گزارش پروژه شماره ۲

پیش‌بینی قیمت مسکن بوستون

نام ونام خانوادگی: فرزانه اکبرنژاد

عنوان پروژه: پیش‌بینی قیمت مسکن بوستون

زبان برنامه‌نویسی: Python

۱. مقدمه

در این پروژه می‌خواستم با استفاده از اطلاعات مربوط به خانه‌ها و محله‌های بوستون، قیمت مسکن را پیش‌بینی کنم.

برای انجام این کار از مجموعه داده‌ی Boston Housing Dataset استفاده کردم. این مجموعه داده شامل 506 رکورد و اطلاعات مختلفی درباره‌ی خانه‌ها و محله‌های بوستون هست.

اول داده‌ها را وارد کردم و اطلاعات آن‌ها را بررسی کردم. بعد داده‌های گمشده را بررسی کردم و برای اینکه اطلاعات را بهتر متوجه شوم، چند نمودار رسم کردم.

بعد از آن داده‌ها را به دو قسمت آموزش و تست تقسیم کردم و با استفاده از مدل Decision Tree Regressor مدل پیش‌بینی قیمت مسکن را ساختم.

در آخر هم میزان خطای مدل را با معیار MAE بررسی کردم و مدل آموزش‌دیده را ذخیره کردم.

۲. معرفی داده‌ها

داده‌هایی که در این پروژه استفاده کردم مربوط به خانه‌ها و محله‌های شهر بوستون هستند.

این مجموعه داده شامل 13 ویژگی مختلف است که اطلاعاتی درباره‌ی شرایط خانه‌ها و محله‌ها می‌دهند. یک ستون دیگر به نام MEDV هم وجود دارد که قیمت متوسط خانه را مشخص می‌کند و در این پروژه به عنوان مقدار هدف استفاده شده است.

ویژگی‌های موجود در داده:

CRIM: میزان جرم و جنایت در منطقه

ZN: درصد زمین‌های مسکونی با قطعات بزرگ

INDUS: درصد زمین‌های غیرمسکونی مربوط به صنعت

CHAS: مشخص می‌کند منطقه در کنار رودخانه قرار دارد یا نه

NOX: میزان غلظت اکسید نیتریک

RM: میانگین تعداد اتاق‌ها

AGE: درصد خانه‌های قدیمی

DIS: فاصله از مراکز اشتغال

RAD: میزان دسترسی به بزرگراه‌ها

TAX: میزان مالیات مربوط به ملک

PTRATIO: نسبت دانش‌آموز به معلم

B: یک شاخص مربوط به جمعیت

LSTAT: درصد جمعیت با وضعیت اجتماعی و اقتصادی پایین‌تر

MEDV: قیمت متوسط خانه‌ها

مقدار MEDV بر حسب هزار دلار در نظر گرفته شده است.

۳. کتابخانه‌های مورد استفاده

در ابتدای برنامه کتابخانه‌هایی که برای انجام پروژه لازم داشتم را وارد کردم.

از pandas برای خواندن و کار کردن با داده‌ها استفاده کردم.

از numpy برای کار با داده‌ها و محاسبات استفاده کردم.

از matplotlib و seaborn برای رسم نمودارها استفاده کردم.

از train_test_split برای تقسیم داده‌ها به قسمت آموزش و تست استفاده کردم.

برای ساخت مدل پیش‌بینی از DecisionTreeRegressor استفاده کردم.

برای محاسبه‌ی خطای مدل از mean_absolute_error استفاده کردم.

از joblib هم برای ذخیره کردن مدل استفاده کردم.

۴. وارد کردن داده‌ها

در این قسمت فایل housing.data را با استفاده از pandas وارد برنامه کردم.

برای خواندن فایل از read_csv استفاده کردم و اسم ستون‌های داده را هم مشخص کردم.

بعد از اینکه داده‌ها وارد شدند، تعداد رکوردهای موجود را بررسی کردم.

تعداد کل رکوردهای این مجموعه داده 506 است.

۵. بررسی اولیه داده‌ها

بعد از وارد کردن داده‌ها، آن‌ها را بررسی کردم.

اول با استفاده از head() پنج سطر اول داده‌ها را نمایش دادم تا ببینم اطلاعات به چه شکلی هستند.

بعد با استفاده از info() اطلاعات کلی داده‌ها را بررسی کردم. با این دستور می‌توان اطلاعاتی مثل تعداد داده‌ها و نوع ستون‌ها را مشاهده کرد.

بعد از آن از describe() استفاده کردم تا آمار توصیفی داده‌ها را ببینم. این قسمت اطلاعاتی مثل میانگین، کمترین مقدار و بیشترین مقدار را نشان می‌دهد.

۶. بررسی داده‌های گمشده

در مرحله بعد بررسی کردم که داده‌ی گمشده‌ای در مجموعه داده وجود دارد یا نه.

برای این کار از isnull().sum() استفاده کردم. نتیجه نشان داد که در این مجموعه داده داده‌ی گمشده‌ای وجود ندارد، پس نیازی به حذف یا جایگزین کردن داده‌های گمشده نداشتم.

۷. رسم نمودارها

برای اینکه بتوانم داده‌ها را بهتر بررسی کنم، سه نمودار رسم کردم.

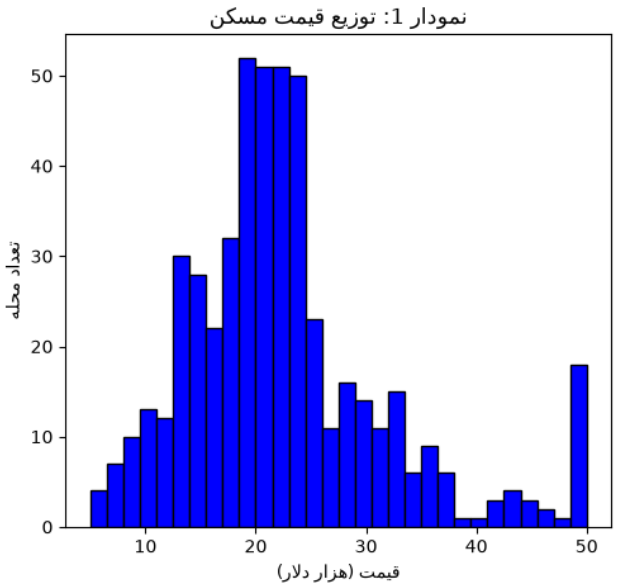
این سه نمودار موارد زیر هستند:

نمودار اول: توزیع قیمت مسکن

در نمودار اول از هیستوگرام استفاده کردم.

در محور افقی قیمت مسکن و در محور عمودی تعداد محله‌ها قرار دارد.

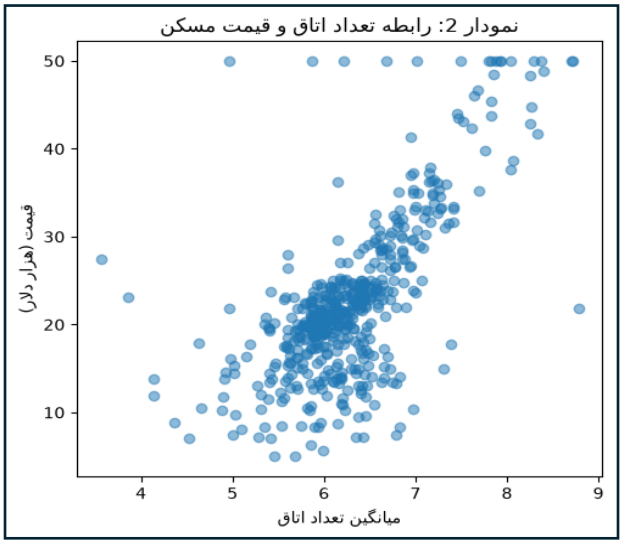
با این نمودار می‌توان توزیع قیمت خانه‌ها را بهتر مشاهده کرد.



نمودار دوم: رابطه تعداد اتاق و قیمت مسکن

در نمودار دوم از نمودار پراکندگی استفاده کردم.

در این نمودار RM یعنی میانگین تعداد اتاق‌ها در محور افقی و MEDV یعنی قیمت مسکن در محور عمودی قرار دارد.هدف از این نمودار بررسی رابطه بین تعداد اتاق‌ها و قیمت مسکن بود.

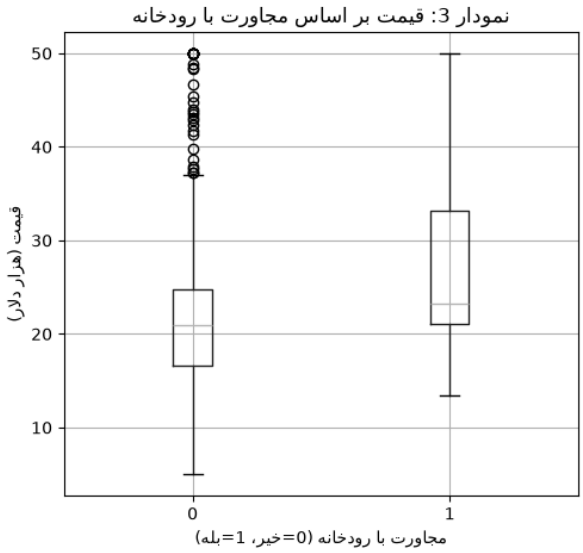


نمودار سوم: قیمت مسکن بر اساس مجاورت با رودخانه

در نمودار سوم از Boxplot استفاده کردم.

در این نمودار قیمت مسکن را بر اساس CHAS مقایسه کردم.

مقدار صفر یعنی محله در کنار رودخانه نیست و مقدار یک یعنی محله در کنار رودخانه قرار دارد.



۸. آماده کردن داده‌ها برای مدل

در این مرحله ویژگی‌ها را از مقدار هدف جدا کردم.

در متغیر x ویژگی‌های مربوط به خانه‌ها و محله‌ها قرار گرفتند.

در متغیر y هم ستون MEDV قرار گرفت که قیمت مسکن را مشخص می‌کند.

به این صورت مدل می‌تواند با استفاده از ویژگی‌های مختلف، قیمت مسکن را پیش‌بینی کند.

۹. تقسیم داده‌ها به آموزش و تست

برای اینکه بتوانم مدل را هم آموزش بدهم و هم عملکرد آن را بررسی کنم، داده‌ها را به دو قسمت تقسیم کردم.

80 درصد داده‌ها برای آموزش و 20 درصد برای تست استفاده شدند.

برای این کار از train_test_split استفاده کردم.

همچنین مقدار random_state را 42 قرار دادم تا در صورت اجرای دوباره برنامه، تقسیم داده‌ها به شکل یکسان انجام شود.

در اجرای برنامه :

تعداد داده‌های آموزش: 404

تعداد داده‌های تست: 102

۱۰. ساخت مدل درخت تصمیم

برای پیش‌بینی قیمت مسکن از مدل Decision Tree Regressor استفاده کردم.

درخت تصمیم یکی از روش‌هایی است که با استفاده از ویژگی‌های داده، مقدار موردنظر را پیش‌بینی می‌کند.

در این پروژه مقدار max_depth را برابر 8 قرار دادم تا عمق درخت مشخص باشد.

بعد از ساخت مدل، آن را با استفاده از داده‌های آموزش، آموزش دادم.

مدل با موفقیت آموزش داده شد.

۱۱. پیش‌بینی قیمت مسکن

بعد از آموزش مدل، از داده‌های تست برای پیش‌بینی استفاده کردم.

مدل با توجه به ویژگی‌های موجود، قیمت مسکن مربوط به داده‌های تست را پیش‌بینی کرد.

نتیجه‌ی این پیش‌بینی‌ها در متغیر y_pred قرار گرفت.

بعد از آن مقدار پیش‌بینی‌شده را با مقدار واقعی مقایسه کردم تا میزان خطای مدل را به دست بیاورم.

۱۲. بررسی خطای مدل با MAE

برای بررسی عملکرد مدل از معیار MAE یا Mean Absolute Error استفاده کردم.

MAE به ما نشان می‌دهد که پیش‌بینی‌های مدل به طور متوسط چقدر با مقدار واقعی تفاوت دارند.

هرچه مقدار MAE کمتر باشد، خطای مدل کمتر است.

نتیجه‌ای که از اجرای برنامه گرفتم:

MAE = 2.28 هزار دلار

طبق خواسته‌ی پروژه، مقدار MAE باید کمتر از 4.0 هزار دلار باشد.

مقدار به‌دست‌آمده یعنی 2.28 کمتر از 4 است، بنابراین مدل توانست این قسمت از پروژه را با موفقیت انجام دهد.

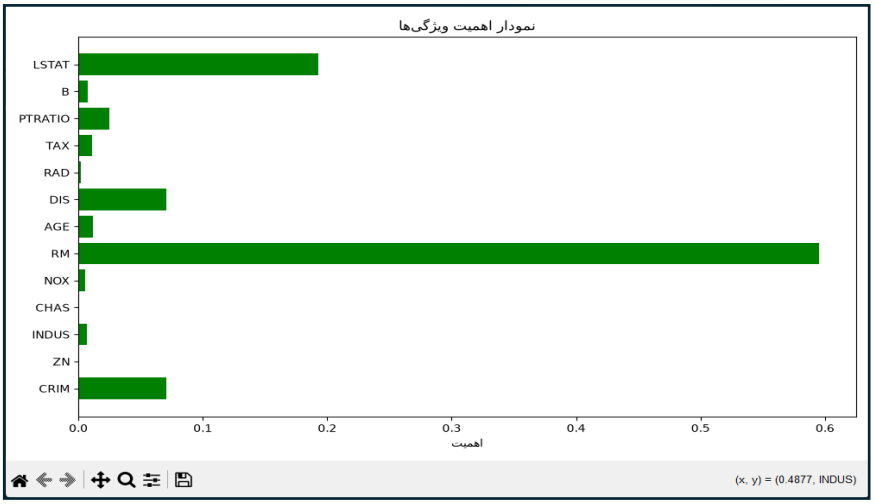
۱۳. بررسی اهمیت ویژگی‌ها

بعد از آموزش مدل، اهمیت ویژگی‌های مختلف را هم بررسی کردم.

برای این کار از feature_importances_ استفاده کردم.

این قسمت نشان می‌دهد که هر ویژگی در تصمیم‌گیری مدل چه مقدار اهمیت داشته است.

برای اینکه مقایسه‌ی ویژگی‌ها راحت‌تر باشد، یک نمودار میله‌ای هم برای اهمیت ویژگی‌ها رسم کردم.



۱۴. مهم‌ترین ویژگی

بعد از بررسی اهمیت ویژگی‌ها، مشخص شد که ویژگی RM بیشترین اهمیت را در مدل دارد.

مقدار اهمیت RM در اجرای من برابر بود با: 0.5953

RM نشان‌دهنده‌ی میانگین تعداد اتاق‌ها است.

بنابراین در مدل من، تعداد اتاق‌ها بیشترین اهمیت را بین ویژگی‌های استفاده‌شده برای پیش‌بینی قیمت مسکن داشته است.

۱۵. ذخیره کردن مدل

بعد از اینکه مدل آموزش داده شد و نتیجه‌ی آن را بررسی کردم، مدل را با استفاده از joblib ذخیره کردم.

نام فایل مدل ذخیره‌شده:

مدل_مسکن_بوستون.joblib است

با ذخیره کردن مدل، می‌توان در آینده از مدل آموزش‌دیده استفاده کرد و لازم نیست هر بار مدل را از اول آموزش داد.

۱۶. فایل‌های تولیدشده:

در پایان اجرای برنامه، فایل‌های زیر ایجاد شدند:

boston_housing.py (کد اصلی)

نمودارها.png (شامل سه نمودار پروژه)

اهمیت_ویژگی‌ها.png (نمودار اهمیت ویژگی‌ها)

مدل_مسکن_بوستون.joblib (مدل آموزش‌دیده و ذخیره‌شده)

۱۷. نتیجه‌گیری

در این پروژه سعی کردم با استفاده از اطلاعات مربوط به خانه‌ها و محله‌های بوستون، قیمت مسکن را پیش‌بینی کنم.

اول داده‌ها را وارد کردم و اطلاعات کلی آن‌ها را بررسی کردم. همچنین بررسی کردم که آیا داده‌ی گمشده‌ای وجود دارد یا نه و مشخص شد که داده‌ی گمشده‌ای وجود ندارد.

بعد برای بررسی بهتر داده‌ها سه نمودار رسم کردم. سپس داده‌ها را به دو قسمت آموزش و تست تقسیم کردم و با استفاده از Decision Tree Regressor مدل پیش‌بینی را ساختم.

برای بررسی عملکرد مدل از معیار MAE استفاده کردم و مقدار خطای مدل من 2.28 هزار دلار شد. این مقدار کمتر از 4 هزار دلار بود، بنابراین نتیجه‌ی مدل در محدوده‌ی موردنظر پروژه قرار گرفت.

همچنین اهمیت ویژگی‌ها را بررسی کردم و مشخص شد که RM با مقدار اهمیت 0.5953 مهم‌ترین ویژگی در مدل من بوده است.

در آخر هم مدل آموزش‌دیده را با استفاده از joblib ذخیره کردم.

نتیجه نهایی:

مدل با موفقیت آموزش داده شد و توانست قیمت مسکن را با MAE برابر با 2.28 هزار دلار پیش‌بینی کند.