به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

درس داده کاوی پیشرفته

تمرین اول

طراحان	بخش نظری: فرشاد حسامی Farcshad@gmail.com بخش عملی: سعید رحیمی Saeedrhimi۲@gmail.com
تاریخ بارگذاری	۱۴۰۳/۱۲/۰۷
مهلت ارسال	۱۴۰۳/۱۲/۱۹

فهرست

جدول‌ها

بخش نظری

۱. سوال اول

در سوال اول قصد داریم انواع مختلف ویژگی‌های داده‌ها و نحوه‌ی برخورد مناسب با آن‌ها را مورد بررسی قرار دهیم. فرض کنید مجموعه‌داده‌ای مرتبط با آگهی‌های لیست‌شده برای فروش خانه‌ها را جمع‌آوری کرده‌ایم که این ستون‌ها(ویژگی‌ها) را شامل می‌شوند.

- سال ساخت
- مساحت (متر مربع)
- طبقه
- تعداد کل طبقات
- نوع خانه
- تعداد اتاق‌ها
- آسانسور دارد
- پارکینگ دارد
- سطح امنیت منطقه
- نوع پوشش کف واحد
- قیمت

۱. نوع هر یک را مشخص کنید. (ابتدا پیوسته یا گسسته بودن و سپس، دودویی، اسمی، ترتیبی و یا عددی بودن آن‌ها را مشخص نمایید)

۲. برای هر ویژگی، از بین نمودارهای زیر، نمودارهای مناسب جهت نمایش آن ویژگی را مشخص کنید.

Histogram, Pie chart, Box plot, Bar chart

۳. فرض کنید می‌خواهید رابطه‌ی بین مساحت خانه و قیمت آن‌را با یک نمودار به تصویر بکشید، بهترین انتخاب برای این‌کار چه نموداری خواهد بود؟

۴. در نظر بگیرید که به صورت کلی با بالا رفتن مساحت خانه، انتظار داریم که قیمت آن نیز افزایش یابد. با این حال، در برخی موارد، به دلایل مختلف خانه‌هایی با مساحت کمتر ممکن است قیمت بیشتری داشته باشند. با این فرض‌ها میزان همبستگی بین قیمت و مساحت را تحلیل کنید. (برای مثال مشخص کنید که این مقدار مثبت است یا منفی، تا چه حد به صفر نزدیک است و یا از صفر فاصله دارد.)

۲. سوال دوم

دانش‌آموزان یک کلاس در دو آزمون (ریاضی و فیزیک) شرکت کرده‌اند. نمره‌ی آزمون ریاضی از ۲۰ و نمره‌ی آزمون فیزیک از ۱۰۰ محاسبه‌شده است. نمرات دانش‌آموزان برای دو آزمون به شرح زیر هستند.

Math = [۱۷.۵ , ۱۶, ۱۷, ۱۵, ۱۵, ۱۶, ۱۸, ۲۰, ۱۳.۵, ۱۷, ۱۶, ۱۸, ۱۳.۵ , ۲۰, ۱۵.۰, ۱۶.۵, ۱۶.۵, ۱۵, ۱۷, ۱۸]

Physics = [۲۰, ۸۱, ۸۰, ۷۰, ۸۵, ۷۵, ۸۷, ۹۷, ۳۵, ۹۰, ۸۳, ۸۸, ۶۹, ۱۰۰, ۸۱, ۱۰, ۷۹, ۷۸, ۸۴, ۹۱]

۱. برای هریک از این داده‌ها میانگین، انحراف‌معیار، میانه، و چارک‌های اول و سوم را محاسبه نمایید.

۲. توضیح‌دهید کدام‌یک از این معیارها برای ارزیابی عملکرد کلی کلاس بهتر است که استفاده شود.

۳. بررسی نمایید که آیا داده‌های داده‌شده شامل هرگونه ناهنجاری هستند و در این‌صورت، این ناهنجاری‌ها چه تاثیری داشته‌اند.

۴. برای هریک هیستوگرام رسم‌کنید، میانگین و میانه را در آن نشان‌داده و سپس توزیع را در هر یک تحلیل نمایید.

۵. نمودار boxplot را برای هر یک رسم و تحلیل کنید.

۶. دو توزیع را نرمال‌سازی کنید و سپس آن‌ها را مقایسه کنید.

۷. در مورد نحوه‌ی استفاده از qq-plot جهت تشخیص نرمال بودن توزیع توضیح دهید، سپس با استفاده از qq-plot تحلیل نمایید که هریک از این دو توزیع، توزیع نرمال هستند یا خیر.

۸. همبستگی نمرات در این دو آزمون را بررسی کنید.

۹. در ادامه‌ی سوال فرض کنید که علاوه بر این داده‌ها، داده‌های دیگری هم وجود داشته‌اند که پاک‌سازی شده‌اند. دلیل حذف‌شدن هریک از این‌داده‌ها را که در ادامه آورده شده‌اند را توجیه کنید.

Math = [N/A, ۲۱, ۰, A+, “۱۹”]

Physics = [۹۰, ۷۶, ۰, B, “۸۴”]

۳. سوال سوم

در این سوال قصد داریم با استفاده از chi-square test، استقلال رشته‌ی دانشگاهی از جنسیت را مورد بررسی قرار دهیم.

فرض کنید ۲۲۰ نفر از ورودی‌های یک دانشگاه مورد بررسی قرار گرفته‌اند. از بین این افراد ۱۲۰ نفر پسر و ۱۰۰ نفر دختر هستند. از بین ۱۲۰ ورودی پسر، ۳۰ نفر به رشته‌ی کامپیوتر، ۴۰ نفر به رشته‌ی برق و ۵۰ نفر به رشته‌ی مکانیک وارد شده اند. از بین ۱۰۰ نفر ورودی دختر این تعداد به‌ترتیب برابر ۵۰، ۳۰ و ۲۰ می‌باشد.

با استفاده از chi-square test مستقل بودن رشته از جنسیت را با significance level ۰.۰۵ بررسی کنید.

بخش عملی

۱. مقدمه

به‌عنوان یک مهندس داده، یکی از مهارت‌های کلیدی شما توانایی جمع‌آوری و پردازش داده‌ها از منابع مختلف است. در این تمرین، شما با یک مجموعه‌داده از آگهی‌های املاک منتشر شده در وب‌سایت دیوار برای پنج شهر ایران کار خواهید کرد. این داده‌ها اخیراً توسط یک کرالر استخراج شده‌اند و به دلیل ماهیت پردازش خودکار، شامل نویز و اشکالات متعددی هستند. هدف شما در این تمرین این است که این مجموعه داده را گام‌به‌گام پاک‌سازی کرده و به فرمی ساختاریافته و قابل استفاده تبدیل کنید. در ادامه، با انجام تحلیل‌های مختلف، الگوهای موجود در داده‌ها را کشف کرده و بینش‌های ارزشمندی استخراج خواهید کرد.

در هر بخش، ابتدا شرایط موجود را به‌دقت تحلیل کنید و مسیر دستیابی به هدف موردنظر را مشخص نمایید. در این مسیر، ممکن است با چالش‌ها و موانعی روبه‌رو شوید؛ وظیفه شما این است که این مشکلات را شناسایی کرده، درباره آن‌ها بیاندیشید، راهکارهای ممکن را بررسی کنید و در نهایت، بهترین تصمیم را برای حل آن‌ها و پیشبرد هدف اتخاذ نمایید.

مهم است که تمام این مراحل، از تحلیل شرایط و شناسایی مشکلات گرفته تا استدلال‌های شما برای انتخاب بهترین راه‌حل، به‌دقت مستند شوند. این مستندسازی به ما کمک می‌کند تا نحوه تفکر، تحلیل و تصمیم‌گیری شما را بهتر درک کنیم و ببینیم چگونه با مسائل برخورد می‌کنید.

توجه داشته باشید که این فرآیند بخش مهمی از ارزیابی عملکرد شما در این تمرین را تشکیل می‌دهد و در نتیجه نهایی شما تأثیر قابل‌توجهی خواهد داشت. بنابراین، سعی کنید تمامی مراحل را با دقت و جزئیات کافی در پاسخ‌های خود منعکس کنید تا توانایی تحلیل و حل مسئله شما به‌خوبی نمایان شود.

۲. توضیحات مجموعه داده

جدول ۱. شرح دادگان موجود در مجموعه داده‌های آگهی ها

نام ستون	توضیحات
Title	عنوان آگهی
PropertySize	مساحت ملک بر حسب متر مربع

TotalPrice	قیمت کل ملک که به تومان نمایش داده می‌شود
PriceperMeter	قیمت هر متر مربع از ملک که به تومان نمایش داده میشود
RoomCount	تعداد اتاق‌های ملک
BuildYear	سال ساخت ملک
FloorNumber	شماره طبقه‌ای که ملک در آن قرار دارد
TotalFloors	تعداد کل طبقات ساختمان
Characteristics	ویژگی‌های کلی واحد
Features	امکانات واحد
Exchangable	فروشنده ملک، معاوضه آن با ملک یا دارایی دیگری را میپذیرد یا خیر
Description	توضیحات تکمیلی مربوط به ملک
URL	لینک آگهی در وبسایت دیوار
CrawlDate	تاریخ دریافت و ذخیره‌سازی آگهی در سیستم

۳. مراحل تمرین

۱٫۳ – بررسی اولیه داده ها

أ. ابتدا به این لینک مراجعه کنید و ۱۰ آگهی را به‌صورت تصادفی مشاهده کنید.

ب. بخش‌های مختلف هر آگهی در سایت دیوار را بررسی کنید تا با ساختار اطلاعاتی آن‌ها آشنا شوید.

ج. در سایت دیوار یک آگهی در دسته فروش مسکونی را تا مرحله‌ی نهایی انتشار پیش ببرید تا با نکات مختلف آن آشنا شوید و همچنین بررسی کنید چه اطلاعاتی به‌صورت اجباری از کاربر دریافت می‌شود.

نکته: این بخش احتیاج به مستند سازی متن یا نموداری در گزارش نداشته و نمره ای هم ندارد، اما انجام آن توصیه می‌شود، زیرا در مراحل بعدی برای اتخاذ تصمیم‌های داده‌محور و شناخت بهتر ساختار داده‌ها به شما کمک خواهد کرد.

۲٫۳ – بارگذاری و نمایش داده ها

أ. دیتاست های داده شده مربوط به شهرهای مختلف را در محیط کاری خود بوسیله کتابخانه Pandas در زبان پایتون به صورت DataFrame بارگذاری کنید.

ب. ۵ نمونه از آگهی‌های ثبت‌شده در هر شهر به همراه نام ستون ها را نمایش دهید تا با فرمت و نحوه ذخیره‌سازی آن‌ها آشنا شوید.

۳٫۳ – تجمیع داده ها

أ. به‌عنوان یک مهندس داده، همواره باید رویکردی تحلیلی و آینده‌نگر داشته باشید و هیچ داده‌ی مفیدی را بی‌دلیل از دست ندهید. با این دیدگاه، تمامی آگهی‌های مربوط به شهر های مختلف که در مجموعه داده های جداگانه در اختیار شما قرار گرفته است را در یک دیتافریم واحد تجمیع کنید. در این فرآیند، ممکن است با چالش‌ها و مشکلاتی مواجه شوید؛ آن‌ها را شناسایی، مستند کرده و راهکارهای مناسب برای رفعشان ارائه دهید.

ب. پس از ادغام داده‌ها، ۵ نمونه از آگهی‌ها را همراه با نام ستون‌ها و تعداد کل آگهی‌های تجمیع‌شده نمایش دهید.

٤٫۳ - شناسایی و حذف داده‌های نادرست

برخی از آگهی‌ها دارای ساختار متفاوتی هستند که باعث شده کرالر برخی اطلاعات را به اشتباه استخراج کند و در نتیجه، داده‌های آن آگهی‌ها نامعتبر شوند. این مشکل به‌ویژه در ستون‌های قیمت کل و سال ساخت دیده می‌شود.

I should check it again

أ. بررسی کنید که داده‌های موجود در این ستون‌ها چه الگوهایی دارند و چه تفاوت‌هایی میان مقادیر معتبر و نامعتبر مشاهده می‌شود. بر اساس این بررسی، معیارهایی برای تشخیص و حذف آگهی‌های نامعتبر تعریف کنید، آگهی های نامعتبر را حذف کنید و فرآیند تصمیم گیری تا اجرا را مستند کنید.

ب. تعداد آگهی ها قبل و بعد از این عمل را گزارش کنید.

راهنما: مقادیر معتبر در این ستون‌ها دارای یک الگوی مشخص هستند.

۵٫۳ - بررسی داده‌های تکراری

أ. بررسی کنید که آیا آگهی‌های تکراری در دیتاست وجود دارند یا نه؟ در صورت وجود تصمیم مناسب در برخورد با آنها چیست؟ آن را اعمال کنید.

ب. توضیح دهید از چه ستون یا ستون هایی برای شناسایی آگهی‌های تکراری استفاده کردید و چرا این انتخاب شما مناسب است.

ج. در هر مرحله تعداد آگهی ها را گزارش کنید.

۶٫۳ – بررسی نوع داده ها

راهنما: مقادیر گمشده در این مجموعه‌داده با عبارت "Not Found" مشخص شده‌اند.

أ. با استفاده از دستور dtype() در پایتون بررسی کنید که هر ستون با چه نوع داده‌ای در دیتافریم ادغام شده‌ی آگهیِ شهرها، ذخیره شده است.

just do .dtype()?

ب. مشخص کنید که هر ستون باید با چه نوع داده‌ای ذخیره شود تا تحلیل ها روی آن‌ها به‌درستی انجام شود و از نظر حافظه نیز بهینه باشد.

did not considered!

ج. موانع احتمالی تغییر نوع داده‌ها را شناسایی کرده و مستند کنید، آن‌ها را برطرف کنید و فرمت داده‌های ستون‌ها را تغییر دهید.

ye toosh koni!:
bayed biam baray Not Found ye fekri konam tokhmi = 0 gozashtam

د. حجم اشغال شده در حافظه برای هر ستون قبل و بعد از تغییر را گزارش کنید

ه. پس از انجام تغییرات، بررسی کنید که نوع داده‌ها به‌درستی اصلاح شده‌اند یا نه و آنها را نمایش دهید.

نکته: ممکن است بعد از تغییر، دستور dtype() پاسخ دقیقی ندهد، در اینصورت روش دیگری برای نمایش و اطمینان از تغییر نوع داده‌ها به کار ببرید.

۷٫۳ - پردازش ستون‌های ویژگی‌ها (Characteristics) و امکانات (Features)

أ. ستون‌های ویژگی‌ها و امکانات را بررسی کنید و ۵ نمونه از آن‌ها را نمایش دهید.

ب. این ستون‌ها شامل مجموعه‌ای از ویژگی ها و امکانات تجمیع‌شده هستند. آیتم های مختلف را از هم جدا کنید و مقادیر یکتای آن‌ها را چاپ کنید.

ج. مشخص کنید که هر آیتم در چه درصدی از آگهی‌ها وجود دارد.

د. تحلیل کنید که کدام آیتم‌ها ارزش تبدیل شدن به یک ستون جداگانه را دارند و سپس آن‌ها را به ستون‌های مستقل تفکیک کرده و به دیتاست اضافه کنید.

۸٫۳- بررسی محتوای ستون‌های توضیحات (Description) و عنوان (Title)

أ. بیش از ۱۰ نمونه تصادفی از ستون‌های "توضیحات" و "عنوان" بررسی کنید تا با انواع حالت های آنها آشنا شوید و سپس ۵ نمونه از آن‌ها را نمایش دهید.

ب. تحلیل کنید که آیا این دو ستون دارای اطلاعات ارزشمندی برای ما هستند که بتوان با کمک آنها اطلاعات دیتاست را بهبود داد؟ در صورت وجود با آوردن حداقل ۲ مثال از بین داده ها، آنها را بیان کنید.

ج. در صورتی که تصمیم بگیریم از محتوای این دو ستون برای بهبود دیتاست استفاده کنیم، با چالش هایی روبرو میشویم. با آوردن حداقل ۲ مثال از بین داده ها، آن ها را بیان کنید.

۹٫۳ – پردازش داده های گمشده

أ. بررسی کنید که چه درصدی از مقادیر هر ستون در دیتاست گمشده است. سپس، روش‌های مختلفی را برای مدیریت داده‌های گمشده پیشنهاد دهید و آن‌ها را پیاده‌سازی کنید. (حداقل ۲ روش)

in 2 bakhsh monde

ب. پس از انتخاب روش مناسب برای هر ستون، آن روش را توضیح داده و دلایل انتخاب خود را شرح دهید.

نکته: ساده‌ترین روش، حذف ردیف‌هایی است که دارای مقدار گمشده هستند، اما این همیشه بهترین انتخاب نیست. برای هر ستون، مناسب‌ترین روش را با در نظر گرفتن ساختار داده‌ها و امکان‌پذیری پیاده‌سازی در این مجموعه داده انتخاب کنید.

۱۰٫۳ – شناسایی مقادیر پرت

أ. با رسم نمودار مناسب، وجود داده‌های پرت را در هر یک از ستون‌های عددی شناسایی و نمایش دهید سپس تصمیم مناسبی در مورد نحوه برخورد با این داده‌ها اتخاذ کنید.

tasmim monaseb monde!

۱۱٫۳ – مصور سازی داده ها

نکته: انتخاب نوع نمودار مناسب در این بخش بر عهده شماست. هدف از انتخاب نمودار، ارائه بهترین امکان مقایسه و تحلیل داده‌ها است. بنابراین، با بررسی انواع مختلف نمودارها، مناسب‌ترین گزینه را انتخاب کنید تا داده‌ها به‌صورت شفاف و قابل تفسیر نمایش داده شوند.

أ. میانگین قیمت هر متر مربع خانه در هر شهر را با نمودار مناسب نمایش دهید.

delete outliers and claculate

ب. میانگین قیمت هر متر مربع خانه را بر اساس سال ساخت برای هر شهر را در نمودار مناسب نمایش دهید و تحلیل کنید.

ج. نمودار توزیع برای سال ساخت، تعداد اتاق، قیمت هر متر مربع، قیمت کل را برای هر شهر نمایش دهید و میانه آنها را با عدد آن روی نمودار مشخص کنید.

د. آیا امکان نمایش داده‌هایی با بیش از دو بُعد در یک نمودار در یک صفحه دوبُعدی وجود دارد؟ اگر پاسخ شما مثبت است، توضیح دهید و برای هر شهر، نموداری ارائه دهید که ابعاد آن شامل تعداد اتاق، مساحت خانه و قیمت هر مترمربع باشد.

۱۲٫۳ – افزودن اطلاعات محله به آگهی‌های شهر تهران (این بخش امتیازی است.)

در کنار مجموعه داده‌هایی که اطلاعات آگهی‌ها را در خود جای داده‌اند، یک دیتاست دیگر در اختیار شما قرار داده شده است که شامل اطلاعات محله‌های تهران و خیابان‌های آن است. هدف ما این است که از این دیتاست استفاده کنیم و یک ستون جدید به آگهی‌های مربوط به شهر تهران اضافه کنیم که محله هر آگهی را مشخص کند.

جدول ۲. شرح دادگان موجود در مجموعه داده محله های تهران

نام ستون	توضیحات
Neighborhood	نام محله
Nearby Streets	نام خیابان های موجود در آن محله

در این بخش، شما وظیفه دارید که محله‌ی هر آگهی را از میان داده‌های موجود استخراج کنید. این کار به شما کمک می‌کند تا بتوانید توزیع جغرافیایی آگهی‌ها را بررسی کنید و تحلیل‌های بعدی را بر اساس مناطق شهری انجام دهید. اگر مراحل قبلی را به‌درستی انجام داده باشید، باید درک روشنی از ساختار آگهی‌ها و اینکه چه ستون یا ستون‌هایی برای این کار مناسب هستند پیدا کرده باشید.

شما می‌توانید برای این بخش از روش‌های کلاسیک استفاده کنید یا روش‌های پیشرفته‌تر، مانند استفاده از یک مدل زبانی بزرگ (LLM). انتخاب روش به عهده شماست. با این حال، ما استفاده از روش‌های کلاسیک داده‌کاوی را با توجه به محتوای این درس می‌پذیریم و می‌دانیم که در مواجهه با این تسک ممکن است نتیجه‌ای به خوبی روش‌های پیشرفته نداشته باشد.

چیزی که اهمیت دارد این است که شما بتوانید به‌دقت تحلیل کنید، مشکلات مختلفی که در مسیرتان قرار دارد را شناسایی کنید و تا حد امکان برای آن‌ها راه‌حل ارائه دهید و پیاده‌سازی

کنید. حتی اگر نتوانید همه مشکلات را برطرف کنید، تلاش کنید درصد بیشتری از محله‌های آگهی‌ها را با دقت بالاتر شناسایی کنید.

أ. ابتدا ۵ ردیف از دیتاست محله ها را نمایش دهید.

ب. آگهی های شهر تهران را انتخاب کنید، ستون‌های مختلف این دیتاست را بررسی کرده و مشخص کنید که کدام ستون یا ستون‌ها اطلاعاتی ارائه می‌دهند که می‌توان از آن‌ها برای تعیین محله استفاده کرد.

ج. بررسی کنید که این ستون یا ستون ها چه ایرادات محتوایی دارند که هنگام استفاده برای رسیدن به این هدف برای ما مشکل ساز خواهند شد. آنها را بیان کنید و تا جای ممکن رفع کنید. (حداقل ٤ مورد)

4th item should be asked

د. با توجه به دو دیتاست موجود، فرآیند تعیین محله برای هر آگهی را آغاز کنید. ابتدا روش مناسب برای این کار را مشخص کنید و گام‌به‌گام پیش بروید. در این مسیر ممکن است با چالش‌های مختلفی مواجه شوید که لازم است آن‌ها را تحلیل کرده و تا حد ممکن برطرف کنید.

ه. مشخص کنید که برای چه درصدی از آگهی‌ها توانسته‌اید محله مربوطه را تعیین کنید.

و. ۵ تا از گران ترین محله های تهران را با توجه به میانگین قیمت هر متر خانه مشخص کنید.

ز. ۵ تا از گران ترین خانه های تهران را نمایش دهید. (فقط ستون های عنوان، قیمت تمام شده و محله را نمایش دهید)

ح. ۲۰ آگهی را به‌صورت تصادفی انتخاب کنید، یک جدول درست کنید و ستون یا ستون‌هایی که برای تعیین محله استفاده شده‌اند را در کنار یکدیگر نمایش دهید و محله‌ای که به هر آگهی نسبت داده شده است را نیز در جدول قرار دهید. (هر ردیف باید مربوط به یک آگهی باشد)

نکته: تا جای ممکن تحلیل ها، مشکلات، روش هایی که برای رفع هر مشکل با آن برخورد کردید یا حتی اگر مشکلی وجود داشت و با ابزار های موجود نمیتوانستید آن را برطرف کنید را مستند کنید.

ملاحظات

تمامی نتایج شما باید در یک فایل فشرده با عنوان DM_CA۱_StudentID تحویل داده شود.

- خوانایی و دقت بررسی‌ها در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. به تمرین‌هایی که به صورت کاغذی تحویل داده شوند یا به صورت عکس در سایت بارگذاری شوند، ترتیب اثری داده نخواهد شد.
- بخش اصلی نمره به گزارش شما تعلق می‌گیرد و دستیاران الزامی برای اجرای تمام کدهای شما در صورتی که در گزارش به آن‌ها اشاره‌ای نکرده باشید ندارند. لطفا تمام موارد مورد نیاز را در گزارش ذکر کنید.
- کدهای نوشته شده برای هر بخش را با نام مناسب مشخص کرده و به همراه گزارش تکلیف ارسال کنید. همه‌ی کدهای پیوست گزارش بایستی قابلیت اجرای مجدد داشته باشند. در صورتی که برای اجرا مجدد آنها نیاز به تنظیمات خاصی می‌باشد بایستی تنظیمات مورد نیاز را نیز در گزارش خود ذکر کنید.
- برای تحویل تمارین از چارچوب قرارداده شده در سامانه، سایت درس به آدرس dm-ut.github.io و یا گروه تلگرام استفاده کنید.
- در صورت قصد ارسال تمرین به صورت دیگر (انگلیسی، latex و ...)، لطفا پیش از ارسال با دستیار مسئول تمرین هماهنگ کنید.
- توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارائه شده باید نتیجه فعالیت فرد نویسنده باشد (هم‌فکری خارج از چارچوب و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تخلف برای همه‌ی افراد مشارکت کننده، نمره تمرین، صفر در نظر گرفته خواهد شد.
- در صورت استفاده از ابزارهای هوش مصنوعی، قوانین استفاده در پایان تمرین را مطالعه کنید.
- در پایان گزارش ارسالی خود، اظهارنامه بند ۸ از قوانین استفاده مسئولانه از هوش مصنوعی را قرار دهید.
- در صورت بروز هرگونه مشکل با ایمیل زیر در ارتباط باشید:

فرشاد حسامی Farcshad@gmail.com

سعید رحیمی Saeedrhimi۲@gmail.com

مهلت تحویل: ۱۹ اسفند ۱۴۰۳

مهلت تحویل با تاخیر: ۲۶ اسفند ۱۴۰۳

استفاده مسئولانه از هوش مصنوعی

۱. هدف و اصول کلی

هدف

- ترویج استفاده **اخلاقی** و **مسئولانه** از LLMها (مانند ChatGPT، Deepseek) به عنوان ابزار کمکی
- اطمینان از **مشارکت فعال** دانشجویان در تکالیف و درک راه‌حل‌های آن‌ها
- حفظ **صداقت علمی** در عین بهره‌گیری از ابزارهای مدرن هوش مصنوعی

اصول کلی

- تمرین باید نتیجه تلاش و زحمت شخصی شما باشد.
- باید به تمام بخش‌های تمرین، اعم از پیاده‌سازی و تحلیل نتایج مسلط باشید.
- تمامی کدها باید توسط خود شما اجرا شده و نتایج قابل مشاهده باشند.
- تمام مراحل انجام تمرین باید مستند و قابل پیگیری باشد.
- هرگونه نتیجه‌گیری و تحلیل باید بر اساس درک شخصی شما باشد.
- **LLMها ممکن است پاسخ‌های نادرست یا قدیمی تولید کنند، اولویت با مطالب و کارگاه‌های درس است.**

موارد ذکر شده در ادامه این سند، به عنوان **راهنمایی بیشتر** برای انجام تمرین آورده شده‌اند. با این حال، **مسئولیت تطبیق کار با اصول کلی فوق بر عهده شماست**. توجه داشته باشید که ممکن است مواردی در ادامه ذکر نشده باشند که با اصول کلی ذکر شده در تضاد باشند. در چنین مواردی به تشخیص دستیار آموزشی و دستیار مسئول، شما موظف به پاسخ‌گویی در قبال تمرین خود هستید. عدم رعایت هر یک از اصول فوق می‌تواند منجر به کسر نمره یا عدم پذیرش تمرین شود.

۲. استفاده مجاز از LLMها

شما می‌توانید از LLMها برای موارد زیر استفاده کنید:

- روشن‌سازی مفاهیم (مثال: "خوشه‌بندی DBSCAN چگونه کار می‌کند؟")
- **کمک** در اشکال‌زدایی (مثال: شناسایی خطاهای گرامری یا منطقی در کد)
- ایده‌پردازی رویکردها (مثال: "روش‌های مدیریت داده‌های missing را پیشنهاد دهید")

الزامات استفاده مجاز:

- ثبت تعاملات اصلی: (به بخش ۴ مراجعه کنید.)
- درک راه‌حل: باید قادر به توضیح هر خط کد یا منطق استفاده شده باشید.

۳. استفاده غیرمجاز از LLMها

اقدامات ممنوع شامل:

- کپی-پیست مستقیم خروجی‌های LLM بدون تغییر
- استفاده از LLMها برای حل اصلی مسائل (مثال: "این سؤال تکلیف را برای من حل کن")

- گرفتن کد از سایر دانشجویان به هر شکل غیر مجاز است، تغییر و پارافریز کردن کد دیگران توسط LLM نیز قابل قبول نیست.
- هرگونه استفاده که منجر به عدم احاطه شما به موضوع تمرین شود.

۴. مستندسازی

ارجاع به مشارکت‌های LLM: افزودن پانویس یا توضیح (مثال: کد با رعایت قوانین به کمک ChatGPT نوشته شده است.)

- نیازی به اشتراک گذاری پرامپت‌ها و سابقه چت نیست.
- مستندسازی تک تک تعاملات با هوش مصنوعی هدف این بخش نیست. اشاره کوتاه و کلی در بخش‌های مورد استفاده کافی است. در نظر داشته باشید که مستندسازی به معنای رفع مسئولیت نبوده و باید **اصول کلی** را رعایت کنید.

۵. آمادگی ارائه شفاهی

آماده دفاع از کار خود باشید: در صورت درخواست دستیار تمرین در بازه زمانی اعلام شده برای ارائه شفاهی، باید:

- رویکرد، کد یا نتایج خود را توضیح دهید.
- درک مفاهیم کلیدی را نشان دهید (مثلاً چرا یک الگوریتم خاص انتخاب شده است)

عدم توضیح کافی کار شما ممکن است منجر به جریمه شود (بخش ۶)

۶. پیامدهای تخلفات

- **تخلفات جزئی (مثل مستندسازی ناقص):** کاهش نمره
- **تخلفات عمده (مثل کپی-پیست بدون تغییر):** نمره ۵۰- در تکلیف
- **تخلفات مکرر:** نمره ۵۰- در تکلیف و گزارش به استاد

۷. موارد تکمیلی

- **از LLMها به عنوان معلم استفاده کنید، نه پاسخ‌نامه تمرین‌ها:** اولویت را به مهارت‌های حل مسئله خود بدهید.
- **خروجی‌ها را متقابلاً تأیید کنید:** پیشنهادات LLM را با کتاب مرجع درس، اسلایدها و کارگاه‌ها مقایسه کنید.
- **از دستیاران آموزشی کمک بگیرید:** اگر پاسخ LLM یا نحوه استفاده شما را گیج می‌کند، در ساعات متعارف از دستیاران آموزشی کمک بگیرید.

۸. اظهارنامه

این عبارت را در تکلیف ارسالی خود قرار دهید:
"تأیید می‌کنم که از LLMها مطابق با دستورالعمل‌های بارگذاری شده در سامانه Elearn درس به طور مسئولانه استفاده کرده‌ام. تمام اجزای کار خود را درک می‌کنم و آماده بحث شفاهی درباره آنها هستم."