سؤال ۱ – تحلیل کوواریانس و همبستگی در دیتاستCalifornia Housing

در این سؤال باید ساختار وابستگی بین ویژگی‌های عددی در دیتاست California Housing را بررسی کنید و مشخص کنید کدام متغیرها بیشتر با «قیمت متوسط خانه» مرتبط هستند.

الف) دیتاست California Housing را از sklearn بارگذاری کنید، آن را به یک DataFrame تبدیل کنید و فقط ویژگی‌های عددی ورودی و متغیر هدف (قیمت) را نگه دارید. خلاصه‌ای از داده (تعداد سطر، ستون، بازه‌ی مقادیر هر ستون) گزارش کنید.

ب) ماتریس کوواریانس و ماتریس همبستگی بین تمام ویژگی‌ها (شامل قیمت) را محاسبه کنید. هر دو ماتریس را به‌صورت مناسب (مثلاً جدول یا نمودار) نمایش دهید و توضیح دهید از نظر عددی چه‌قدر با هم متفاوت به نظر می‌رسند.

ج) بر اساس ماتریس همبستگی، چند ویژگی را که بیشترین همبستگی مثبت یا منفی را با قیمت دارند مشخص کنید و در چند جمله تحلیل کنید که این روابط از نظر شهودی و منطقی چه معنایی برای بازار مسکن کالیفرنیا دارند.

د) یک یا دو زوج ویژگی پیدا کنید که کوواریانس آن‌ها بزرگ است اما همبستگی آن‌ها چندان بزرگ نیست (یا برعکس) و توضیح دهید چگونه مقیاس و پراکندگی داده‌ها می‌تواند باعث شود که تفسیر کوواریانس به‌تنهایی گمراه‌کننده باشد.

هـ) توضیح دهید اگر واحد اندازه‌گیری یکی از ویژگی‌ها تغییر کند (مثلاً از کیلومتر به متر یا ضرب در یک ثابت)، چه اتفاقی برای کوواریانس و چه اتفاقی برای همبستگی می‌افتد و در تحلیل روابط بین ویژگی‌ها ترجیح می‌دهید بیشتر به کدام‌یک تکیه کنید و چرا.

سؤال ۲ – مقایسه‌ی انواع فاصله‌ها در فضای چندبعدی دیتاست Breast Cancer

در این سؤال می‌خواهید ببینید انتخاب نوع فاصله، چگونه برداشت شما از «شباهت» بین نمونه‌های پزشکی را تغییر می‌دهد. داده‌ی Breast Cancer در sklearn شامل ویژگی‌های عددی استخراج‌شده از تصاویر توده‌های خوش‌خیم و بدخیم است.

الف) دیتاست Breast Cancer را از sklearn بارگذاری کنید و فقط ویژگی‌های عددی ورودی و برچسب کلاس را نگه دارید. شکل کلی داده (تعداد ویژگی‌ها، تعداد نمونه‌ها) را گزارش کنید و چند سطر اول را مرور کنید تا یک حس اولیه از مقادیر داشته باشید.

ب) چند نمونه از کلاس «benign» و چند نمونه از کلاس «malignant» انتخاب کنید (مثلاً ۳ نمونه از هر کلاس) و آن‌ها را به‌عنوان نمونه‌های مرجع برای مقایسه نگه دارید. برای همه‌ی جفت‌های این نمونه‌ها، فاصله‌های Euclidean ، Manhattan، Chebyshev، Cosine و Mahalanobis را محاسبه کنید و نتایج را در یک جدول خلاصه کنید تا بتوانید مقایسه‌ای عددی انجام دهید.

ج) بر اساس این جدول، توضیح دهید که هر یک از فاصله‌ها چگونه رفتار می‌کنند؛ برای نمونه روشن کنید که فاصله‌ی اقلیدسی و منهتن چگونه به اندازه‌ی مطلق ویژگی‌ها حساس هستند، فاصله‌ی کسینوسی چگونه بیشتر به «جهت» بردار ویژگی‌ها اهمیت می‌دهد تا به «مقیاس»، و فاصله‌ی ماهالانوبیس چگونه با استفاده از ماتریس کوواریانس، پراکندگی و همبستگی بین ویژگی‌ها را در نظر می‌گیرد.

د) یک سناریو طراحی و توصیف کنید که در آن یک سیستم هشدار اولیه قرار است فقط با استفاده از فاصله، نمونه‌های مشکوک به بدخیمی را علامت‌گذاری کند. توضیح دهید در چنین شرایطی ترجیح می‌دهید از کدام فاصله استفاده کنید، انتخاب هر فاصله چه مزایا و چه ریسک‌هایی دارد، و چرا استفاده‌ی ناآگاهانه از فاصله‌ی اقلیدسی در داده‌ی پزشکی می‌تواند منجر به تصمیم‌های اشتباه و خطرناک شود.

هـ) در چند جمله، تفاوت مفهومی بین «مقیاس» و «جهت» در بردارهای ویژگی را بیان کنید و با استناد به نتایج خود نشان دهید در داده‌ی Breast Cancer این دو مفهوم چگونه روی مقدار فاصله‌ها اثر می‌گذارند.

سؤال ۳ – کاهش بعد و تحلیل مؤلفه‌های اصلی در دیتاستDiabetes

در این تمرین می‌خواهید ببینید روش PCA چگونه می‌تواند ساختار پنهان بین ویژگی‌های عددی دیتاست Diabetes را آشکار کند و در عین حال بعد داده را کاهش دهد.

الف) دیتاست Diabetes را از sklearn بارگذاری کنید و فقط ویژگی‌های عددی ورودی را (بدون ستون هدف) نگه دارید. داده را استاندارد کنید تا هر ویژگی میانگین صفر و انحراف معیار یک داشته باشد و خلاصه‌ای از داده استانداردشده (میانگین و انحراف معیار هر ستون) را گزارش کنید.

ب) روش PCA را روی داده‌ی استانداردشده اجرا کنید و برای هر مؤلفه‌ی اصلی، مقدار واریانس توضیح داده شده و واریانس تجمعی را استخراج کنید. نتایج را در قالب یک جدول یا نمودار نمایش دهید و مشخص کنید برای پوشش تقریباً چه درصدی از واریانس (مثلاً ٪۸۰ یا ٪۹۰) به چند مؤلفه نیاز است.

ج) بار (loading) هر ویژگی روی چند مؤلفه‌ی اول (مثلاً ۲ یا ۳ مؤلفه‌ی اول) را بررسی کنید و توضیح دهید این مؤلفه‌ها چه ترکیب‌هایی از ویژگی‌ها را نمایندگی می‌کنند؛ برای مثال می‌توانید توضیح دهید که مؤلفه‌ی اول بیشتر ترکیبی از چه نوع عوامل (مثل مقادیر مرتبط با قند خون، فشار، شاخص‌های بدنی و...) است و از روی آن چه تفسیر کلی می‌توان داشت.

د) با استفاده از دو مؤلفه‌ی اولPCA ، یک نمایش دوبعدی از داده ایجاد کنید (مثلاً با رنگ‌بندی براساس مقدار هدف یا یک تقسیم‌بندی ساده) و به‌صورت کیفی توضیح دهید آیا در این فضا الگوهای معنی‌داری مثل نواحی فشرده‌تر یا نواحی پراکنده‌تر دیده می‌شود یا خیر؛ نیازی به استفاده از مفاهیم پیشرفته تری مثل خوشه‌بندی ندارید، فقط توصیف بصری و شهودی کافی است.

هـ) در پایان، در چند پاراگراف کوتاه مزایا و معایب استفاده از فضای کاهش‌یافته‌ی PCA به‌عنوان ورودی یک مدل یادگیری را جمع‌بندی کنید؛ توضیح دهید حذف مؤلفه‌های با واریانس پایین‌تر چه نوع اطلاعاتی را ممکن است از بین ببرد و در مقابل، کاهش بعد چه کمکی به ساده‌تر شدن مدل و کاهش نویز می‌کند.

سؤال ۴ – فشرده‌سازی و بازسازی تصاویر چهره با PCA در دیتاست LFW

در این سؤال با استفاده از دیتاست چهره‌های «Labeled Faces in the Wild (LFW)» در sklearn ، باید توانایی PCA را در فشرده‌سازی و بازسازی تصاویر بررسی کنید و به‌صورت کیفی نقش مؤلفه‌های اصلی را در نمایش ساختار چهره تحلیل نمایید.

الف) دیتاست LFW را از sklearn بارگذاری کنید و زیرمجموعه‌ای از افراد را که تعداد مناسبی تصویر دارند انتخاب کنید. تصاویر را به شکل بردارهای عددی در یک ماتریس داده قرار دهید و ابعاد این ماتریس (تعداد تصاویر و تعداد پیکسل در هر تصویر) را گزارش کنید.

ب) روش PCA را روی این ماتریس اجرا کنید و چند مؤلفه‌ی اول مثلاً PC1 تا PC4 را به‌صورت تصویر نمایش دهید .(eigenfaces) در چند جمله توضیح دهید هر یک از این مؤلفه‌ها چه الگوی بصری کلی را نشان می‌دهند؛ برای نمونه می‌توانید به تغییرات روشنایی، نواحی پررنگ‌تر صورت یا شکل کلی مرز چهره اشاره کنید.

ج) یکی از تصاویر را انتخاب کنید و آن را با تعداد مؤلفه‌های مختلف مثلاً k های **۲۰ و ۸۰ و ۱۵۰** بازسازی کنید. تصاویر بازسازی‌شده را کنار هم قرار دهید و از نظر وضوح، جزئیات و شباهت به تصویر اصلی مقایسه کنید. اگر مایلید می‌توانید یک معیار کمی ساده مثل میانگین مربع خطا (MSE) نیز برای هر k محاسبه و گزارش کنید.

د) یکی از تصاویر را نویزی کنید (مثلاً با اضافه کردن نویز گاوسی به پیکسل‌ها) و سپس با استفاده از مؤلفه‌های اصلی، نسخه‌ی بازسازی‌شده‌ی این تصویر نویزی را تولید کنید. سه تصویر «اصلی»، «نویزی» و «بازسازی‌شده» را کنار هم نمایش دهید و توضیح دهید PCA تا چه حد توانسته نویز را کاهش دهد و چرا مؤلفه‌های اصلی اولیه بیشتر ساختار کلی چهره را نگه می‌دارند و مؤلفه‌های رده‌پایین‌تر بیشتر حامل نویز و جزئیات ناپایدار هستند.

در نهایت در چند جمله محدودیت‌های استفاده از PCA را در کاربردهای واقعی تشخیص چهره بیان کنید؛ به‌خصوص توضیح دهید اگر تعداد مؤلفه‌ها را بیش از حد کم انتخاب کنیم، چه نوع اطلاعات مهمی ممکن است از دست برود، حتی اگر تصویر بازسازی‌شده در نگاه اول «قابل قبول» به نظر برسد.