

السلام عليكم ورحمة الله وبركاته

تحية طيبة للدكتور ياسر

اكتب لك خطوات العمل مع روابط الداتا والتطبيق المشروع

طبعاً أول شيء بدأنا بتحميل البيانات من Kaggle ، واستخدمنا بيانات Backblaze الخاصة بأقراص التخزين. بعد قراءة ملف البيانات لاحظنا أن عدد الصفوف كان كبير جداً، تقريباً ثلاثة ملايين صف، وعدد الأعمدة حوالي ٩٥ عمود.

بعد ذلك بدأنا مرحلة تنظيف البيانات. حذفنا الأعمدة التي لا تفيد في التنبؤ مثل التاريخ والرقم التسلسلي ونوع الموديل، لأنها لا تساعد النموذج على فهم احتمالية الفشل. كذلك تعاملنا مع القيم المفقودة، فحذفنا الأعمدة التي تحتوي على نسبة كبيرة من القيم الفارغة، واستبدلنا القيم الناقصة في الأعمدة الأخرى بصفر حتى تصبح البيانات جاهزة للتدريب.

ثم لاحظنا مشكلة مهمة جداً، وهي أن البيانات غير متوازنة. أغلب الأقراص كانت سليمة، ونسبة الأعطال كانت قليلة جداً. هذا يجعل النموذج يتعلم أن يقول دائماً "لا يوجد عطل"، وبالتالي يعطي دقة عالية لكنها غير حقيقية. لذلك قمنا بأخذ عينة من الأجهزة السليمة واحتفظنا بكل حالات الفشل، حتى نجعل البيانات أكثر توازناً أثناء التدريب.

بعد تجهيز البيانات، قسمناها إلى جزئين: جزء للتدريب وجزء للاختبار. استخدمنا ٨٠٪ من البيانات لتدريب النموذج و ٢٠٪ لاختباره، حتى نقيم أداءه بطريقة صحيحة.

جربنا نموذجين للتنبؤ:

الأول Random Forest لأنه نموذج قوي ومستقر ويعطي نتائج جيدة كبدائية. والثاني XGBoost لأنه عادةً يتعامل بشكل أفضل مع البيانات المعقدة ويحقق أداء أعلى في كثير من الحالات.

بعد التدريب، قمنا بتقييم الأداء باستخدام عدة مقاييس مهمة. حسبنا Precision لمعرفة دقة تنبؤات العطل، و Recall لمعرفة كم حالة فشل اكتشفها النموذج فعلياً، وكذلك حسبنا AUC لمعرفة قدرة النموذج على التفريق بين القرص السليم والمعتّل.

لاحظنا أن النموذج كان يعطي Precision عالي، لكنه يفشل في اكتشاف عدد كافٍ من الأعطال، أي أن الـ Recall كان منخفض. لذلك قمنا بتعديل قيمة العتبة (Threshold) بدلاً من تركها على القيمة الافتراضية. بعد تقليل العتبة، ارتفع الـ Recall بشكل واضح، وهذا مهم جداً في أنظمة الصيانة التنبؤية لأن اكتشاف العطل مبكراً أهم من تقليل عدد الإنذارات الكاذبة.

في النهاية قمنا بتحليل أهم الخصائص المؤثرة في التنبؤ باستخدام XGBoost. وجدنا أن بعض مؤشرات SMART مثل أخطاء القراءة، وعدد ساعات التشغيل، ودورات تحميل الرؤوس، والقطاعات المعلقة كانت من أكثر العوامل تأثيراً. وهذا منطقي من الناحية الهندسية لأن هذه المؤشرات تعكس إجهاد القرص واحتمالية تعطله.

روابط الملفات

رابط الداتا

https://drive.google.com/file/d/1FgpjrSlzLzPLv_OBgoFHlDyP5TOKSna3/view?usp=sharing

رابط المشروع المطبق على google colap

<https://colab.research.google.com/drive/1TlShwLzlZLFYdsXCKA6miBvCv0XW3yTP?usp=sharing>