Task 5

الخطوة الأولى: الحصول على البيانات

تم استخدام مكتبة pandas للحصول على البيانات من ملف الاكسل .

الخطوة الثانية: استكشاف وتصور البيانات

- تم استخدام "head" لعرض أول خمسة صفوف من DataFrame.
- "describe" لعرض إحصاءات ملخصة للبيانات مثل المتوسط والانحراف المعياري وأكبر وأصغر قيمة .
- "info" لعرض معلومات حول البيانات مثل عدد الصفوف والأعمدة وأنواع البيانات.

الخطوة الثالثة : استكشاف البيانات بشكل أعمق
من خلال إجراء تحليلات إحصائية
تم استخدام رسوم بيانية مختلفة مثل الرسوم البيانية الشريطية والرسوم البيانية الدائرية والرسوم البيانية القطعية لتصور العلاقات بين المتغيرات بشكل أكثر تفصيلاً.
استخدامنا مكتبة matplotlib ,seaborn لإنشاء هذه الرسوم البيانية.

الخطوة الرابعة: تحضير البيانات

- تنظيف البيانات:

تنظيف البيانات وتحضيرها للتحليل والتدريب تم فحص البيانات للعثور على القيم المفقودة باستخدام

isna()sum().

تم معالجة القيم المفقودة في العمود total_bedroomsباستخدام الخيار الثالث، والذي يتمثل في ملء القيم الناقصة بوسيطة العمود.

- تحليل البيانات:

  تم فحص البيانات للتأكد من عدم وجود قيم

  صفرية غير مقبولة في البيانات.

- تجميع السمات:

  تم إنشاء سمات جديدة تمثل علاقات بين السمات الموجودة

  مثل rooms_per_household و

  bedrooms_per_room و
  .population_per_household

- تدقيق الارتباط:

  تم فحص الارتباط بين السمات باستخدام مصفوفة الارتباط،

  حيث يمكن أن يوفر هذا الفحص فهمًا أفضل لعلاقات البيانات

  وتأثيرها على بعضها البعض.

- تحسين البيانات:

  تمت إزالة السمات القديمة

  total_bedrooms)

  (, total_rooms, population, households

  التي تم استبدالها بسمات جديدة.

- معالجة البيانات النصية :

تم تشفير المتغير الفئوي

ocean_proximity

باستخدام تقنية ترميز OneHotEncoder.

الخطوة الخامسة: تقسيم البيانات
تم استخدام دالة train_test_split من مكتبة scikit-learn لتقسيم مجموعة البيانات إلى مجموعة تدريب ومجموعة اختبار بنسبة محددة، في هذه الحالة تم استخدام نسبة 80% من البيانات للتدريب و20% للاختبار.