**وصف المشروع: تحليل بيانات المباني وبناء نموذج تنبؤي للكفاءة الطاقية**

**1. مقدمة**

هذا المشروع يركز على **تحليل بيانات المباني المتعلقة بالاستهلاك الطاقي والمائي** وبناء **نموذج تنبؤي لتقدير ENERGY STAR Score** لكل مبنى. يعتمد المشروع على بيانات حقيقية لمباني مدينة نيويورك، ويطبق سلسلة من الخطوات العملية لمعالجة البيانات، استكشافها، تدريب نموذج قوي، وتفسير نتائجه.

**2. معالجة البيانات**

1. **تحميل البيانات:**
   * قراءة ملف CSV يحتوي على معلومات استهلاك الطاقة والمياه وبيانات المباني.
2. **تنظيف البيانات:**
   * تحويل القيم النصية "Not Available" إلى قيم مفقودة (NaN).
   * التعامل مع الأعمدة الرقمية:
     + تحويل النصوص الرقمية إلى أرقام.
     + ملء القيم المفقودة بالمتوسط.
     + معالجة القيم المتطرفة باستخدام طريقة IQR (Interquartile Range).
   * التعامل مع الأعمدة الزمنية: تحويلها إلى نوع بيانات التاريخ وملء القيم المفقودة بالوسيط.
   * التعامل مع الأعمدة الجغرافية: حذف الصفوف التي تفتقد لإحداثيات خطوط الطول والعرض.
   * التعامل مع الأعمدة النصية: ملء القيم المفقودة بـ "Unknown".
3. **استكشاف البيانات:**
   * رسم توزيع ENERGY STAR Score لفهم كيفية انتشار الدرجات بين المباني.

**3. تجهيز البيانات للنموذج**

1. **الميزات الرقمية:**
   * أخذ اللوغاريتم لبعض الأعمدة لتقليل تأثير القيم المتطرفة.
2. **الميزات الفئوية:**
   * تحويل الأعمدة الفئوية مثل "Borough" و"Largest Property Use Type" إلى تمثيل عددي باستخدام One-Hot Encoding.
3. **العمود الهدف:**
   * العمود الهدف هو ENERGY STAR Score.
   * إزالة العمود الهدف من مجموعة الميزات.
4. **تقسيم البيانات:**
   * تقسيم البيانات إلى مجموعة تدريب (70%) ومجموعة اختبار (30%).
5. **ملء القيم المفقودة وتطبيع البيانات:**
   * استخدام SimpleImputer لملء القيم المفقودة في الميزات.
   * استخدام MinMaxScaler لتطبيع البيانات بين 0 و1.

**4. بناء النموذج وتدريبه**

* **النموذج المستخدم:** GradientBoostingRegressor، أحد النماذج القوية لتقدير القيم العددية.
* **الإعدادات الرئيسية:**
  + loss='absolute\_error'
  + max\_depth=5
  + n\_estimators=500
  + معلمات أخرى لضبط العمق وعدد العينات لكل تقسيم.
* **التدريب:** النموذج يتعلم من بيانات التدريب بعد المعالجة والتطبيع.

**5. تقييم النموذج وأهمية الميزات**

* استخراج **أهمية الميزات** لمعرفة العوامل الأكثر تأثيرًا على التنبؤ.
* عرض **أهم 10 ميزات** بشكل بياني لتوضيح تأثير كل ميزة.
* تحليل العينة ذات أكبر خطأ في التنبؤ لتحديد نقاط ضعف النموذج.

**6. تفسير النموذج بصريًا**

1. **عرض شجرة فردية:**
   * استخراج شجرة واحدة من نموذج Gradient Boosting وتحويلها إلى صورة .png لفهم القرارات الداخلية.
2. **استخدام LIME:**
   * تفسير التنبؤات الفردية، خصوصًا العينة ذات أكبر خطأ.
   * عرض تأثير كل ميزة على التنبؤ بطريقة مرئية واضحة.

**7. الخلاصة**

* المشروع يغطي **المسار الكامل من معالجة البيانات → استكشافها → تدريب النموذج → تفسير النتائج**.
* النموذج:
  + **قوي وواقعي**، يعتمد على بيانات حقيقية.
  + يعطي معلومات عملية عن **العوامل الأكثر تأثيرًا على الكفاءة الطاقية للمباني**.
* الاستخدامات العملية:
  + مساعدة المهندسين والمخططين الحضريين في تقييم كفاءة المباني.
  + تقديم نتائج يمكن شرحها للمستخدمين النهائيين أو أصحاب القرار باستخدام أدوات تفسير مثل LIME.

💡 **ملاحظة:** يمكن تطوير المشروع أكثر باستخدام تقنيات إضافية مثل:

* تحسين معلمات النموذج (Hyperparameter Tuning).
* استخدام Cross-Validation لتقييم النموذج بدقة أكبر.
* مقارنة النموذج مع نماذج أخرى مثل Random Forest أو XGBoost.