

دانشگاه بوقعلی سینا

پروژه هوش مصنوعی



پیش بینی قیمت مسکن در تهران با الگوریتم های یادگیری ماشین



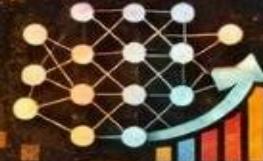
وب کاوی



پردازش داده



مهندسی ویژگی



آموزش مدل



ویژگی های خانه

ورودی مدل ، یک خانه

خروجی مدل تخمینی خانه



پروژه درس هوش مصنوعی

پیش‌بینی قیمت مسکن از آگهی‌های دیوار

دانشگاه بوعلی سینا همدان

دکتر دزفولیان

پدیدآورندگان: محمد متین پریان، امین شهابی، سعید مظاہری،

یاسمن صفریان، سارا راشدی، آریا نادری

۱۴۰۱ ترم

هدف پروژه

هدف این پروژه آشنایی عملی با یک مسئله‌ی واقعی یادگیری ماشین است که شامل جمع‌آوری داده از وب (Web Crawling)، پیش‌پردازش داده، مهندسی ویژگی (Feature Engineering)، آموزش مدل و تحلیل نتایج می‌باشد.

در این پروژه شما با چالش‌های دنیای واقعی داده‌ها (نویز، داده‌های ناقص، متن آزاد و ...) مواجه می‌شوید و یاد می‌گیرید چگونه یک مسئله‌ی خام را به یک مسئله‌ی قابل حل با مدل‌های هوش مصنوعی تبدیل کنید.

صورت مسئله

در این پروژه باید با استفاده از آگهی‌های فروش مسکن در وبسایت دیوار، قیمت یک واحد مسکونی در یک منطقه‌ی مشخص از تهران را بر اساس ویژگی‌های آن پیش‌بینی کنید.

ورودی مدل: ویژگی‌های یک خانه

خروجی مدل: قیمت (یا قیمت متری) تخمینی خانه

محدودیت دامنه

برای کنترل پیچیدگی پروژه و عملکرد بهتر مدل:

- فقط یک منطقه مشخص از تهران به دلخواه انتخاب شود (مثلاً: صادقیه، پونک، تهرانپارس، نارمک، ...)
- فقط آگهی‌های فروش آپارتمان مسکونی مجاز هستند
- حداقل تعداد آگهی مورد نیاز: ۳۰۰ آگهی



جمعآوری داده (Data Collection)

در این مرحله باید با استفاده از یکی از روش‌های زیر داده‌ها را جمعآوری کنید:

- استفاده با پایتون Web scrabing
- استفاده از API‌های رسمی/غیررسمی (در صورت وجود)

پیش‌پردازش داده (Data Cleaning)

در این مرحله انتظار می‌رود:

- حذف داده‌های ناقص یا غیرمنطقی
- حذف آگهی‌های تکراری
- شناسایی و حذف outlierهای شدید (مثلاً قیمت‌های غیرواقعی)
- یکسان‌سازی واحدها (مثلاً تومان / متر مربع)

پیشنهاد می‌شود در رابطه با پیش‌پردازش اصولی چنین پژوهه‌هایی تحقیق و جست و جو کنید.

مهندسی ویژگی (Feature Engineering)

دیتاست شما شامل دو بخش است. بخش اول ویژگی هایی که مستقیماً در آگهی ذکر شده است و بخش دوم ویژگی هایی که از قسمت توضیحات و قسمت جزئیات آگهی استخراج می‌کنید.

ویژگی‌های ساختاریافته (بخشی که به طور مستقیم در آگهی ذکر شده است)
این ویژگی‌ها باید به طور مستقیم از آگهی استخراج شوند:

نوع	ویژگی
عددی	متراز
عددی	تعداد اتاق خواب
عددی	سن بنا
عددی	طبقه
categorical	منطقه
عددی	قیمت کل
عددی	قیمت متری



ویژگی‌های استخراج شده از توضیحات (description) یا بخش جزئیات

به دلیل ساختار آزاد متن توضیحات (description)، استفاده از NLP آزاد مجاز نیست. دقت داشته باشید باید فقط از لیست از پیش‌تعریف شده زیر برای ویژگی‌ها استفاده کنید و آن‌ها را به صورت باینری (۰/۱) استخراج نمایید.

لیست ویژگی‌های مجاز:

توضیح	ویژگی
has_elevator	وجود آسانسور
has_parking	پارکینگ
has_storage	انباری
is_new	نوساز یا کلیدنخورده
is_renovated	بازسازی شده
has_balcony	بالکن
good_light	نورگیر
luxury_keywords	وجود کلمات لوکس (VIP، لاکچری، ...)
has_doc	وجود سند

هر ویژگی باید با استفاده از کلیدواژه‌ها و قوانین ساده (Rule-Based) از توضیحات استخراج شود.

اگر مایل به اضافه کردن ویژگی جدید بودید در گزارش کار خود ذکر کنید.

(Modeling) مدل‌سازی

باید حداقل سه مدل مختلف را پیاده‌سازی و مقایسه کنید.

خروجی می‌تواند یکی از حالت‌های زیر باشد:

- پیش‌بینی قیمت کل
- پیش‌بینی قیمت متری

(Evaluation) ارزیابی مدل

حداقل یکی از معیارهای زیر باید گزارش شود:

- MAE
- RMSE
- R^2

همچنین:

- مقایسه‌ی مدل‌ها الزامی است
- تحلیل Feature Importance امتیاز مثبت دارد

تحلیل و گزارش نهایی

گزارش نهایی باید شامل بخش‌های زیر باشد:

1. شرح مختصر مسئله
2. روش جمعآوری داده
3. مراحل پیش‌پردازش
4. توضیح ویژگی‌ها
5. مدل‌ها و تنظیمات
6. نتایج و مقایسه
7. تحلیل خطاهای و محدودیت‌ها
8. دیتاست استفاده شده

بخش امتیازی (Bonus)

- Visualization جامع
- بر اساس خلاقیت خودتون و تایید منتور ها

