BLACKBIRD FOR TRAINING

دورة علم البيانات للمهنيين: التحليل وتعلم الآلة ونشر النماذج (أسبوعين)



دورة علم البيانات للمهنيين: التحليل وتعلم الآلة ونشر النماذج (أسبوعين)

علم البيانات هو جمع البيانات وتنظيفها وتحليلها بالإحصاء وتعلم الآلة للإجابة عن الأسئلة وبناء تنبؤات تدعم القرارات. تمتد هذه الدورة أسبوعين، وهي موجهة إلى المهنيين وقادة الفرق الذين يريدون مهارات عملية بلغة Python لا مجرد نظرة عامة، وتنتهي بمشروع ختامي يضم نموذجا تنبؤيا مختبرا وتقرير تقييم ولوحة معلومات وخطة نشر.

مقدمة

تنظم الدورة ثلاثة مراجع. الأول CRISP-DM، وهو المنهج القياسي العابر للصناعات في التنقيب عن البيانات، ويحدد ست مراحل من فهم الأعمال حتى النشر. والثاني منظومة Python للبيانات، وأساسها pandas لمعالجة البيانات scikit-learng لتعلم الآلة، وتستخدم في كل تمرين. والثالث Apache Spark الذي يعرف بالمعالجة الموزعة للبيانات التي تتجاوز سعة جهاز واحد. المستوى تطبيقي. ومقارنة بشهادة الأيام الخمسة، تضيف الأيام العشرة تدريبا على البرمجة والتعلم الموجه وغير الموجه وتقييم النماذج ومشروعا ختاميا، مع مجموعات بيانات من التجزئة والمصارف والخدمات العامة تعالج في دفاتر برمجية.

أهداف الدورة

- تطبيق مراحل CRISP-DM لتخطيط مشروع علم بيانات من السؤال حتى النشر.
- تفسير الإحصاءات الاستكشافية واختبارات الفرضيات لتحديد المتغيرات التي تستحق النمذجة.
- التمييز بين الانحدار والتصنيف واختيار الخوارزمية المناسبة لكل مشكلة أعمال.
- المقارنة بين الدقة والاستدعاء وتبرير الموازنة التي تناسب تكلفة كل نوع من الأخطاء.
- الاختيار بين مجمعات التجميع والتعزيز وضبطها بالتحقق المتقاطع.
- تعريف خطة نشر ومراقبة تحافظ على فائدة النموذج بعد إطلاقه.

الفئة المستهدفة

- المحللون المنتقلون إلى علم البيانات الذين يقررون أسلوب النمذجة المناسب لسؤال العمل.
- قادة وحدات التحليل والتقارير الذين يحكمون على موثوقية النموذج قبل استخدامه.
- المهندسون ومختصو تقنية المعلومات الذين يقررون طريقة تشغيل خطوط البيانات وتقييم النماذج في بيئة الإنتاج.
- الباحثون والمخططون في الجهات العامة الذين يختارون بين الاختبارات الإحصائية والنماذج التنبؤية.
- مختصو المالية والتسويق والعمليات الذين يقررون أي التنبؤات يستحق البناء.



محاور الدورة

اليوم 1: أدوار علم البيانات ودورة حياة المشروع

- CRISP-DM مقابل التحليل العشوائي: منهج من ست مراحل مقابل استكشاف غير مخطط
- عالم البيانات مقابل محلل البيانات مقابل مهندس البيانات: المسؤوليات والمهارات
- دورة حياة البيانات: مراحل الجمع والتخزين والتحضير والنمذجة والنشر والإيقاف
- أنواع البيانات وبنياتها: السجلات الرقمية والفئوية والنصية والزمنية
- إعداد بيئة Python: دفاتر Jupyter ومكتبة pandas وأول ملخص لمجموعة بيانات

اليوم 2: اكتساب البيانات وإدارتها

- المصادر المهيكلة مقابل غير المهيكلة: جداول قواعد البيانات مقابل المستندات والصور والسجلات
- خطوط ETL ومستودعات البيانات: الاستخراج والتحويل والتحميل إلى مخازن التحليل
- حوكمة البيانات والخصوصية والامتثال: الملكية وحقوق الوصول وقواعد الاحتفاظ
- أبعاد جودة البيانات: تطبيق فحوص الدقة والاكتمال والحداثة والاتساق
- الاتصال بمجموعات بيانات حقيقية: واجهات API واستعلامات SQL واستيراد الجداول في Python

اليوم 3: التحليل الاستكشافي والتنظيف

- تحليل البيانات الاستكشافي: الإحصاءات الموجزة والتوزيعات ورصد الأنماط الأولية
- القيم المفقودة مقابل القيم المتطرفة: خيارات التعويض مقابل قواعد الكشف والمعالجة
- التكرارات والشذوذ: تحديد السجلات غير المتسقة ومعالجتها باستخدام pandas
- هندسة الخصائص: الترميز والتحجيم وإنشاء متغيرات مشتقة للنماذج
- التحليلات المرئية للاستكشاف: المدرجات التكرارية والمخططات الصندوقية والخرائط الحرارية والمخططات الزوجية

اليوم 4: الإحصاء واختبار الفرضيات

- الإحصاء الوصفي مقابل الاستدلالي: تلخيص العينات مقابل استنتاج خصائص المجتمع
- التوزيعات الاحتمالية والمعاينة: شرح التوزيع الطبيعي وذي الحدين وخطأ المعاينة
- اختبار الفرضيات: الفرضية الصفرية والقيمة الاحتمالية واختبار t واختبار مربع كاي
- الارتباط مقابل السببية: لماذا لا يحرك المتغيران المرتبطان أحدهما الآخر بالضرورة
- دراسة حالة إحصائية: اختبار أثر تغيير السعر على مبيعات التجزئة



اليوم 5: التعلم الموجه

- الانحدار مقابل التصنيف: التنبؤ بالكميات المتصلة مقابل تحديد الفئات المعنونة
- تقسيم البيانات إلى تدريب وتحقق واختبار: حجز بيانات لتقييم نزيه
- الانحدار الخطي واللوجستي: نماذج أساسية للتعلم الموجه مبنية بمكتبة scikit-learn
- اختيار الخصائص: أساليب الترشيح والتغليف والتضمين لاختيار المدخلات
- بناء نموذج تنبؤي: أول نموذج تعلم موجه على بيانات العملاء

اليوم 6: النماذج المجمعة والشبكات العصبية والتعلم غير الموجه

- التجميع مقابل التعزيز: الغابات العشوائية مقابل التعزيز المتدرج للبيانات الجدولية
- أشجار القرار: قواعد التفرع وحدود العمق وقابلية التفسير لمستخدمي الأعمال
- مفاهيم الشبكات العصبية والتعلم العميق: الطبقات ودوال التنشيط وأساسيات التدريب
- العنقدة وتقليص الأبعاد: خوارزمية K-Means وتحليل PCA للبيانات غير المعنونة
- دراسة حالة لسلوك العملاء: مقارنة التنبؤ بتسرب العملاء واكتشاف الشرائح

اليوم 7: تقييم النماذج وضبطها

- الدقة مقابل الاستدعاء: موازنة الأخطاء حين تختلف تكلفة الإنذار الكاذب عن الإخفاق
- موازنة التحيز والتباين: تشخيص فرط التخصيص ونقصه من منحنيات التعلم
- التحقق المتقاطع وضبط المعاملات الفائقة: شرح البحث الشبكي والبحث العشوائي
- مقاييس الانحدار: MAE وRMSE ومعامل التحديد للتنبؤات المتصلة
- منحنى ROC ومساحة AUC: مقارنة نماذج التصنيف بمعزل عن العتبة

اليوم 8: التصوير المرئي وسرد البيانات

- المرئيات الاستكشافية مقابل التفسيرية: رسوم المحللين مقابل رسوم الإدارة العليا
- مبادئ تصميم لوحات المعلومات: التخطيط والتسلسل البصري واختيار نوع الرسم
- Python Matplotlib Seaborng: مقارنة بأدوات ذكاء الأعمال Power Blg Tableaug
- كتابة التقارير التنفيذية: تحويل نتائج النماذج إلى روايات أعمال واضحة
- بناء لوحة معلومات تفاعلية: المرشحات والتعمق في بيانات المبيعات



اليوم 9: البيانات الضخمة وتحليل النصوص والنشر

- Hadoop مقابل Spark: المعالجة الدفعية على الأقراص مقابل الحوسبة الموزعة في الذاكرة
- التحليلات السحابية والفورية: خطوط البيانات المتدفقة وخدمات التحليل المدارة
- أساسيات معالجة اللغة الطبيعية وتحليل النصوص: التقطيع وقياس المشاعر ونمذجة الموضوعات
- نشر النماذج: التقييم الدفعي وواجهات API ومراقبة انجراف النموذج
- الدمج في الأعمال: مؤشرات الأداء وقياس العائد وأتمتة سير العمل ونمذجة السيناريوهات

اليوم 10: تنفيذ مشروع علم البيانات الختامي

- صياغة مشكلة المشروع الختامي: سؤال العمل ومقياس النجاح وخطة البيانات
- نمذجة المشروع الختامي: تطبيق التنظيف والخصائص والخوارزميات على مجموعة بيانات واحدة
- نقد الزملاء مقابل التقييم الذاتي: مراجعة خيارات النموذج واختبار المعرفة النهائي
- خارطة التحول بالبيانات: التخطيط المؤسسي ومسارات التعلم المستمر
- النموذج الختامي وخطة النشر: عرض تقرير التقييم ولوحة المعلومات

المهارات التي ستكتسبها

- معالجة البيانات باستخدام pandas
- هندسة الخصائص
- اختبار الفرضيات
- التعزيز المتدرج
- العنقدة بخوارزمية K-Means
- التحقق المتقاطع
- مراقبة انجراف النماذج
- السرد بلوحات المعلومات

لماذا تحضر هذه الدورة

- من القراءة عن تعلم الآلة إلى كتابة دفاتر Python تنظف البيانات وتدرب النماذج.
- من الحكم على النموذج بالدقة الإجمالية وحدها إلى اختيار مقاييس تعكس التكلفة الفعلية للأخطاء.
- من تحليلات منفردة إلى هيكل مشروع متكرر وفق CRISP-DM.
- من تمارين متفرقة إلى نموذج تنبؤي ختامي مع تقرير تقييم ولوحة معلومات وخطة نشر.



الخاتمة

يحول علم البيانات السجلات الخام إلى تنبؤات وتفسيرات مختبرة يمكن لصناع القرار الوثوق بها. وتوضح الدورة الحد الفاصل بين الارتباط والسببية، وبين التعلم الموجه وغير الموجه، وبين نموذج يحقق نتيجة جيدة في الدفتر البرمجي ونموذج يصمد بعد النشر. وهي مناسبة للمهنيين الذين يتعاملون مع البيانات أصلا ويريدون أسبوعين من البرمجة والنمذجة ومشروعا ختاميا بدلا من مقدمة أقصر للمفاهيم.

الأسئلة الشائعة عن الدورة

ما هو علم البيانات للمهنيين؟

هو استخدام البرمجة والإحصاء وتعلم الآلة من قبل العاملين في الأعمال أو الحكومة أو المجالات الفنية لتنظيف البيانات واكتشاف الأنماط وبناء نماذج تتنبأ بنتائج تخص عملهم.

ما الفرق بين الدقة والاستدعاء في علم البيانات؟

الدقة هي نسبة التنبؤات الإيجابية الصحيحة من كل ما تنبأ به النموذج إيجابيا، والاستدعاء هو نسبة الحالات الإيجابية الفعلية التي يكتشفها النموذج، ورفع أحدهما يخفض الآخر عادة، فيتوقف الاختيار على الخطأ الأعلى تكلفة.

ما المنهج الذي تتبعه دورة علم البيانات للمهنيين؟

تتبع CRISP-DM في هيكلة المشروع، وتستخدم Python مع pandas scikit-learng للتحليل والنمذجة، مع تعريف بمنصة Spark Apache للبيانات الضخمة.

هل تناسب دورة علم البيانات لعشرة أيام محلل الأعمال؟

نعم، لمحلل الأعمال الذي يعمل بالجداول أو أدوات التقارير ويريد بناء نماذج تنبؤية وتقييمها بلغة Python.

ماذا ينتج المشارك في دورة علم البيانات للمهنيين؟

يكمل كل مشارك مشروعا ختاميا يضم نموذجا تنبؤيا مدربا ومقيما على مجموعة بيانات واحدة، وتقرير تقييم مكتوبا، ولوحة معلومات تفاعلية، وخطة للنشر والمراقبة.

