۱) فهم مسئله:

۱. **انگیزه اصلی چنین پروژهای چیست؟** اهداف کسب و کار انگیزه اصلی پروژه است. اهدافی مانند این که ویژگی های مشترک مشتریانی که اخیرا از دست داده ایم و از خدمات و محصولات شرکتهای رقیب استفاده می کنند چیست ؟ و یا هریک از مشتریان شرکت دارای چه ارزشی برای ما هستند.

۲. **خروجی چنین پروژهای برای چه مواردی ممکن است کاربرد داشته باشد؟** ۱- اینکه یک مقاله خبری به چه دسته ای تعلق دارد۲-پیش بینی حجم فروش برای تاریخ معین آینده۳- رفتارهای مشتریان را میتوان مشخص کرد

۳. **چه کسانی ممکن است به نتایج این پروژه علاقمند باشند؟ چرا؟** صاحبان کسب وکار(پیش بینی نرخ فروش...)

---

۲) درک داده:

۱. **دادهها از کجا بدست آمده اند و چگونه جمع آوری شده اند؟** براساس یک فایل اکسل که مقادیر و مشخصات قیمت خانه ها در یک منطقه را بررسی میکند

۲. **هر یک از متغیرها چه چیزی را اندازه گیری میکنند؟** متغیرهای شماره ۱ تا ۵ (مستقل) هستند و متغیر شماره ۶ (price) متغیر وابسته است.براساس محور X مقادیر محور Y مشخص میشوند یعنی محور X متغیر مستقل است و متغیر Y وابسته است. و متغیر address راهم حذف کردیم چون نوع آن object بود. در مدل زدن باید متغیرهای وابسته و مستقل مشخص شوند با دستور iloc مشخص میکنیم کدام وابسته و کدام مستقل باشد.

۱) 'Avg. Area Income': Avg. Income of residents of the city house is located in.درآمد متوسط ساکنان منطقه
۲) 'Avg. Area House Age': Avg Age of Houses in same city عمرمتوسط خانه ها
۳) 'Avg. Area Number of Rooms': Avg Number of Rooms for Houses in same city متوسط تعداد اتاقهای هر خانه
۴) 'Avg. Area Number of Bedrooms': Avg Number of Bedrooms for Houses in same city متوسط تعداد اتاق خوابهای هر خانه
۵) 'Area Population': Population of city house is located in جمعیت شهر
۶) 'Price': Price that the house sold at قیمت فروخته شده هر خانه
۷) 'Address': Address for the house آدرس

۳. **آیا ابهامی در تعاریف داده ها وجود دارد؟** خیر هیچ ابهامی وجود ندارد داده خالی یا missing value و مقدار تکراری هم ندارد و همه ی داده ها از یک نوع بودند ونیازی به تغییرنوع داده هم نبود

۴. **آیا ممکن است در اندازه گیری متغیرها و یا ثبت داده ها خطایی وجود داشته باشد؟** بله امکان خطا هست ولی در داده های ما هیچ خطا یا missing value ای وجود ندارد چون همه ستونها دارای ۵۰۰۰ رکورد هستند.

۵. **چه متغیرهای دیگری اگر وجود داشتند، میتوانست به حل مسئله کمک کند؟** اگر فقط مساحت را داشته باشیم از الگوریتم رگرسیون خطی استفاده میشود اگر چند متغیره باشد از الگوریتم polynomial regression(الگوریتم چند متغیره) استفاده میکنیم.در رگرسیون خطی مدل و شیب و خط برازش را تخمین میزند.یکسری متغیر واقعی داریم و یکسری متغیری که پیش بینی شده$y^{\wedge}$ و به ما در حل مسئله کمک میکند برای ارزیابی باید (متغیر پیش بینی شده – متغیر واقعی) را استفاده کنیم. متغیر x (مساحت) و متغیر(y) قیمت خانه است اینها متغیرهای واقعی هستند.

متغیرهای پیش بینی شده y^ میباشد .مقدار β_1شیب خط است و x مقدار مستقل است و y مقدار وابسته است. $\beta.$ ثابت میباشد. و برای ارزیابی مقدار پیش بینی شده از فرمول $y^{\wedge}-y$ استفاده میشود (متغیر پیش بینی شده – متغیر واقعی) . هرچه قدر فاصله کمتر باشد پیش بینی درست تر است

$$y = \beta. + \beta_1 x$$

---

٦. **متغیرهای موجود از کدام نوعند (رسته ای – عددی)؟** عددی

٧. **خلاصه آماری متغیرهای موجود چیست؟** برای تحلیل آماری از تابع pairplot و distplot استفاده میکنیم. برای اینکه ببینیم قیمت خانه نرمال هست یا خیر از تابع distplot استفاده کردیم.(خانه ها از یک میلیون تا دو میلیون هستند وبیشترین قیمت خانه روی ١.٥ میلیارد است و این میانه دیتا است و این توزیع نرمالی ست)چولگی به چپ و راست ندارد- میانگین(mean) و انحراف معیار(std) دیتا فریم را بدست میاوریم.

٣) آماده سازی داده

١. **آیا نیاز به درآمیختن داده ها است؟ اقدامات و نتایج گزارش شود.**

٢. **آیا نیاز به پاکسازی داده است؟ اقدامات و نتایج گزارش شود.** بله.برای استاندارد کردن داده ها از **standardize** استفاده میکنیم تعداد اتاق خوابها ٣ تاست و جمعیت ۴۴ هزار تاست و داده ها نرمال نیستند و خیلی پایین و بالا هستند و باید داده ها استاندارد شوند.که از کتابخانه **sklearn.preprocessing** و از تابع **standardscaler** استفاده میکنیم و داده را استاندارد و یکسان میکنیم.

٣. **آیا نیاز به تبدیل داده است؟ اقدامات و نتایج گزارش شود.** فقط ستون Adress از نوع **object** است که انرا حذف میکنیم بقیه ستونها نوعشان ٦٤ **float** است که نیازی به تبدیل ندارند.با دستور **()info** اطلاعات کاملی از نوع و خالی بودن سطرها نشان داده میشود.

٤. **آیا نیاز به کاهش داده است؟ اقدامات و نتایج گزارش شود.** خیربا استفاده از دستور **duplicate** داده های تکراری را پیدا میکنیم که در اینجا داده تکراری نداشت

---

۴) مدلسازی:

روی داده های آموزش ،الگوریتم رگرسیون ساخته شود .اقدامات و نتایج گزارش شود.الگوریتم بانظارت را استفاده کردیم پس باید به ماشین مدل **x,y** را بدهیم تا یاد بگیره الگویی را کشف کند از اینکه $\beta.$, β_1چند باشد

$$y = \beta. + \beta_1 x$$

---

۵) ارزیابی:

۶) **مدلهای ارائه شده، روی دادههای آزمایش با استفاده از شاخصهای ارزیابی رگرسیون خطی در یادگیری ماشین ارزیابی شوند .اقدامات و نتایج گزارش شود.** انواع یادگیری ١- نظارت شده(دیتاهایی که برچسب دارند و به مدل آموزش میدهند ومدل رابطه بین الگوها را کشف

میکند و براساس الگو خروجی جدیدی میدهد) ۲- بدون نظارت(هیچ لیبلی ندارد براساس دیتایی که به آن داده میشود خوشه هایی را مشخص میکند) ۳- تقویتی(براساس محیط است) که پروژه رگرسیون خطی براساس یادگیری نظارت شده پیاده سازی شده است.اگر خط برازش خط خوبی باشد پس مدل ما مدل خوبی است. میانگین خطای نهایی را بدست میاورند براساس آن نشان میدهند کدام روش مناسب است. برای ارزیابی روشهای متفاوتی داریم(متریک قدرمطلق خطا- متریک جذر خطا- جذر میانگین توان دو خطا)را داریم.

در مدل زدن باید متغیرهای وابسته و مستقل مشخص شوند با دستور iloc مشخص میکنیم کدام وابسته و کدام مستقل باشد. برای اینکه ببینیم مدلمان مدل خوبی هست یا نه آموزش و آزمایش train and test split تقسیم میکنیم.برای اینکه ببینیم مدلمان خوب کار کرده یک مقدار برای تست میزاریم. ۶۰ درصد آموزش و ۴۰ درصد تست.از کتابخانه sklearn استفاده میکنیم.

۷) ۲.چه پیشنهاداتی دارید تا نتایج در محیط واقعی، آزمایش گردد؟

۸) استقرار:

۹) حال اگر بخواهید چنین الگوریتمی را در مقیاس صنعتی توسعه دهید، به این فکر کنید با چه چالش هایی مواجه خواهید شد و

۱۰) برای آن چه راهکارهایی دارید. موارد زیر را گزارش کنید:

۱. **چالشهای توسعه الگوریتم را بررسی کنید.** ۱- نداشتن داده های کافی برای آموزش ۲- وجود داده های نویزی و بی کیفیت ۳- بیش برازش ۴- کم برازش
۲. **چه راهکارهایی برای حل آن ها دارید؟**
۳. **چه ملزوماتی برای ارائه آن راهکارها نیاز دارید؟**

۱۱) نتیجه گیری:

۱. **انجام این پروژه چه یادگیری برای شما داشت؟**
۲. **با چه چالشهایی مواجه شدید؟ چگونه آنها را حل کردید؟**

۱۲) ارزیابی پروژه:

۱۳) کسانی که به بیش از ۶۰ درصد موارد بالا پاسخ صحیح داده باشند، نمره قبولی پروژه را دریافت خواهند کرد.

۱۴) پروژه را در سامانه تربیت مربی وگیت ها ب خود بارگذاری کنید word ، html **فایل**

۱۵) لینک گیت هاب خود را هم در سامانه قرار دهید.