- **Problem 1: Describe the California housing dataset.**

مجموعه داده مسکن کالیفرنیا یک مجموعه داده محبوب است که در یادگیری ماشین و وظایف تجزیه و تحلیل داده استفاده می شود و شامل اطلاعاتی در مورد قیمت مسکن و ویژگی های مختلف خانه ها در مناطق مختلف کالیفرنیا است. مجموعه داده شامل ویژگی هایی مانند میانگین ارزش خانه، درآمد متوسط، تعداد متوسط اتاق ها، جمعیت و غیره است. هدف از استفاده از این مجموعه داده معمولاً پیش‌بینی ارزش متوسط خانه بر اساس سایر ویژگی‌ها است. چون متغیر هدف ارزش متوسط خانه داخل خود دیتاست هست، یک تسک با ناظر محسوب می شود و از آنجایی که لیبل (ارزش متوسط خانه) یک مقدار از یک بازه پیوسته است، برای مدلسازی و پیش بینی قیمت خانه از تسک رگرسیون خطی استفاده شده است.

مجموعه داده به صورت عمومی در دسترس است و از منابع مختلف، از جمله مخزن github قابل دسترسی است که در اینجا دیتاست housing.csv از مخزن گیت هاب دانلود شده و در درایو گوگل خودم قرار دادم و لینک مربوطه در فایل کد ها قرار داده شده است. گیت هاب یک سرویس ابری است که به توسعه دهندگان امکان ذخیره و مدیریت و کنترل مداوم ورژنهای مختلف کد را می دهد و معمولا برای میزبانی پروژه های توسعه نرم افزار منبع باز استفاده می شود و یکی از مخازنی هست که میتوان از آن دیتاستهای مختلف را دانلود کرد و معمولاً برای تمرین تکنیک های تجزیه و تحلیل داده ها و یادگیری ماشین و همچنین برای اهداف آموزشی استفاده می شود.

مجموعه داده شامل ۲۰٦٤۰ رکورد و ۱۰ ویژگی است:

۱. **median_income: درآمد متوسط خانوارهای منطقه.**

۲. **housing_median_age: میانه سن خانه های منطقه.**

۳. **total_rooms: میانگین تعداد اتاق در هر خانه.**

٤. **total_bedrooms: میانگین تعداد اتاق خواب در هر خانه.**

٥. **Population: جمعیت منطقه.**

٦. **households: میانگین تعداد ساکنین در هر خانوار.**

۷. **Latitude: مختصات عرض جغرافیایی موقعیت منطقه.**

۸. **Longitude: مختصات طول جغرافیایی موقعیت منطقه.**

۹. **median_house_value: میانه ارزش خانه در منطقه (متغیر هدف).**

۱۰. **ocean_proximity: نزدیکی به اقیانوس یا سایر آب ها.**

یکی از چالش‌های کار با مجموعه داده مسکن کالیفرنیا، مقابله با مقادیر خالی (missing values) است که نیاز به بررسی دقیق دارد چون می تواند بر دقت و قابلیت اطمینان تحلیل و مدل های ساخته شده با استفاده از مجموعه داده تأثیر بگذارد. می توان از تکنیک های مختلفی مانند انتساب یا حذف مقادیر از دست رفته استفاده کرد و در این کد در مرحله preprocessing بجای مقادیر خالی، میانگین همان ستون قرار داده شده است.

- **Problem 2: Describe a dataset that is appropriate for a regression task.**

مجموعه داده الماس یک مجموعه داده کلاسیک ومحبوب در علم داده و یادگیری ماشین است و شامل اطلاعاتی در مورد ویژگی های مختلف الماس، مانند وزن قیراط، کیفیت برش، رنگ، وضوح، عمق، قیمت و سایر ویژگی های تقریباً ٥٤٠٠٠ الماس است. مجموعه داده های الماس اغلب برای پیش بینی قیمت یک الماس بر اساس ویژگی های مختلف آن استفاده می شود. این مجموعه داده شامل۱۰ فیچر است که قیمت بعنوان ستون هدف است. مجموعه داده الماس به صورت عمومی در دسترس است و از منابع مختلف، از جمله مخزن Kaggle قابل دسترسی است که در اینجا دیتاست diamonds.csv از مخزن کگل دانلود شده و لینک مربوطه در فایل کد ها قرار داده شده است. کگل یک پلتفرم رقابت آنلاین دانشمندان داده و متخصصان یادگیری ماشین است که به کاربران امکان می دهد مجموعه داده های مختلف را پیدا و منتشر کنند، مدل ها را در یک محیط علم داده مبتنی بر وب بسازند و با دیگر دانشمندان داده و مهندسین یادگیری ماشین کار کنند. از آنجایی که لیبل (قیمت) یک مقدار از یک بازه پیوسته است، برای مدلسازی و پیش بینی قیمت الماس ها از تسک رگرسیون خطی استفاده شده است. در اینجا شرحی از ستون های مجموعه داده الماس آورده شده است:

۱. **price: قیمت به دلار آمریکا (\$326--\$18,823) (ستون لیبل)**
۲. **Carat: وزن الماس که بر حسب قیراط اندازه گیری می شود. یک قیراط برابر با ۲۰۰ میلی گرم است.**
۳. **Cut: کیفیت تراش الماس که تعیین می کند الماس چقدر نور را منعکس می کند. مقادیر ممکن عبارتند از: منصفانه، خوب، بسیار خوب، ممتاز و ایده آل.**
٤. **Color: درجه رنگ الماس از J (بدترین) تا D (بهترین)**
٥. **Clarity: درجه شفافیت الماس که وجود عیوب داخلی یا خارجی را اندازه گیری می کند. (I۱ (بدترین)، SI۲، SI۱، VS۲، VS۱، VVS۲، VVS۱، IF (بهترین))**
٦. **x: طول الماس بر حسب میلی متر**
۷. **Y: عرض بر حسب میلی متر**
۸. **Z: عمق بر حسب میلی متر**
۹. **depth: درصد عمق الماس که به عنوان نسبت عمق به قطر متوسط الماس محاسبه می شود.**
۱۰. **table: عرض بالای الماس نسبت به پهن ترین نقطه الماس که به صورت درصد بیان می شود.**

یک چالش در تجزیه و تحلیل مجموعه داده های الماس مقابله با داده های پرت است، داده هایی که به طور قابل توجهی از توزیع نرمال مجموعه داده منحرف می شوند و می توانند تأثیر قابل توجهی بر تحلیل و مدل سازی آماری داشته باشند. در مورد مجموعه داده های الماس، ممکن است نقاط پرت در وزن، قیمت یا سایر ستون های عددی قیراط رخ دهد که با نمودار باکس پلات آنها را شناسایی و چون تعداد آنها نسبت به کل مجموعه ناچیز است، آنها را حذف میکنیم.

- **Problem 3: Choose an arbitrary dataset which you like and describe it in details.**

مجموعه داده باغ وحش یک مجموعه داده محبوب است که در یادگیری ماشین و برای تسک کلاس بندی چندکلاسی استفاده می شود و شامل اطلاعاتی در مورد حیوانات مختلف و طبقه بندی آنها به دسته های مختلف با استفاده از رگرسیون لجستیک است. مجموعه داده شامل ۱۰۱ سمپل (ردیف) و ۱۸ فیچر (ستون)، شامل ویژگی های کمی و کیفی است. مجموعه داده باغ وحش به صورت عمومی در دسترس است و از منابع مختلف، از جمله مخزن UCI قابل دسترسی است که در اینجا دیتاست zoo.csv از مخزن uci دانلود شده و لینک مربوطه در فایل کد ها قرار داده شده است.

در اینجا توضیح مختصری از ویژگی های مجموعه داده باغ وحش آورده شده است:

۱. animal: نام حیوان .

۲. hair: این که آیا حیوان مو دارد یا نه .

۳. feathers: اینکه حیوان پر داشته باشد یا نداشته باشد .

٤. eggs: اینکه آیا حیوان تخم می گذارد یا نه .

٥. milk: این که آیا حیوان شیر تولید می کند یا نه .

٦. airborne: اینکه آیا حیوان می تواند پرواز کند یا نه .

۷. aquatic: این که آیا حیوان در آب زندگی می کند یا نه .

۸. predator: این که آیا حیوان شکارچی است یا نه .

۹. toothed: اینکه حیوان دندان داشته باشد یا نداشته باشد .

۱۰. backbone: این که آیا حیوان دارای ستون فقرات است یا نه .

۱۱. breathes: آیا حیوان هوا را تنفس میکند یا نه .

۱۲. venomous: اینکه حیوان سمی باشد یا نباشد .

۱۳. fins: این که آیا حیوان بال دارد یا نه .

۱٤. legs: تعداد پاهایی که حیوان دارد (عددی: ۰، ۲، ٤، ٥، ٦، ۸).

۱٥. tail: اینکه حیوان دم داشته باشد یا نداشته باشد .

۱٦. domestic: این که حیوان اهلی باشد یا نباشد .

۱۷. catsize: این که آیا حیوان اندازه گربه است یا نه .

۱۸. type: نوع کلاس حیوان (عدد صحیح: ۱-۷). (ستون لیبل)

یکی از چالش‌های کار با مجموعه داده باغ وحش، مقابله با توزیع کلاس نامتعادل(imbalanced) است. توزیع کلاس نامتعادل زمانی اتفاق می افتد که تعداد نمونه ها در یک کلاس به طور قابل توجهی بیشتر یا کمتر از تعداد نمونه های کلاس های دیگر باشد. در مجموعه داده باغ وحش، ۷ کلاس مختلف وجود دارد که نشان دهنده انواع مختلف حیوانات است. ممکن است برخی از کلاس ها تعداد نمونه های بسیار بیشتری نسبت به سایرین داشته باشند. توزیع نامتعادل کلاس می‌تواند چالش‌هایی را در آموزش مدل‌های یادگیری ماشین ایجاد کند، زیرا آنها تمایل دارند نسبت به طبقه اکثریت تعصب داشته باشند. این می تواند منجر به عملکرد ضعیف و پیش بینی های نادرست برای طبقات اقلیت شود. برای مقابله با این چالش، تکنیک هایی مانند نمونه برداری بیش از حد از کلاس اقلیت، کم نمونه برداری از کلاس اکثریت با استفاده ازcross validationو stratified k-foldبرای متعادل کردن توزیع کلاس و بهبود عملکرد مدل استفاده شده است.

- در ادامه علاوه بر تسک های تمرین، در فایل کد هر سه دیتاست سعی شده تا سایر مفاهیم آموزش داده شده در طی دوره علم ۱ از جمله مدلسازی با انواع روشهای یادگیری نظارت شده و بدون ناظر مثل معادله نرمال و گرادیان کاهشی و رگرسیون لجستیک، نرمال سازی داده ها، متریک های مختلف و ارزیابهای مدل، کراس ولیدیشن، غلبه بر بیش برازش، رگولاریزیشن، تعیین هایپرپارامترها، منحنی یادگیری، ماتریس درهم ریختگی، تکنیکهای پیش‌پردازش مثل تبدیل داده کتگوریکال به داده نیومریکال و هندل کردن داده های پرت و تکراری و خالی و رسم انواع نمودارها، استفاده شود.