به نام خدا

«پروژه اول»

بوت‌کمپ هوش مصنوعی کوئرا

بهار ۱۴۰۴

مهلت ارسال پاسخ: تا ساعت ۲۳:۵۹ روز پنج‌شنبه ۱۵ خرداد

زمان ارائه‌ی گروهی: یکشنبه ۱۸ و دوشنبه ۱۹ خردادماه

بخش اول:معرفی داده

[جهت دریافت مجموعه‌داده بخش کلیک کنید.]

مجموعه داده‌ی **«آگهی‌های املاک دیوار»** شامل بیش از یک میلیون آگهی واقعی منتشر شده در پلتفرم دیوار است که اطلاعات جامع و متنوعی از املاک مسکونی و تجاری را در بازه‌های زمانی مختلف و در مناطق گوناگون کشور ثبت کرده‌است. این مجموعه داده شامل ویژگی‌هایی مانند موقعیت جغرافیایی (شهر، محله، طول و عرض جغرافیایی)، مشخصات فیزیکی ملک (متراژ، تعداد اتاق، سال ساخت، امکانات)، جزئیات مالی (مبلغ رهن، اجاره، قیمت کل)، نوع سند، نوع آگهی‌دهنده و سایر ویژگی‌های کاربردی می‌باشد. تنوع بالا در نوع املاک و پراکندگی جغرافیایی، این مجموعه داده را به یکی از منابع غنی برای تحلیل بازار مسکن ایران تبدیل کرده است.

بررسی این مجموعه داده از جهات مختلفی اهمیت دارد؛ نخست، می‌توان از آن برای تحلیل روندهای قیمتی در بازار املاک، پیش‌بینی قیمت‌ها، شناسایی مناطق گران یا ارزان و الگوهای عرضه و تقاضا استفاده کرد. همچنین می‌توان با تحلیل این داده‌ها، ترجیحات کاربران در انتخاب ملک (مانند داشتن آسانسور، پارکینگ یا متراژ مشخص) را دریافت و آگهی‌های دقیق‌تر و منطبق با نیاز کاربر را به وی نمایش داد. از سوی دیگر، این مجموعه داده یک منبع مهم برای پروژه‌های یادگیری ماشین و تحلیل مکانی محسوب می‌شود که می‌تواند به توسعه سامانه‌های هوشمند توصیه‌گر، سیستم‌های ارزش‌گذاری ملک و پلتفرم‌های تحلیل بازار کمک کند.

در ادامه لیستی از ویژگی‌های موجود در این مجموعه داده را بررسی و هر کدام را معرفی می‌کنیم:

نام ستون(انگلیسی)	نام ستون (فارسی)	توضیح
cat2_slug	دسته‌بندی سطح ۲	زیرمجموعه‌ای از دسته‌بندی کلی آگهی
cat3_slug	دسته‌بندی سطح ۳	دسته‌بندی جزئی‌تر
city_slug	شهر	نام شهر محل ملک

neighborhood_slug	محله	نام محله یا منطقه‌ای از شهر
created_at_month	تاریخ ایجاد (ماه)	ماهی که آگهی ثبت شده است
user_type	نوع کاربر	نوع آگهی‌دهنده (شخص، مشاور، مالک)
description	توضیحات	توضیحاتی که کاربر نوشته است
title	عنوان	عنوان آگهی
rent_mode	حالت اجاره	نوع اجاره (روزانه، ماهانه)
rent_value	مبلغ اجاره	مبلغ اجاره اعلام‌شده
rent_to_single	اجاره به مجرد	اجاره به افراد مجرد یا نه
rent_type	نوع اجاره	نوع قرارداد اجاره
price_mode	نوع قیمت	نحوه تعیین قیمت
price_value	مبلغ قیمت	مبلغ کل قیمت
credit_mode	نوع رهن	نحوه تعیین رهن
credit_value	مبلغ رهن	مبلغ رهن اعلام‌شده
rent_credit_transform	تبدیل اجاره و رهن	ترکیب اجاره و رهن به عدد قابل مقایسه
transformable_price	قیمت قابل تبدیل	قیمت تبدیل‌شده برای مدل‌سازی
transformable_credit	رهن قابل تبدیل	رهن تبدیل‌شده برای مدل‌سازی
transformed_credit	رهن نرمال‌شده	رهن پس از اعمال تبدیل
transformable_rent	اجاره قابل تبدیل	اجاره تبدیل‌شده برای مدل‌سازی
transformed_rent	اجاره نرمال‌شده	اجاره پس از تبدیل
land_size	متراژ زمین	مساحت زمین ملک
building_size	زیربنا	مساحت بنای ساخته‌شده
deed_type	نوع سند	نوع سند ملک
has_business_deed	دارای سند تجاری	آیا سند تجاری است
floor	طبقه	طبقه ملک

rooms_count	تعداد اتاق	تعداد اتاق خواب‌ها
total_floors_count	تعداد کل طبقات	کل طبقات ساختمان
unit_per_floor	واحد در هر طبقه	تعداد واحدهای هر طبقه
has_balcony	دارای بالکن	آیا ملک بالکن دارد
has_elevator	دارای آسانسور	آیا ملک آسانسور دارد
has_warehouse	دارای انباری	آیا ملک انباری دارد
has_parking	دارای پارکینگ	آیا ملک پارکینگ دارد
construction_year	سال ساخت	سال ساخت ملک
is_rebuilt	بازسازی شده	آیا ملک بازسازی شده است
has_water	دارای آب	آیا ملک آب دارد
has_warm_water_provider	دارای آبگرم	آیا سیستم آب گرم دارد
has_electricity	دارای برق	آیا ملک برق دارد
has_gas	دارای گاز	آیا ملک گاز دارد
has_heating_system	سیستم گرمایشی	آیا سیستم گرمایشی دارد
has_cooling_system	سیستم سرمایشی	آیا سیستم سرمایشی دارد
has_restroom	دارای سرویس بهداشتی	آیا سرویس بهداشتی دارد
has_security_guard	نگهبان	آیا ملک نگهبان دارد
has_barbecue	باربیکیو	آیا باربیکیو دارد
building_direction	جهت ملک	جهت جغرافیایی ملک
has_pool	استخر	آیا استخر دارد
has_jacuzzi	جکوزی	آیا جکوزی دارد
has_sauna	سونا	آیا سونا دارد
floor_material	جنس کف‌پوش	نوع متریال کف
property_type	نوع ملک	نوع ملک (آپارتمان، ویلا،...)

regular_person_capacity	ظرفیت نفرات عادی	ظرفیت عادی نفرات
extra_person_capacity	ظرفیت نفرات اضافی	ظرفیت افراد اضافه
cost_per_extra_person	هزینه هر نفر اضافه	هزینه اضافی هر نفر اضافه
rent_price_on_regular_days	اجاره در روزهای عادی	مبلغ اجاره در روزهای عادی
rent_price_on_special_days	اجاره در روزهای خاص	مبلغ اجاره در ایام خاص
rent_price_at_weekends	اجاره در آخر هفته	مبلغ اجاره در آخر هفته
location_latitude	عرض جغرافیایی	عرض جغرافیایی موقعیت ملک
location_longitude	طول جغرافیایی	طول جغرافیایی موقعیت ملک
location_radius	شعاع مکان	شعاع مکانی ملک

توجه کنید که در آگهی‌های املاک، قیمت‌ها بسته به نوع قرارداد (فروش، رهن، اجاره یا ترکیبی از رهن و اجاره) به شیوه‌های متفاوتی اعلام می‌شوند؛ به همین دلیل، مقایسه مستقیم آن‌ها ممکن نیست. برای رفع این مسئله، از مفهوم "قیمت تبدیل‌شده"(**transformable_price**) استفاده می‌شود. در این فرآیند، با استفاده از یک نسبت تبدیل مشخص (مثلاً تبدیل هر یک میلیون تومان رهن به معادل سی هزار تومان اجاره)، مبالغ رهن و اجاره به یک واحد مالی مشترک تبدیل می‌شوند تا بتوان آن‌ها را با یکدیگر مقایسه یا در مدل‌سازی‌های تحلیلی و یادگیری ماشین استفاده کرد. در مقابل، "قیمت اصلی" همان مبلغ خام و ثبت‌شده توسط کاربر در آگهی است که بدون هیچ‌گونه نرمال‌سازی یا تبدیل عددی درج شده است. بنابراین به طور کلی قیمت یک ملک را در صورتی که برای فروش گذاشته شده قیمت تبدیل شده و در صورتی که برای اجاره قرار داده شده رهن تبدیل شده در نظر گرفت.

بخش دوم: تحلیل‌های آماری

در این بخش، به کمک دانش آماری می‌خواهیم به تعدادی از سوال‌ها پاسخ دهیم؛ این سوالات به منظور درک و یافتن شهود از مجموعه داده و نیز بررسی بعضی از فرضیه‌های رایج بازار مسکن پرسیده شده است.

آمار توصیفی

1. توزیع آگهی‌های موجود در دسته‌های مختلف را برای دسته‌بندی سطح دو و سطح سه رسم کنید.
2. هیستوگرام سال ساخت را رسم کنید.
3. تعداد آگهی‌های منتشر شده در ماه‌های مختلف را برای فروش و اجاره بررسی کنید. آیا تعداد آگهی‌های فروش و اجاره در زمان‌های مشخصی از سال افزایش چشم‌گیری داشته است؟
4. توزیع قیمت فروش(price_value) را برای دسته‌بندی‌های سطح سه در یک نمودار رسم کنید.
5. بر روی نقشه‌ی جغرافیایی heatmap آگهی‌های هر منطقه را رسم کنید. تراکم آگهی‌ها کدام منطقه بیشتر است؟
6. ترند میانگین قیمت اجاره بر حسب ماه‌های قرار گرفتن آگهی‌ها رسم کنید.(دقت کنید که ماه‌ها باید به تاریخ شمسی و خوانا باشند.)
7. در طول زمان قیمت‌های اسمی افزایش پیدا می‌کنند اما این افزایش لزوما به معنی بالارفتن ارزش واقعی ملک نیست و می‌تواند ناشی از تورم باشد. به ازای میانگین مبلغ قیمت(price_value) در سال‌های ۱۴۰۰ تا ۱۴۰۳ قیمت حقیقی را محاسبه کنید و بررسی کنید ترند قیمت حقیقی چگونه است.
8. ماتریس هم‌بستگی را برای مبلغ قیمت، متراژ زمین، زیربنا، ظرفیت نفرات، تعداد اتاق‌ها و طول و عرض جغرافیایی رسم نمایید.
9. می‌خواهیم بررسی کنیم خانه‌هایی که دارای بالکن، آسانسور، نگهبان، باربیکیو و استخر هستند عمدتا در کدام مناطق قرار دارند. با نمودار مناسب این موضوع را نشان دهید.

آزمون فرض

- با توجه به رشد مهاجرت افراد از شهرهای کوچکتر به کلان‌شهرها و تراکم جمعیت در این نواحی، تصور می‌شود که میانگین مساحت خانه‌های مسکونی در کلان‌شهرها نسبت به شهرهای کوچک و روستاها کمتر است. آیا مجموعه داده این فرضیه را پشتیبانی می‌کند؟(برای دسته‌بندی شهرها به کلان‌شهر و شهر کوچک می‌توانید از این مجموعه داده استفاده کنید.)
- معمولا این جمله را می‌شنویم که «قدیما خونه‌ها دلبازتر بود!» برای بررسی این فرضیه، آیا میانگین مساحت خانه‌های قدیمی‌ساخت نسبت به خانه‌های جدید ساخت بیشتر است؟(خانه‌های قدیمی‌ساخت را خانه‌هایی در نظر بگیرید که قبل از سال ۹۶ ساخته شده‌اند.)
- داشتن سند تجاری (یا هر نوع سند ملکی) در املاک به این معنی است که سند مالکیت معتبر، رسمی و قانونی برای ملک تجاری دارید. این سند نشان می‌دهد که شما صاحب قانونی ملک تجاری هستید و می‌توانید از حقوق مالکیت آن استفاده کنید. بررسی کنید که آیا داشتن سند تجاری(has_business_deed) بر میانگین قیمت فروش ملک تجاری تاثیر معناداری دارد؟
- در دسته‌بندی امکانات موجود در آگهی‌ها می‌توانیم آنها را به دو دسته‌ی امکانات لاکچری(استخر، باربیکیو، سونا، جکوزی) و امکانات غیر لاکچری تقسیم کنیم. فرضیه‌ی ما این است که میانگین مبلغ قیمت برای وجود ویژگی‌های لاکچری افزایش چشم‌گیری دارد. اما آیا این میانگین برای وجود امکانات غیرلاکچری نیز تفاوت معناداری دارد؟

بخش سوم:یادگیری ماشین

مسئله‌ی ۱: ساخت سیستم توصیه‌گر

خوشه‌بندی املاک موجود در مجموعه‌داده برای توسعه‌ی سیستم‌های توصیه‌گر ملک یکی از اساسی‌ترین استفاده‌های مجموعه‌داده‌ی در اختیار قرار داده شده است. به همین منظور ابتدا با ارائه‌ی تحلیل مناسب یک مجموعه از ویژگی‌های مهم برای تعیین خوشه‌ها را انتخاب کنید. دقت کنید که در یک سیستم توصیه‌گر خوشه‌ها بر مبنای سلیقه و نیاز کاربر ساخته می‌شوند بنابراین انتخاب ویژگی‌های کلیدی برای ساخت خوشه‌های معنادار به شدت حائز اهمیت است.(دقت کنید که ویژگی‌های انتخابی نباید آنقدر زیاد باشند که دچار نفرین ابعاد شویم.)

بخش ۱)

حال الگوریتم خوشه‌بندی K-means را تنها بر حسب با **۱۰ خوشه** برای این مجموعه‌داده اجرا کنید. سپس ابتدا مختصات جغرافیایی را به فرمت UTM درآورید و برای دو ویژگی قیمت و مختصات UTM یک اسکترپلات رسم کنید. بر روی اسکتر پلات رسم‌شده مشخص کنید کدام نقاط مربوط به کدام خوشه هستند و مرکز هر خوشه را نیز رسم کنید. به انتخاب رنگ، مارکر، نام‌گذاری محورها و به‌طور کلی قابل درک بودن تصویر دقت داشته باشید.

بخش ۲)

پس از آن الگوریتم K-means را برای k هایی از ۱ تا ۲۰ اجرا کرده و با محاسبه‌ی *مجموع مجذورات درون خوشه‌ای (Within-Cluster Sum of Square)*، مقدار مناسبی برای هایپرپارامتر k انتخاب کنید. توجه کنید که بخش زیادی از نمره‌ی این بخش مربوط به نحوه‌ی انتخاب مقدار k است و چنان‌چه روش‌های تدریس‌شده و معمول پاسخگوی حل مسئله نبود،انتظار می‌رود با جستجو و مطالعه‌ی بیشتر، روشی مناسب برای رفع چالش‌های احتمالی پیشنهاد دهید.

بخش ۳)

در آخرین گام از این سوال از شما می‌خواهیم که ابتدا دو ویژگی مختصات UTM و قیمت قابل تبدیل را در نظر بگیرید و سپس با استفاده از روش DBScan داده‌ها را فقط با در نظر گرفتن این دو ویژگی خوشه‌بندی کنید و هایپرپارامترها را به‌نحوی تغییر دهید که ۳ کلاستر بامعنا در خروجی تولید شود. اسکتر پلات داده‌ها

و نحوه‌ی خوشه‌بندی آن‌ها را مطابق بخش ۱ رسم کنید. نحوه‌ی اثرگذاری هر یک از هایپرپارامترها بر خروجی را توضیح دهید.

مسئله‌ی ۲: پیش‌بینی

پیش‌بینی قیمت ملک یکی دیگر از کاربردهای مجموعه‌داده‌ی این پروژه است. چالش اصلی در پیش‌بینی قیمت توجه به حالت‌های مختلف معامله‌ی املاک از جمله رهن اجاره و فروش است که مدل باید به درستی مبلغ را برای حالت عرضه‌ی ملک پیش‌بینی کند.

در این بخش از شما می‌خواهیم مدلی آموزش دهید که با توجه به اطلاعات دریافتی از مشخصات ملک و نوع معامله پیش‌بینی کند قیمت اجاره یا فروش ملک مربوطه چقدر است؟

شما مجاز هستید از هر کدام از الگوریتم‌های یادگیری ماشین که تاکنون در کلاس‌های بوت‌کمپ آموخته‌اید برای مدل‌سازی استفاده کنید.

> **توجه:** استفاده از الگوریتمی غیر از الگوریتم‌های اصلی‌ای که در کلاس‌ها آموزش داده شده‌اند در بخش اصلی مجاز نیست. در صورت علاقه و تسلط می‌توانید از آن‌ها برای بخش امتیازی استفاده کنید. البته توجه داشته باشید که نیاز است تمام اعضای گروه نسبت به نحوه‌ی کار آن الگوریتم دانش کافی داشته باشند.

در صورت نیاز می‌توانید هر ویژگی دلخواهی را به مجموعه‌داده اضافه کنید یا آن‌ها را مهندسی کنید. البته دقت کنید که **ویژگی‌های ورودی مدل منجر به نشت متغیر هدف نشود.**

با مقایسه‌ی پیش‌بینی مدل خود با مقادیر حقیقی برای داده های تست معیار ارزیابی r2_score و MAE و MSE را گزارش کنید. نیاز است در زمان ارائه تحلیل مناسبی از نتایج به‌دست‌آمده ارائه دهید.

> **توجه:** در آزمایش‌های خود و انتخاب مدل و هایپرپارامترهای آن نباید از داده‌های آزمون (Test) استفاده کنید، بلکه این کار باید با داده‌های اعتبارسنجی (Validation) انجام گیرد. تنها پس از دست‌یابی به مدل نهایی خود از مجموعه‌ی آزمون بهره ببرید.

نکته‌های کلی

- کدهای خود را خوانا و تمیز بنویسید.
- مهم‌ترین بخش این پروژه، تحلیل و تفسیر شما از شرایط مسئله و نتایج آن است. باید بتوانید برای هر کدام از انتخاب‌های خود در طول مسیر، دلیلی موجه و علمی داشته باشید. ارائه‌ی شما نیز باید بر همین محور باشد، یعنی روند حل مسئله، نتایج و تحلیل و تفسیر را ارائه دهید، نه توضیح کد.
- به نکات ذکر شده در ارتباط با نحوه‌ی ارسال فایل در صفحه‌ی پروژه در کلاس توجه فرمایید.

بخش امتیازی (بیشینه: ۲۵ نمره)

- مستندسازی غنی و مناسب در نت‌بوک‌ها (۲ نمره)
- استفاده از گیت و مشارکت فعال در آن (۲ نمره)
- استخراج و اضافه کردن ویژگی‌های مناسب و بامعنا (۲ نمره)
- استفاده از مدل‌های حرفه‌ای‌تر و دست‌یابی به نتایج بهتر با تسلط کامل اعضای گروه به الگوریتم (۷ نمره)
- طرح مسئله‌ای جدید با توجه به داده‌های موجود و مرتبط (با تایید منتور) و دست‌یابی به نتایج قابل قبول و تفسیرپذیر (۱۰ نمره)`````
- ارائه‌ای جذاب با بهره‌گیری از خط داستانی و استفاده از ابزارهای مناسب ارائه همچون اسلاید (۲ نمره)

موفق باشید 🤩