گزارش تمرین دیتا ماینینگ :**بهنام بشیری ۴۰۳۲۱۴۰۰۲**

تمرین یک:

هدف :این کد برای بارگذاری و بررسی اولیه داده‌ها استفاده می‌شود. هدف این است که با استفاده از pandas اطلاعات پایه‌ای مانند تعداد سطرها و ستونها، اطلاعات اولیه از داده‌ها و تعداد مقادیر یکتای هر ستون را نمایش دهد.

مراحل کد:

۱. **بارگذاری داده‌ها** :ابتدا داده‌ها از فایل CSV بارگذاری می‌شوند.
۲. **اطلاعات کلی دیتافریم** :اطلاعاتی مانند تعداد سطرها، ستونها و اطلاعات مربوط به نوع داده‌ها (عدد یا رشته و ...) به‌طور کلی نمایش داده می‌شود.
۳. **نمایش ۵ ردیف اول** :برای آشنایی اولیه با داده‌ها، پنج ردیف اول دیتافریم نمایش داده می‌شود.
۴. **نمایش تعداد مقادیر یکتا** :برای هر ستون تعداد مقادیر یکتا نمایش داده می‌شود تا نوع داده‌های موجود در هر ستون بهتر شناخته شود.

توضیحات:

- مقادیر یکتا در این کد به تعداد انواع مختلف داده در هر ستون اشاره دارد. به عنوان مثال اگر ستون "Gender" دو مقدار "Male"و "Female" داشته باشد، تعداد مقادیر یکتا برای این ستون برابر با ۲ خواهد بود.

تمرین دو:

هدف :هدف این کد شناسایی و بررسی مقادیر گمشده در دیتافریم است. این کار با استفاده از متد ()isnull انجام می‌شود که مقادیر گمشده را شناسایی کرده و با متد ()sum تعداد آنها در هر ستون شمارش می‌کند.

مراحل کد:

۱. **بازخوانی داده‌ها** :داده‌ها دوباره از فایل CSV بارگذاری می‌شوند تا مطمئن شویم که دیتافریم (df) به درستی تعریف شده است.
۲. **بررسی مقادیر گمشده** :تعداد مقادیر گمشده در هر ستون با استفاده از ()isnull().sum شمارش می‌شود.
۳. **فیلتر کردن ستونهای دارای مقادیر گمشده** :تنها ستونهایی که دارای مقادیر گمشده هستند، نمایش داده می‌شوند.
۴. **پیام‌های مختلف** :اگر مقادیر گمشده وجود داشته باشد، آنها به همراه نام ستونها نمایش داده می‌شود. در غیر این صورت، پیامی مبنی بر عدم وجود مقادیر گمشده چاپ می‌شود.

توضیحات:

- مقادیر گمشده به معنای مقادیری هستند که در یک سطر خاص برای یک ستون وجود ندارند و معمولاً با NaN نشان داده می‌شوند.

تمرین سه:

هدف :هدف این کد حذف ستون id از دیتافریم است. در برخی موارد ممکن است نیاز باشد که ستونهایی که برای تحلیل یا مدلسازی اهمیت ندارند، مانند شناسه‌ها(IDs) ، حذف شوند تا تمرکز روی داده‌های واقعی افزایش یابد.

مراحل کد:

۱. **بارگذاری داده‌ها** :داده‌ها از فایل CSV بارگذاری می‌شوند تا به دیتافریم تبدیل شوند.
۲. **حذف ستون :id** در این بخش، ستون id با استفاده از متد ()drop از دیتافریم حذف می‌شود.
۳. **نمایش نتایج** :بعد از حذف ستون، لیست نام ستونهای باقی‌مانده با استفاده از ()df.columns.tolist نمایش داده می‌شود.

توضیحات:

- حذف ستونها در پردازش داده‌ها امری رایج است، به‌ویژه وقتی که ستونها هیچ اطلاعاتی برای تحلیل ندارند.

تمرین چهار:

هدف :این کد برای پردازش و تحلیل داده‌های مربوط به نمرات CGPA در دیتافریم استفاده می‌شود. هدف اصلی این است که نمرات CGPA دانشجویان را به ۳ گروه مختلف دسته‌بندی کرده و سپس ستونهای اضافی که به تحلیل‌ها کمک نمی‌کنند، حذف شوند.

مراحل کد:

۱. **بارگذاری داده‌ها** :داده‌ها از فایل CSV به دیتافریم df بارگذاری می‌شوند.
۲. **حذف ستون :id** ستون id حذف می‌شود زیرا برای تحلیل‌های بعدی نیازی به آن نداریم.
۳. **دسته‌بندی نمرات :CGPA** از تابع qcut برای تقسیم نمرات CGPA به ۳ گروه مساوی استفاده می‌شود.
۴. **شمارش تعداد دانشجویان در هر گروه** :پس از دسته‌بندی، تعداد دانشجویانی که در هر گروه قرار دارند با استفاده از ()value_counts شمارش می‌شود.
۵. **حذف ستون :CGPA** پس از انجام دسته‌بندی، ستون اصلی نمرات CGPA حذف می‌شود.

توضیحات:

- تابع qcut در pandas برای تقسیم داده‌ها به گروه‌هایی با اندازه‌های تقریباً برابر استفاده می‌شود.
- drop(columns=["id"])برای حذف ستون id از دیتافریم استفاده می‌شود.

تمرین پنج:

بخش ۱:

هدف :هدف این کد، تحلیل داده‌ها با استفاده از نمودار هیستوگرام است. این کد ویژگی‌های عددی موجود در دیتافریم را شناسایی کرده و برای هر یک هیستوگرام رسم می‌کند تا توزیع مقادیر آنها بررسی شود.

مراحل کد:

۱. **بارگذاری داده‌ها** :داده‌ها از فایل CSV بارگذاری می‌شوند و به یک دیتافریم به نام df تبدیل می‌شوند.
۲. **حذف ستون :id** ستون id که برای شناسایی یکتای دانشجویان استفاده می‌شود، حذف می‌شود.
۳. **شناسایی ویژگی‌های عددی و متنی** :ویژگی‌های عددی و متنی شناسایی می‌شوند.
۴. **رسم هیستوگرام** :برای هر ویژگی عددی، هیستوگرام با تعداد bins معین رسم می‌شود.
۵. **تنظیمات نمایش نمودار** :با استفاده از ()plt.suptitle عنوان برای نمودار اضافه می‌شود.

توضیحات:

- هیستوگرام‌ها برای نمایش توزیع داده‌ها استفاده می‌شوند.
- bins=۳۰برای تقسیم داده‌ها به ۳۰ بخش استفاده می‌شود.

بخش ۲:

هدف :هدف این کد رسم نمودارهای میله‌ای (Bar Charts) برای ویژگی‌های متنی (categorical) است.

مراحل کد:

۱. **پیمایش ویژگی‌های دسته‌ای** :برای هر ویژگی متنی، یک نمودار میله‌ای رسم می‌شود.
۲. **رسم نمودار میله‌ای** :از متد ()value_counts برای شمارش تعداد وقوع هر مقدار استفاده می‌شود.
۳. **تنظیمات نمودار** :تنظیماتی برای عنوان برچسب‌ها و چیدمان نمودار اعمال می‌شود.

توضیحات:

- نمودار میله‌ای برای نمایش توزیع مقادیر یک ویژگی دسته‌ای استفاده می‌شود.

بخش ۳:

هدف :این کد برای شناسایی و نمایش مقادیر پرت (outliers) در ویژگی‌های عددی مختلف دیتافریم استفاده می‌شود.

مراحل کد:

۱. **تعریف تابع :detect_outliers_IQR** مقادیر پرت با استفاده از IQR شناسایی می‌شوند.
۲. **شناسایی مقادیر پرت برای هر ویژگی** :مقادیر پرت با استفاده از تابع detect_outliers_IQR شناسایی می‌شوند.
۳. **چاپ اطلاعات** :تعداد مقادیر پرت و نمونه‌هایی از آنها چاپ می‌شود.
۴. **رسم نمودار :Boxplot** برای هر ویژگی عددی، یک نمودار Boxplot رسم می‌شود.

توضیحات:

- IQR (Interquartile Range)برای شناسایی مقادیر پرت استفاده می‌شود.
- Boxplotنموداری برای نمایش مقادیر پرت است.

تمرین شش:

بخش ۱:

هدف :این کد برای تقسیم داده‌ها به دو مجموعه‌ی آموزشی (train) و آزمایشی (test) با استفاده از Stratified Samplingاست.

مراحل کد:

۱. **بارگذاری داده‌ها** :داده‌ها از فایل CSV بارگذاری می‌شوند و ستون id حذف می‌شود.
۲. **تقسیم داده‌ها با Stratified Sampling** :داده‌ها به دو بخش آموزشی و آزمایشی تقسیم می‌شوند.
۳. **چاپ نتایج** :اندازه مجموعه‌های آموزشی و آزمایشی نمایش داده می‌شود.

توضیحات:

- Stratified Samplingبرای حفظ توزیع داده‌ها در مجموعه‌های آموزشی و آزمایشی مفید است.

بخش ۲:

هدف :این کد برای تقسیم داده‌ها به دو مجموعه‌ی آموزشی (train) و آزمایشی (test) به‌طور تصادفی است.

مراحل کد:

۱. **تقسیم داده‌ها به‌صورت تصادفی** :داده‌ها به‌صورت تصادفی به دو بخش تقسیم می‌شوند.
۲. **چاپ نتایج** :توزیع جنسیت در داده‌های آموزشی و آزمایشی نمایش داده می‌شود.

توضیحات:

- در این روش، توزیع ویژگی‌ها در مجموعه‌های آموزشی و آزمایشی تغییر می‌کند.

بخش ۳:

هدف :این کد برای مقایسه درصد جنسیت‌ها در مجموعه‌های داده مختلف استفاده می‌شود.

مراحل کد:

۱. **تقسیم داده‌ها به دو روش مختلف** :داده‌ها به دو روش Stratified Sampling و Random Sampling تقسیم می‌شوند.
۲. **محاسبه درصد جنسیت‌ها** :درصد جنسیت‌ها برای هر مجموعه محاسبه می‌شود.
۳. **ساخت و نمایش جدول مقایسه‌ای** :یک جدول مقایسه‌ای برای نمایش درصد جنسیت‌ها ساخته می‌شود.

توضیحات:

- این مقایسه

کمک می‌کند تا انتخاب روش تقسیم‌بندی داده‌ها بر اساس ویژگی‌های مهم نظیر جنسیت، بهتر انجام شود.

تمرین هفت:

هدف :هدف این کد، بررسی روابط میان ویژگی‌های مختلف داده‌ها و شبیه‌سازی یک مدل پیش‌بینی با استفاده از الگوریتم‌های مختلف است.

مراحل کد:

۱. **انتخاب ویژگی‌ها** :ویژگی‌هایی که قرار است برای پیش‌بینی استفاده شوند، انتخاب می‌شوند.
۲. **ساخت مدل** :از الگوریتم‌های مختلف مثل Regression یا Classification برای ساخت مدل استفاده می‌شود.
۳. **آموزش مدل** :مدل با استفاده از داده‌های آموزشی آموزش داده می‌شود.
۴. **ارزیابی مدل** :پس از آموزش، مدل با استفاده از داده‌های آزمایشی ارزیابی می‌شود و نتایج پیش‌بینی بررسی می‌شود.

توضیحات:

- الگوریتم‌های مختلف بسته به نوع مشکل (رجعتی یا طبقه‌بندی) انتخاب می‌شوند.
- ارزیابی مدل معمولاً شامل مقایسه پیش‌بینی‌ها با مقادیر واقعی است.

تمرین هشت:

هدف :این تمرین به تحلیل ویژگی‌ها و وابستگی‌های بین آنها می‌پردازد و از تکنیک‌های مختلف آماری برای تحلیل داده‌ها استفاده می‌کند.

مراحل کد:

۱. **تحلیل همبستگی** :از روش‌های آماری مانند همبستگی پیرسون برای بررسی روابط میان ویژگی‌های مختلف استفاده می‌شود.
۲. **رسم ماتریس همبستگی** :ماتریس همبستگی رسم می‌شود تا وابستگی‌های مختلف بین ویژگی‌ها به‌طور واضح نمایش داده شود.
۳. **بررسی نتایج** :نتایج تحلیل همبستگی می‌تواند به تصمیم‌گیری برای انتخاب ویژگی‌های مهم برای مدل کمک کند.

توضیحات:

- تحلیل همبستگی به شناسایی روابط خطی میان ویژگی‌ها کمک می‌کند. این تحلیل می‌تواند برای کاهش ابعاد داده یا انتخاب ویژگی‌ها مفید باشد.

تمرین نه:

هدف :این کد به شبیه‌سازی مدل‌های طبقه‌بندی (مثل KNN یا Random Forest) و پیش‌بینی داده‌ها پرداخته و سپس نتایج را ارزیابی می‌کند.

مراحل کد:

۱. **تقسیم داده‌ها به مجموعه آموزشی و آزمایشی** :داده‌ها به دو بخش تقسیم می‌شوند تا مدل بر اساس داده‌های آموزشی آموزش داده شود.
۲. **آموزش مدل طبقه‌بندی** :از الگوریتم‌هایی مانند KNN یا Random Forest برای ساخت مدل استفاده می‌شود.

۳. **ارزیابی مدل** :دقت مدل و سایر متریک‌های ارزیابی مانندF۱-Score ، Precisionو Recall بررسی می‌شود.
۴. **مقایسه با مدل‌های دیگر** :نتایج مدل‌های مختلف با یکدیگر مقایسه می‌شوند تا بهترین مدل انتخاب شود.

توضیحات:

- مدل‌های طبقه‌بندی معمولاً برای پیش‌بینی کلاس‌های مختلف (مثلاً "بله" یا "خیر") استفاده می‌شوند.
- ارزیابی با استفاده از معیارهای مختلف دقت پیش‌بینی را اندازه‌گیری می‌کند.

تمرین ده:

هدف :هدف این تمرین، پیاده‌سازی و ارزیابی مدل‌های رگرسیونی (مثل Linear Regression یا Decision Tree Regressor) برای پیش‌بینی یک متغیر عددی است.

مراحل کد:

۱. **آماده‌سازی داده‌ها** :داده‌ها پیش‌پردازش شده و ویژگی‌های مربوط به مدل انتخاب می‌شوند.
۲. **ساخت مدل رگرسیونی** :از الگوریتم‌های مختلف رگرسیون (مثل Linear Regression یا Decision Tree Regressor) استفاده می‌شود.
۳. **آموزش مدل** :مدل با استفاده از داده‌های آموزشی آموزش داده می‌شود.
۴. **ارزیابی مدل** :مدل با استفاده از معیارهایی مانند Mean Squared Error (MSE) یا R^2 ارزیابی می‌شود.

توضیحات:

- در رگرسیون، هدف پیش‌بینی یک مقدار پیوسته است (مثل قیمت خانه).
- معیارهای ارزیابی مانند MSE به سنجش دقت پیش‌بینی‌های مدل کمک می‌کند.