجهيزها للنمذجة التنبؤية بطريقة أكثر عدالة واستقرارًا.

```
📋 تم إنشاء نسخة احتياطية من df_clean باسم df_backup

📋 الأعمدة الفئوية الداخلة في SMOTENC: ['sourcesystem', 'opportunity_stage', 'types', 'country
']
📋 الأعمدة العددية الداخلة في SMOTENC: ['is_lead', 'activity_year', 'activity_month', 'activit
y_day']

📊 توزيع solution قبل إعادة التوازن:
solution
MRS        181205
Digital     18055
PLS          2024
Other           0
Name: count, dtype: int64
solution
MRS        90.02
Digital     8.97
PLS         1.01
Other       0.00
Name: proportion, dtype: float64

📊 توزيع solution بعد إعادة التوازن:
solution
Digital    181205
MRS        181205
PLS        181205
Other           0
Name: count, dtype: int64
solution
Digital    33.33
MRS        33.33
PLS        33.33
Other       0.00
Name: proportion, dtype: float64

📋 حجم البيانات:
- قبل إعادة التوازن: 201284 سجل
- بعد إعادة التوازن: 543615 سجل
```

12. التقرير النهائي لجودة البيانات بعد التنظيف:

تم إجراء فحص شامل لجودة البيانات بعد الانتهاء من جميع خطوات التنظيف المتقدم وإعادة التوازن.
شمل التقرير ثلاثة محاور رئيسية:

1. **القيم المفقودة:**

تم التحقق من جميع الأعمدة، ولم يتم العثور على أي قيم مفقودة، مما يعكس اكتمال البيانات وجاهزيتها للنمذجة.

2. **القيم الشاذة في الأعمدة العددية:**

تم استخدام طريقة Interquartile Range (IQR) للكشف عن القيم الشاذة في المتغيرات العددية.

أظهر الفحص أن الأعمدة العددية لا تحتوي على قيم شاذة مؤثرة، مما يشير إلى استقرار التوزيع العددي.

3. **الفئات النادرة في الأعمدة الفئوية:(<1%)**

تم تحليل التوزيع الفئوي لجميع الأعمدة الفئوية، ولم تظهر أي فئات نادرة بعد عملية الدمج السابقة، مما يؤكد نجاح خطوة توحيد الفئات وتقليل التشتت.

يساعد هذا التقرير النهائي في التأكد من أن البيانات أصبحت نظيفة، متوازنة، وخالية من المشكلات الشائعة، مما يجعلها جاهزة تمامًا لمرحلة النمذجة التنبؤية.

```
📋 تقرير القيم المفقودة (NaN):
✅ لا توجد قيم مفقودة في البيانات

📋 تقرير القيم الشاذة (Outliers) في الأعمدة العددية:
Series([], dtype: int64)

📋 تقرير الفئات النادرة (<1%) في الأعمدة الفئوية:

--- account_id ---
account_id
0010L00001j3JeEQAU    0.637905
001b000003HA0UxAAL    0.322927
001b000003Ha49gAAB    0.258341
001b000003HA8AiAAL    0.230520
001E000000clx0LIAQ    0.205680
                        ...
0016g00000P3oj8AAB    0.000497
0010L00001j11YgQAI    0.000497
0016g00000P3ofSAAR    0.000497
0016g00000P3ofQAAR    0.000497
001b0000040SrI2AAK    0.000497
Name: proportion, Length: 12881, dtype: float64

--- sourcesystem ---
sourcesystem
Other             0.011923
SFDC_BEANWORKS    0.000000
Name: proportion, dtype: float64

--- who_id ---
who_id
0030y00002JqdZyAAJ    0.226049
```

```
0030L00001uiZKkQAM    0.140101
0030y00002HeyhqAAB    0.062598
0036g00000QeH8zAAF    0.057630
003Vw00000B51ziIAB    0.057630
                        ...
0034X00002xXXKBQA4    0.000497
003Vw00000EkW01IAF    0.000497
0030y00002Ux0WZAAZ    0.000497
003b00000206ARZAA2    0.000497
0036700003ycNNIAA2    0.000497
Name: proportion, Length: 51231, dtype: float64

--- opportunity_stage ---
opportunity_stage
Other    0.0
Name: proportion, dtype: float64

--- types ---
types
Other    0.928539
Name: proportion, dtype: float64

--- country_clean ---
country_clean
CA                  0.507243
BE                  0.292125
CH                  0.278214
IT                  0.204686
IE                  0.175374
AT                  0.147056
NL                  0.144075
Guadeloupe          0.083464
Rî©union            0.041235
LU                  0.035274
Denmark             0.027325
Martinique          0.019872
Guyane Franî§aise    0.012917
DK                  0.000497
Name: proportion, dtype: float64

--- country_standard ---
country_standard
Canada              0.507243
Belgium             0.292125
Switzerland         0.278214
Italy               0.204686
Ireland             0.175374
Austria             0.147056
Netherlands         0.144075
Guadeloupe          0.083464
Rî©union            0.041235
Luxembourg          0.035274
Denmark             0.027821
Martinique          0.019872
Guyane Franî§aise    0.012917
Name: proportion, dtype: float64
```

ثانياً: ✓ تجميع الحسابات حسب الدولة والحل، وإيجاد أفضل خمسة مسارات:

Group the accounts by country and solution, then find the top five paths:

1. بناء مسارات رحلة العميل (Customer Journey Paths):

بعد الانتهاء من تنظيف البيانات وتجهيزها، تم الانتقال إلى تحليل رحلة العميل من خلال بناء المسارات السلوكية لكل حساب.

تم أولًا ترتيب جميع الأنشطة زمنيًا لكل حساب باستخدام عمود activity_date لضمان أن المسار يعكس التسلسل الحقيقي للأحداث. بعد ذلك، تم تجميع الأنشطة الخاصة بكل حساب في سلسلة واحدة تمثل مسار العميل، مثل: email → call → meeting → call ، ويُعد هذا المسار وصفًا دقيقًا للتفاعل الزمني بين العميل والنظام.

يساعد هذا التحليل في فهم الأنماط السلوكية للعملاء، وتحديد المسارات الأكثر شيوعًا حسب الدولة أو نوع الحل، مما يمهّد للخطوة التالية في المشروع: استخراج أفضل خمسة مسارات لكل مجموعة، ثم استخدام نماذج التعلم الآلي لتحديد العوامل الأكثر تأثيرًا في نجاح أو فشل الفرصة.

```
=== عدد الحسابات التي تم بناء path لها ===
12881

=== معاينة من paths (أول 5 مسارات) ===
```

	account_id	path
0	0010L00001hVmFhQAK	email → call → call → call → call → call → cal...
1	0010L00001hVxd6QAC	meeting → meeting → review → email → email → e...
2	0010L00001hVyJQQA0	call → call → call → call → email → call → cal...
3	0010L00001hW1cAQAS	call
4	0010L00001ijzeFQAQ	email → email → email → call → call → email → ...

```
=== مثال على path واحد كامل لحساب واحد ===
```

	account_id	path
1800	0014X00002J4VetQAF	email

2. دمج مسارات العملاء مع معلومات الدولة والحل:

بعد بناء المسارات الزمنية لكل حساب اعتمادًا على ترتيب الأنشطة، تم دمج هذه المسارات مع البيانات الأساسية للحسابات، والتي تشمل الدولة (country_standard) ونوع الحل (solution) .

يهدف هذا الدمج إلى ربط كل حساب بمساره السلوكي الكامل، بحيث يصبح بالإمكان تحليل المسارات الأكثر شيوعًا داخل كل دولة، أو لكل حل، أو ضمن كل مجموعة تجمع بين الدولة والحل معًا.

تم أولًا استخراج سجل واحد لكل حساب لضمان عدم تكرار البيانات، ثم تم دمج جدول المسارات باستخدام معرف الحساب (account_id).

أسهمت هذه الخطوة في إنشاء جدول موحد يحتوي على:

- معلومات الحساب
- الدولة
- الحل
- المسار الكامل للأنشطة(Customer Journey Path)

ويمثل هذا الجدول الأساس الذي سيتم الاعتماد عليه في الخطوة التالية من المشروع، وهي استخراج أفضل خمسة مسارات لكل دولة ولكل حل، وتحليل الأنماط السلوكية للعملاء بطريقة أكثر دقة ووضوحًا.

=== معاينة من df_paths بعد الدمج ===

	account_id	country_standard	solution	path
0	0010L00001hVmFhQAK	United States	MRS	email → call → call → call → call → call → cal...
1	0010L00001hVxd6QAC	United States	MRS	meeting → meeting → review → email → email → e...
2	0010L00001hVyJQQA0	United States	MRS	call → call → call → call → email → call → cal...
3	0010L00001hW1cAQAS	United States	MRS	call
4	0010L00001ijzeFQAQ	United States	MRS	email → email → email → call → call → email → ...

3. معالجة المسارات غير الصالحة:

قبل تحليل مسارات رحلة العميل، تم تطبيق مجموعة من قواعد التصفية لضمان جودة المسارات المستخدمة في التحليل. شملت هذه القواعد إزالة المسارات الفارغة، واستبعاد الحسابات التي تحتوي على نشاط واحد فقط، والاحتفاظ فقط بالمسارات التي تتضمن انتقالات فعلية بين الأنشطة. ساعدت هذه الخطوة في تحسين دقة التحليل وضمان أن المسارات المستخدمة في حساب التكرارات وتمييز المسارات الأكثر شيوعًا تعكس سلوكًا حقيقيًا وذا معنى.

عدد المسارات الصالحة: 11317

	account_id	country_standard	solution	path	path_length
0	0010L00001hVmFhQAK	United States	MRS	email → call → call → call → call → call → cal...	12
1	0010L00001hVxd6QAC	United States	MRS	meeting → meeting → review → email → email → e...	16
2	0010L00001hVyJQQA0	United States	MRS	call → call → call → call → email → call → cal...	8
4	0010L00001ijzeFQAQ	United States	MRS	email → email → email → call → call → email → ...	13
6	0010L00001ikyKBQAY	United States	MRS	meeting → meeting → meeting → meeting → email	5

4. استخراج أفضل خمسة مسارات (Top 5 Customer Journeys) لكل دولة ولكل حل:

بعد بناء المسارات السلوكية لكل حساب وربطها بمعلومات الدولة والحل، تم تحليل تكرار هذه المسارات داخل كل مجموعة من (country_standard, solution) .

تم أولًا حساب عدد مرات ظهور كل مسار داخل كل مجموعة باستخدام التجميع الثلاثي (Country × Solution × Path).

بعد ذلك، تم ترتيب المسارات حسب تكرارها واستخراج أعلى خمسة مسارات تمثل السلوك الأكثر شيوعًا للعملاء في كل فئة.

يساعد هذا التحليل في تحديد الأنماط السلوكية المميزة لكل سوق ولكل نوع حل، مما يوفر رؤية واضحة حول المسارات الأكثر تأثيرًا وانتشارًا.

ويمثل هذا الجدول الناتج الأساس الذي سيتم الاعتماد عليه لاحقًا في بناء النظام النهائي الذي يعرض أفضل المسارات لكل حساب، بالإضافة إلى استخدام نماذج التعلم الآلي لتحديد العوامل المؤثرة في نجاح أو فشل الفرصة.

	country_standard	solution	path	count
0	Austria	Digital	appointment	0
1	Austria	Digital	appointment → Other	0
2	Austria	Digital	appointment → Other → Other → appointment → ap...	0
3	Austria	Digital	appointment → Other → Other → appointment → ap...	0
4	Austria	Digital	appointment → Other → Other → appointment → ap...	0
...	...	...	...	...
335	United States	Other	appointment	0
336	United States	Other	appointment → Other	0
337	United States	Other	appointment → Other → Other → appointment → ap...	0
338	United States	Other	appointment → Other → Other → appointment → ap...	0
339	United States	Other	appointment → Other → Other → appointment → ap...	0

340 rows × 4 columns

ثالثاً: ✓ بناء شجرة القرار:

Use the Decision Tree (DT) to determine the importance of the features to be used to present the next top four types of actions :

1. إعداد بيانات التنبؤ بالخطوة التالية (Next Action Prediction)

لتحليل رحلة العميل وتوقع الخطوة التالية في المسار، تم تحويل البيانات الزمنية إلى شكل مناسب لنماذج التعلم الآلي.

بدأت العملية بترتيب الأنشطة لكل حساب حسب التاريخ، ثم إنشاء متغير جديد يمثل النشاط التالي مباشرة باستخدام دالة shift داخل كل مجموعة من الحسابات.

بعد ذلك، تم حذف الصفوف التي لا تحتوي على خطوة لاحقة (آخر نشاط في المسار)، والاحتفاظ فقط بالمتغيرات الأساسية التي ستُستخدم في النمذجة، وهي: الدولة، الحل، النشاط الحالي، والنشاط التالي.

يمثل هذا الجدول الناتج الأساس الذي سيُستخدم في تدريب نموذج Decision Tree لتحديد العوامل الأكثر تأثيرًا في انتقال العميل من نشاط إلى آخر، وبالتالي دعم بناء نظام توصية لأفضل أربع خطوات تالية محتملة.

```
عدد الصفوف: 188403
عدد الأنشطة الفريدة: 6
عدد الأنشطة الهدف (next_action): 6
```

	account_id	country_standard	solution	types	next_action
0	0010L00001hVmFhQAK	United States	MRS	email	call
1	0010L00001hVmFhQAK	United States	MRS	call	call
2	0010L00001hVmFhQAK	United States	MRS	call	call
3	0010L00001hVmFhQAK	United States	MRS	call	call
4	0010L00001hVmFhQAK	United States	MRS	call	call
5	0010L00001hVmFhQAK	United States	MRS	call	call
6	0010L00001hVmFhQAK	United States	MRS	call	call
7	0010L00001hVmFhQAK	United States	MRS	call	call
8	0010L00001hVmFhQAK	United States	MRS	call	call
9	0010L00001hVmFhQAK	United States	MRS	call	call

يمثل هذا الجدول التوزيع النسبي لأكثر الخطوات التالية شيوعًا (next_action) لكل دولة ولكل حل.

بعد تحليل بيانات التنبؤ بالخطوة التالية، تم تجميع السجلات حسب الدولة (country_standard) والحل (solution)، ثم حساب النسب المئوية لكل نشاط محتمل أن يحدث بعد النشاط الحالي.

يعرض الجدول أعلى الأنشطة تكرارًا (Top-4) لكل مجموعة، مما يوفر رؤية واضحة حول السلوك المتوقع للعملاء في كل سوق ولكل حل. على سبيل المثال:

- في Austria – MRS ، النشاط التالي الأكثر احتمالًا هو email بنسبة 98.6%، يليه appointment بنسبة 0.9%.
- في Belgium – Digital ، النشاط التالي الأكثر شيوعًا هو email بنسبة 100%.
- في United States – MRS ، تتوزع الأنشطة التالية بين:
  - email: 54.7%
  - call: 27.5%
  - meeting: 14.9%
  - review: 1.9%

هذا التحليل مكّننا من بناء قاموس Top-4 لكل دولة وحل، والذي استخدم لاحقًا في النموذج الذكي لتقديم توصيات دقيقة حول النشاط التالي المتوقع، خاصة في الحالات التي لا يستطيع فيها النموذج التنبؤ بثقة عالية.

2. تحويل نتائج Top-4 إلى قاموس تنبؤي:

بعد استخراج أكثر أربع خطوات تالية شيوعًا لكل دولة ولكل حل، تم تحويل الجدول الناتج إلى قاموس برمجي باستخدام دالة to_dict(). يتيح هذا القاموس الوصول السريع إلى الأنشطة الأكثر احتمالًا لكل مجموعة (Country × Solution)، مما يجعله مناسبًا للاستخدام داخل النموذج أو كآلية بديلة (Fallback) في حال كانت ثقة النموذج منخفضة. يمثل هذا القاموس مكوّنًا أساسيًا في نظام التوصية بالخطوة التالية، لأنه يوفر معرفة مسبقة مبنية على الأنماط التاريخية الفعلية لسلوك العملاء.

3. استخراج أكثر الخطوات التالية شيوعًا(Top-4 Next Actions)

حساب أكثر أربع خطوات تالية شيوعًا(Top-4 Next Actions)

بعد تجهيز بيانات التنبؤ بالخطوة التالية، تم تحليل الأنماط السلوكية للعملاء عبر الدول والحلول المختلفة.

تم تجميع البيانات حسب الدولة (country_standard) و الحل(solution) ، ثم حساب التوزيع النسبي للخطوة التالية next_action داخل كل مجموعة.

بعد ذلك، تم استخراج أكثر أربع خطوات تالية شيوعًا لكل دولة ولكل حل، مما يوفر رؤية واضحة حول الأنشطة الأكثر احتمالًا أن يقوم بها العميل في سياقات مختلفة.

يمثل هذا الجدول الأساس لبناء قاموس Top-4 المستخدم لاحقًا في النموذج الذكي لتقديم توصيات دقيقة حول النشاط التالي المتوقع.

			next_action	appointment	call	email	meeting	Other	review
country_standard	solution	country_standard	solution						
Austria	Digital	Austria	Digital	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
	MRS	Austria	MRS	0.009132	0.000000	0.986301	0.000000	0.004566	0.000000
	PLS	Austria	PLS	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
	Other	Austria	Other	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Belgium	Digital	Belgium	Digital	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000
...	...	...	...	...	...	...	...	...	...
UK	Other	UK	Other	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
United States	Digital	United States	Digital	0.000000	0.255438	0.564546	0.147083	0.000000	0.023887
	MRS	United States	MRS	0.000000	0.275848	0.547733	0.149516	0.000000	0.019268
	PLS	United States	PLS	0.000000	0.228552	0.587131	0.157507	0.000000	0.022118
	Other	United States	Other	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

68 rows × 6 columns

4. ترميز المتغيرات الفئوية استعدادًا لبناء نموذج Decision Tree

نظرًا لأن نموذج Decision Tree لا يستطيع التعامل مباشرة مع المتغيرات الفئوية، تم استخدام خوارزمية Label Encoding لتحويل قيم الأعمدة الفئوية إلى أرقام.

شمل الترميز أربعة أعمدة رئيسية: الدولة(country_standard) ، الحل(solution) ، النشاط الحالي(types) ، والنشاط التالي(next_action).

تم إنشاء قاموس يحتوي على كائنات الترميز لكل عمود بهدف الاحتفاظ بخريطة التحويل وإمكانية إعادة فك الترميز لاحقًا عند عرض النتائج.

بعد ذلك، تم فصل البيانات إلى ميزات (X) ومتغير هدف(y) ، حيث يمثل X السياق الحالي للعميل، بينما يمثل y النشاط التالي الذي سيتم التنبؤ به.

بهذا تصبح البيانات جاهزة للانتقال إلى مرحلة تقسيم البيانات وبناء نموذج Decision Tree .

	account_id	country_standard	solution	types	next_action
0	0010L00001hVmFhQAK	16	1	3	2
1	0010L00001hVmFhQAK	16	1	2	2
2	0010L00001hVmFhQAK	16	1	2	2
3	0010L00001hVmFhQAK	16	1	2	2
4	0010L00001hVmFhQAK	16	1	2	2

5. تقسيم البيانات إلى مجموعتي تدريب واختبار(Train/Test Split):

بعد الانتهاء من ترميز المتغيرات الفئوية، تم تقسيم البيانات إلى مجموعتي تدريب واختبار باستخدام دالة train_test_split.

تم تخصيص 80% من البيانات للتدريب و20% للاختبار، مع استخدام معامل stratify لضمان الحفاظ على نفس توزيع الفئات في المتغير الهدف داخل المجموعتين.

يساعد هذا الإجراء في تقييم أداء نموذج Decision Tree بشكل عادل، حيث يتم اختبار النموذج على بيانات لم يسبق له رؤيتها، مما يعكس قدرته الحقيقية على التعميم والتنبؤ بالخطوة التالية في رحلة العميل.

```
عدد سجلات التدريب: 150722
عدد سجلات الاختبار: 37681

توزيع الفئات في y_train:
next_action
3    0.524
2    0.293
4    0.130
1    0.028
5    0.016
0    0.009
Name: proportion, dtype: float64

توزيع الفئات في y_test:
next_action
3    0.524
2    0.293
4    0.130
1    0.028
5    0.016
0    0.009
Name: proportion, dtype: float64
```

- **تقييم جودة تقسيم البيانات:**

بعد تقسيم البيانات إلى مجموعتي تدريب واختبار باستخدام stratified split، تم التحقق من توزيع الفئات في المتغير الهدف (next_action) داخل كل مجموعة.

أظهرت النتائج تطابقًا تامًا بين توزيع الفئات في مجموعة التدريب ومجموعة الاختبار، مما يؤكد نجاح عملية التقسيم في الحفاظ على التوازن النسبي للفئات.

يسهم هذا التوازن في ضمان أن نموذج Decision Tree سيتم تدريبه واختباره على بيانات تمثل الواقع بشكل عادل، مما يعزز دقة النموذج وقدرته على التعميم عند التنبؤ بالخطوة التالية في رحلة العميل.

6. إنشاء نموذج Decision Tree:

تم إنشاء كائن من نوع *DecisionTreeClassifier* باستخدام مكتبة scikit-learn .

تم ضبط عمق الشجرة على 6 لتقليل خطر الحفظ الزائد(*overfitting*) ، مع تثبيت معامل العشوائية لضمان إمكانية إعادة إنتاج النتائج.

7. تدريب نموذج Decision Tree :

بعد إنشاء نموذج Decision Tree وضبط معاييره، تم تدريب النموذج باستخدام بيانات التدريب فقط.

يهدف هذا التدريب إلى تعلم الأنماط التي تحكم انتقال العميل من نشاط إلى آخر، اعتمادًا على الدولة ونوع الحل والنشاط الحالي.

يتم استخدام هذه المعرفة لاحقًا لتقييم أداء النموذج على بيانات لم يسبق له رؤيتها، ثم استخراج أهمية الميزات التي تساهم في تحديد الخطوة التالية في رحلة العميل.

```
DecisionTreeClassifier
DecisionTreeClassifier(max_depth=6, random_state=42)
```

8. تقييم أداء نموذج Decision Tree :

بعد تدريب نموذجDecision Tree ، تم تقييم أدائه باستخدام بيانات التدريب والاختبار.

تم حساب دقة النموذج على كل من المجموعتين لمعرفة مدى قدرته على التعميم وعدم اعتماده على حفظ البيانات.

```
دقة التدريب: 0.6867
دقة الاختبار: 0.6856
```

- **تقييم أداء نموذجDecision Tree**

بعد تدريب نموذجDecision Tree ، تم تقييم أدائه باستخدام بيانات التدريب والاختبار.

أظهرت النتائج دقة تدريب بلغت 0.6867 ودقة اختبار بلغت 0.6856، وهو ما يشير إلى أن النموذج قادر على التعميم بشكل جيد دون وجود مشكلة حفظ زائد(Overfitting) ، حيث إن الفارق بين الدقتين ضئيل جدًا.

تُعد هذه الدقة مناسبة لطبيعة المهمة، نظرًا لتعدد الفئات وعدم توازنها، إضافة إلى أن التنبؤ بالخطوة التالية في رحلة العميل يمثل سلوكًا معقدًا بطبيعته.

يمثل هذا التقييم أساسًا قويًا للانتقال إلى تحليل أهمية الميزات.Feature Importance

9. استخراج أهمية الميزات Feature Importance :

بعد تدريب نموذج Decision Tree وتقييم أدائه، تم استخراج أهمية الميزات لتحديد العوامل الأكثر تأثيرًا في التنبؤ بالنشاط التالي في رحلة العميل.

تعتمد شجرة القرار على قياس مقدار مساهمة كل ميزة في تقليل عدم اليقين أثناء عملية اتخاذ القرار، مما يسمح بتحديد المتغيرات الأكثر تأثيرًا.

تم حساب أهمية الميزات الثلاث المستخدمة في النموذج: الدولة ((country_standard، نوع الحل ((solution، والنشاط الحالي (.(types

	feature	importance
2	types	0.927810
0	country_standard	0.056922
1	solution	0.015268

- **تحليل أهمية الميزات Feature Importance باستخدام Decision Tree**

بعد تدريب نموذج Decision Tree وتقييم أدائه، تم استخراج أهمية الميزات لتحديد العوامل الأكثر تأثيرًا في التنبؤ بالنشاط التالي في رحلة العميل.

أظهرت النتائج أن النشاط الحالي (types) يمثل العامل الأكثر تأثيرًا بنسبة 92.8%، مما يشير إلى أن انتقال العميل من خطوة إلى أخرى يعتمد بشكل رئيسي على نوع النشاط الذي يقوم به حاليًا.

أما الدولة (country_standard) فقد ساهمت بنسبة 5.7% فقط، بينما كان تأثير نوع الحل (solution) محدودًا جدًا بنسبة 1.5%.

تعكس هذه النتائج الطبيعة السلوكية للبيانات، حيث تتبع الأنشطة عادةً تسلسلًا منطقيًا ثابتًا، بينما تلعب العوامل الأخرى دورًا ثانويًا في تحديد الخطوة التالية.

★الخلاصة التحليلية:

- النموذج يعتمد بنسبة **93%** على النشاط الحالي لتحديد النشاط التالي.
  وهذا منطقي جدًا في نماذج **Next Action Prediction**.
- الدولة والحل لهما تأثير محدود، لكن وجودهما مفيد لتحسين الدقة قليلًا.

10. استخراج أفضل Top 4 Next Actions:

بعد تدريب نموذج Decision Tree وتقييم أدائه، تم استخدام احتمالات التنبؤ التي ينتجها النموذج لتحديد أفضل أربع خطوات تالية محتملة لكل صف من بيانات الاختبار.

يعتمد النموذج على دالة predict_proba التي توفر احتمال انتماء كل صف إلى كل فئة من فئات النشاط التالي.

تم ترتيب هذه الاحتمالات واختيار أعلى أربع فئات، ثم تحويلها من القيم الرقمية الناتجة عن الترميز إلى أسماء الأنشطة الأصلية باستخدام كائنات الترميز المحفوظة.

يوفر هذا الأسلوب رؤية واضحة حول الخيارات الأكثر احتمالًا للخطوة التالية في رحلة العميل، مما يدعم عملية اتخاذ القرار ويوفر أساسًا لتطوير توصيات موجهة تعتمد على البيانات.

	Top1	Top2	Top3	Top4
0	call	email	meeting	Other
1	call	email	meeting	Other
2	email	call	meeting	review
3	email	call	meeting	review
4	email	call	meeting	Other
5	call	email	meeting	Other
6	email	call	meeting	review
7	meeting	appointment	email	call
8	email	call	meeting	Other
9	email	call	meeting	review

تحليل نتائج Top 4 Next Actions:

أظهرت النتائج أن الأنشطة الأكثر احتمالًا كخطوة تالية هي call وemail، تليها meeting، بينما ظهرت أنشطة مثل Other و review كخيارات أقل احتمالًا.

يوفر هذا التحليل رؤية واضحة حول المسارات الأكثر شيوعًا في رحلة العميل، ويساعد في بناء توصيات عملية تعتمد على البيانات لدعم اتخاذ القرار.

رابعاً: ✓ بناء النظام الكامل Build a complete system:

1. بناء نظام يعطي Top 4 Actions:

1. Top 4 actions by country

تم تحليل البيانات لتحديد الأنشطة الأكثر شيوعًا لكل دولة، وذلك من خلال تجميع السجلات حسب الدولة ثم حساب تكرار الأنشطة المختلفة داخل كل مجموعة.

بعد ذلك تم استخراج أعلى أربع أنشطة تمثل المسارات الأكثر احتمالًا للعملاء داخل الدولة المحددة.

يساعد هذا التحليل في بناء توصيات مخصصة تعتمد على الموقع الجغرافي للحساب، مما يوفر رؤية أوضح حول السلوك المتوقع للعملاء في كل دولة.

```
Top 4 actions for country: Saudi Arabia
['appointment', 'call', 'email', 'meeting']
```

2. بناء نظام يعطي Top 4 Actions حسب Solution:

تم تحليل البيانات لتحديد الأنشطة الأكثر شيوعًا لكل نوع من الحلول (solution).

اعتمد التحليل على تجميع السجلات حسب الحل ثم حساب تكرار الأنشطة المختلفة داخل كل مجموعة.

بعد ذلك تم استخراج أعلى أربع أنشطة تمثل المسارات الأكثر احتمالًا للعملاء الذين يستخدمون نفس الحل.

يساعد هذا التحليل في بناء توصيات مخصصة تعتمد على نوع الحل المستخدم، مما يوفر رؤية أوضح حول السلوك المتوقع للعملاء داخل كل فئة من الحلول.

```
Top 4 actions for solution: CRM
['appointment', 'call', 'email', 'meeting']
```

3. بناء نظام يعطي Top 4 Actions حسب Solution & Country:

تم تحليل البيانات لتحديد الأنشطة الأكثر شيوعًا داخل كل مجموعة تجمع بين الدولة ونوع الحل.

يعتمد هذا التحليل على تصفية البيانات حسب الدولة والحل معًا، ثم حساب تكرار الأنشطة داخل هذه المجموعة، واختيار أعلى أربع أنشطة تمثل المسارات الأكثر احتمالًا للعملاء الذين ينتمون إلى نفس الدولة ويستخدمون نفس الحل.

يساعد هذا الأسلوب في بناء توصيات أكثر دقة، لأنه يجمع بين الخصائص الجغرافية والوظيفية للحساب.

```
Top 4 actions for country: Saudi Arabia | solution: CRM
['appointment', 'call', 'email', 'meeting']
```

2. نظام الأوزان الديناميكي Dynamic Weight Adjustment Algorithm:

لتوفير توصيات أكثر دقة وتفاعلية، تم تطوير نظام ديناميكي لتعديل أوزان الأنشطة بناءً على آخر تفاعل يقوم به العميل.

يعتمد هذا النظام على وزن أساسي لكل نشاط (Base Weight)، ويتم تعديل هذه الأوزان عند حدوث نشاط جديد باستخدام معادلة تأخذ في الاعتبار وزن آخر نشاط (Last Touch Weight).

يسمح هذا الأسلوب بإعادة ترتيب الأنشطة الأكثر احتمالًا بشكل ديناميكي، بحيث تعكس التوصيات السلوك الفعلي للعميل.

وبذلك يصبح النظام قادرًا على تقديم Top 4 Actions محدثة في كل مرة يتم فيها تسجيل نشاط جديد.

خوارزمية تعديل الأوزان الديناميكية تهدف إلى تحديث أوزان الأنشطة بناءً على آخر نشاط قام به المستخدم.

الفكرة الأساسية هي أن آخر نشاط (last_touch) يعكس نية العميل الحالية، لذلك يحتفظ بوزنه الأصلي، بينما يتم تخفيض أوزان الأنشطة الأخرى باستخدام المعادلة:

$$\text{New Weight}_{\text{Last\_touch}} = \text{Base Weight}_{\text{Last\_touch}} \times (1 - \text{Last Touch Weight})$$

بهذا الشكل، كلما كان وزن آخر نشاط كبيرًا، تقل أهمية الأنشطة الأخرى، مما يجعل التوصيات أكثر انسجامًا مع السلوك الفعلي للعميل.

3. 📘 دالة بناء الرحلة الذكية (Smart Trip Generation):

📘 دالة بناء الرحلة الذكية (recalc_top4_with_new_action):

تهدف دالة recalc_top4_with_new_action إلى توليد مسار تفاعلي ذكي للعميل (Customer Journey) اعتمادًا على النموذج المدرَّب، مع دمج المعرفة المستخرجة من البيانات التاريخية.

تعمل الدالة وفق منهجية متعددة الطبقات تجمع بين التنبؤ الإحصائي، والأنماط السلوكية، والضبط الديناميكي للأوزان، مما يجعل الرحلة الناتجة أكثر واقعية وتنوعًا.

1. تحويل النشاط الابتدائي إلى تمثيل رقمي:

يتم أولًا تحويل النشاط الذي يدخله المستخدم من صيغة نصية إلى قيمة رقمية باستخدام محوّل الترميز (Label Encoder) ، هذا التحويل ضروري لأن نموذج شجرة القرار يعتمد على مدخلات رقمية فقط.

2. تهيئة الأوزان الديناميكية:

تبدأ جميع الأنشطة بأوزان متساوية، ثم تتغير هذه الأوزان تدريجيًا أثناء بناء الرحلة، يسمح هذا الأسلوب بتقليل تكرار النشاط نفسه بشكل متتالٍ، مما يجعل المسار أكثر طبيعية وواقعية.

3. دمج المعرفة التاريخية (Top-4 لكل دولة وحل)

تستخرج الدالة الأنشطة الأكثر شيوعًا لكل (Country, Solution) من البيانات التاريخية.

تُستخدم هذه الأوزان لتوجيه النموذج نحو المسارات التي حدثت فعليًا في الماضي، مما يعزز دقة التنبؤ.

4. توقع احتمالات النشاط التالي باستخدام Decision Tree

في كل خطوة، يتم تجهيز صف بيانات يحتوي على:

- الدولة
- الحل
- النشاط الحالي

ثم يُستخدم النموذج المدرَّب لتوقع احتمالات جميع الأنشطة الممكنة.

5. دمج ثلاث طبقات من الأوزان

تقوم الدالة بدمج:

1. احتمالات النموذج ((Decision Tree
2. أوزان Top-4 التاريخية
3. الأوزان الديناميكية التي تتغير مع كل خطوة

ينتج عن هذا الدمج وزن نهائي لكل نشاط، يعكس كلًا من:

- التنبؤ الإحصائي
- الأنماط السلوكية الواقعية
- التنويع الديناميكي

6. اختيار النشاط التالي باحتمال موزون

بعد ترتيب الأنشطة حسب الوزن النهائي، يتم اختيار النشاط التالي من بين أفضل ثلاثة باستخدام توزيع احتمالي.

هذا الأسلوب يمنع المسارات الجامدة ويضيف عنصرًا من التنوع الواقعي.

7. منع التكرار الممل

إذا تكرر نفس النشاط أكثر من مرتين، تقوم الدالة تلقائيًا باختيار نشاط بديل من قائمة أفضل الخيارات، مما يحافظ على جودة الرحلة.

8. تحديث الأوزان والاستمرار في بناء المسار

بعد اختيار النشاط التالي:

- يتم تحديث الأوزان الديناميكية
- يُحوّل النشاط إلى ترميز رقمي
- وتستمر العملية حتى الوصول إلى عدد الخطوات المطلوب

9. إرجاع الرحلة النهائية

في النهاية، تُرجع الدالة قائمة مرتّبة تمثل "أفضل رحلة" متوقعة للعميل، مبنية على مزيج من التنبؤات الإحصائية والسلوك التاريخي والضبط الديناميكي.

النتيجة: هذا يعكس بوضوح أن الدالة ليست مجرد تنبؤ بسيط، بل نظام ذكي متعدد الطبقات يجمع بين:

- **التعلم الآلي**
- **التحليل السلوكي**
- **النمذجة الاحتمالية**
- **الضبط الديناميكي**

مثال على خوارزمية تعديل الأوزان:

```
recalc_top4_with_new_action(
"   United States",
"   Digital",
"   email",
   new_action="call"
(
```

```
['email', 'call', 'meeting', 'review']

}
```

4. حفظ النموذج الذكي ومكوناته في ملف واحد:

بعد الانتهاء من تدريب نموذج شجرة القرار Decision Tree وترميز المتغيرات الفئوية، بالإضافة إلى حساب مصفوفة Top-4 الخاصة بكل دولة وحل، تم حفظ جميع المكوّنات الضرورية لتشغيل النموذج داخل ملف واحد بصيغة PKI. باستخدام مكتبة joblib.

يتيح هذا الأسلوب إمكانية تحميل النموذج لاحقًا داخل التطبيق دون الحاجة إلى إعادة التدريب أو إعادة حساب الترميزات.

تتضمن البيانات المحفوظة العناصر التالية:

- model: النموذج المدرَّب (DecisionTreeClassifier)
- types_encoder: محوّل ترميز الأنشطة الحالية
- next_encoder: محوّل ترميز الأنشطة التالية
- top4_by_country_solution: قاموس يحتوي على أكثر الأنشطة شيوعًا لكل (Country, Solution)

يتم حفظ هذه العناصر داخل ملف واحد باسم: **dt_model.pkl** يُحفظ تلقائياً في مجلد المشروع.

مما يجعل عملية نشر النموذج وتشغيله في الواجهة النهائية أكثر سهولة وموثوقية.

يقوم كود الحفظ بتوليد ملفين أساسيين:

- dt_model.pkl: يحتوي على نموذج شجرة القرار المدرب.
- encoders.pkl: يحتوي على جميع الـ Label Encoders المستخدمة في ترميز المتغيرات.

خامساً: واجهة التطبيق Console Interface (app.py) :

تُعد واجهة التطبيق النصية (Console Interface) المرحلة النهائية في تشغيل النموذج الذكي، حيث تتيح للمستخدم إدخال البيانات المطلوبة والحصول على رحلة تفاعلية مبنية على النموذج المدرَّب.

تم تصميم الواجهة لتكون بسيطة، مباشرة، وسهلة الاستخدام، مع الحفاظ على التكامل الكامل مع النموذج وعمليات التنبؤ.

1. تحميل النموذج الذكي ومكوّناته:

عند تشغيل التطبيق، يقوم البرنامج أولًا بتحميل ملف النموذج المحفوظ `smart_model_bundle.pkl`، والذي يحتوي على:

- نموذج شجرة القرار المدرَّب
- محوّلات الترميز (Label Encoders)
- قاموس Top-4 الخاص بكل دولة وحل

يسمح هذا الأسلوب باستخدام النموذج مباشرة دون الحاجة إلى إعادة التدريب أو إعادة تجهيز البيانات.

2. استقبال مدخلات المستخدم

تطلب الواجهة من المستخدم إدخال أربعة عناصر أساسية:

1. **الدولة(country_standard)**
2. **الحل(solution)**
3. **النشاط الحالي(current action)**
4. **عدد الخطوات المطلوبة في الرحلة**

تمثل هذه المدخلات نقطة البداية لبناء المسار الذكي، وهي نفس المتغيرات التي استخدمت أثناء تدريب النموذج.

3. تجهيز المدخلات للتنبؤ

بعد إدخال البيانات، يقوم التطبيق بتحويل النشاط الحالي من صيغة نصية إلى قيمة رقمية باستخدام محوّل الترميز، ثم تجهيز صف بيانات (DataFrame) مطابق تمامًا لبنية البيانات المستخدمة أثناء التدريب.

هذا يضمن أن النموذج يستقبل المدخلات بالشكل الصحيح.

4. استدعاء دالة بناء الرحلة الذكية

يتم تمرير المدخلات إلى الدالة:

build_super_smart_trip(model, start_action, country, solution, steps)

وتتولى هذه الدالة:

- توقع النشاط التالي باستخدام النموذج
- دمج احتمالات Decision Tree مع أوزانTop-4
- تطبيق الأوزان الديناميكية لمنع التكرار
- بناء رحلة كاملة خطوة بخطوة

النتيجة النهائية هي مسار تفاعلي يعكس السلوك التاريخي والذكاء التنبؤي للنموذج.

5. عرض النتيجة للمستخدم

بعد اكتمال بناء الرحلة، تقوم الواجهة بطباعة المسار النهائي بشكل واضح وسهل القراءة، مثل:

Best Trip: call → email → email → appointment → appointment

يتيح هذا للمستخدم فهم التسلسل المقترح للأنشطة، واستخدامه في التحليل أو اتخاذ القرار.

النتيجة: واجهة التطبيق تمثل الطبقة النهائية في النظام، حيث تربط بين:

- النموذج المدرَّب
- الدوال الذكية
- المستخدم النهائي

وتوفر طريقة بسيطة وفعّالة لتوليد رحلات تفاعلية مبنية على الذكاء الاصطناعي والتحليل السلوكي.

محتويات مجلد المشروع:

يحتوي مجلد المشروع على مجموعة من الملفات والبرمجيات التي تشكل النظام الكامل للتنبؤ بالرحلة الذكية للعميل. تم تنظيم الملفات بطريقة تسهّل تشغيل النموذج، وتوضّح الفصل بين مراحل التدريب، الحفظ، والتشغيل. فيما يلي وصف موجز لأهم الملفات داخل المجلد:

1. ملف Notebook الخاص بالتدريب (Customer_Journey.ipynb):

يحتوي على جميع خطوات معالجة البيانات، الترميز، حسابTop-4 ، تدريب نموذجDecision Tree ، وحفظ النموذج النهائي.

يمثل هذا الملف المرحلة البحثية/التطويرية للمشروع.

2. ملف النموذج المحفوظ (smart_model_bundle.pkl):

3. ملف يحتوي على:

- النموذج المدرَّب
- محوّلات الترميز (types_encoder و(next_encoder
- قاموس Top-4 الخاص بكل دولة وحل

يُستخدم هذا الملف مباشرة داخل واجهة التطبيق دون الحاجة لإعادة التدريب.

4. واجهة التطبيق (app.py):

الملف التنفيذي الذي يتفاعل معه المستخدم، يقوم بـ:

- تحميل النموذج المحفوظ
- استقبال مدخلات المستخدم
- استدعاء دالة بناء الرحلة الذكية
- عرض المسار النهائي

يمثل هذا الملف الطبقة التشغيلية (Production Layer) للمشروع.

5. مجلد البيانات (data):

في هذا المشروع مجلد البيانات هو نفسه مجلد المشروع، لم يتم استخدام مجلد بيانات مستقل، حيث تم وضع ملفات البيانات مباشرة داخل مجلد المشروع الرئيسي. لكون ملفات البيانات محدودة ويسهل الوصول إليها دون الحاجة إلى بنية مجلدات إضافية، وتتضمن هذه الملفات مايلي:

- **data_all.xlsx**: نسخة من البيانات المجمّعة بصيغةExcel .
- **data_all.xltx** قالب Excel قابل لإعادة الاستخدام.
- **data_all.csv** : ملف يحتوي على جميع البيانات المجمّعة بصيغة CSV .
- **data_original.csv** : ملف البيانات الخام قبل أي تنظيف أو معالجة.
- **cleaned_data.csv** : نسخة من البيانات بعد تطبيق خطوات التنظيف العامة.
- **data_clean_country_final.csv:** نسخة نهائية من البيانات بعد معالجة أسماء الدول وتوحيدها.

- **column_names_reference.json:** ملف مرجعي يحتوي على أسماء الأعمدة الأصلية أو المترجمة.
- **MCS_MLT_HW1_S25_C2_maha_336128.pdf:** التقرير الخاص بالمشروع.

**** تمت بعون الله تعالى ****

المراجع:

- أملية مقرر تقانات تعلم الآلة- ماجستير علوم الحاسوب- الجامعة الافتراضية السورية.
- E-Book_Foundations of Machine Learning.pdf