هدف این تمرین ارزیابی عملکرد روش‌های دسته‌بندی دوگانه و چندگانه است. برای این کار مسالهٔ پایش سلامت و تشخیص عیب ابزار برش یک دستگاه فرز در نظر گرفته شده است.

دادگان مورد نظر (فایل milling_machine.csv) شامل ۱۰۰۰ داده است که هر داده، وضعیت ابزار فرز (ستون ششم) را بر اساس پنج ویژگی ("دمای هوا"، "دمای فرآیند"، "سرعت چرخشی ابزار فرز"، "گشتاور وارد به محور ابزار" و "مدت زمان قرارگیری ابزار در معرض سایش") (ستون‌های اول تا پنجم) نشان می‌دهد.

الف) بررسی داده‌های خام

برای آشنایی بهتر با دادگان مورد نظر:

الف-۱) ساختار کلی داده‌ها را با روش‌های info و describe بدست آورید.

الف-۲) برای هر ویژگی، تعداد و نسبت مقادیر ناموجود (missing values) را بدست آورید.

الف-۳) correlation دو به دو ویژگی‌ها را بررسی کنید و با تشکیل ماتریس همبستگی تعیین کنید که وضعیت ابزار فرز به کدام ویژگی‌ها وابستگی بیشتری دارد.

الف-۴) برای سه ویژگی دارای بیشترین تاثیر بر خروجی (بر پایهٔ بند الف-۳) نمودار تعداد مشاهدات هر مقدار منحصر به فرد را رسم کنید.

ب) پیش پردازش داده‌ها

در این بخش لازم است کاستی‌های احتمالی موجود در دادگان (مانند مقادیر خارج از بازهٔ مجاز یا مقادیر ناموجود یا داده‌های پرت) که می‌تواند بر همگرایی و تعمیم‌پذیری مدل تاثیر منفی بگذارد برطرف شود.

ب-۱) ابتدا بررسی کنید که کدام داده‌ها بیشترین میزان مقادیر ناموجود (missing values) را دارند و سپس با توجه به توضیحاتی که در ویدیوی تکمیلی در این باره داده شده است مشکل مقادیر ناموجود را برای همهٔ ویژگی‌ها حل کنید (با ذکر روش بکار رفته و دلیل انتخاب آن).

ب-۲) فرآیندهای standardizing و normalizing را (برای ویژگی‌های کمی) توضیح دهید. آیا در این تمرین نیاز به این فرآیندها هست؟ اگر نیاز به این کار هست آن را اعمال کنید.

ج) دسته‌بندی دوگانه

هدف این بخش دسته‌بندی دوگانهٔ داده‌هاست به گونه‌ای که بتوان سالم یا معیوب بودن ابزار را تشخیص داد.

ج-۱) ابتدا در محیط پایتون یک ستون به ستون‌های فایل دادگان اضافه کنید که محتوای آن برای ابزارهای سالم برچسب No Failure و برای ابزارهای آسیب دیده برچسب Failure باشد. به این ترتیب داده‌ها به دو دستهٔ سالم و معیوب تقسیم می‌شوند که اگر هدف خود را این ستون جدید قرار دهید یک مسئلهٔ دسته‌بندی دوگانه خواهید داشت.

ج-۲) با رسم نمودار میله‌ای (Chart Bar) برای توزیع دسته‌های دوگانهٔ جدید، عدم توازن احتمالی داده‌ها رانشان دهید.

ج-۳) توضیح دهید که عدم توازن داده‌ها چه مشکلی برای مدل ایجاد می‌کند.

ج-۴) با استفاده از روش‌های متوازن‌سازی داده‌ها (مانند smote) مشکل یاد شده را برطرف کنید.

ج-۵) دادگان پالایش شده را به صورت تصادفی به دو بخش آموزش (۸۰٪) و آزمون (۲۰٪) تقسیم کنید(Random_State = 42) و با استفاده از کتابخانه scikit-learn مدل‌های زیر را آموزش دهید:

Logistic Regression
K-Nearest-Neighbors
Support Vector Machine

* برای مدل SVM هسته‌های خطی و غیرخطی را بکار ببرید و تفاوت احتمالی نتایج را تفسیر کنید.

ج-۶) با تشکیل ماتریس آشفتگی (confusion matrix) ، دقت (accuracy) و Classification Report هر مدل را بدست آورید و نتایج را در جدولی ارایه کنید.

ج-۷) برای مدل‌های یاد شده دو پارامتر را از میان هایپرپارامترها انتخاب کرده و آن‌ها را تغییر دهید (در مورد مدل KNN تنها مقدار K را تغییر دهید) و به کمک تابع GridSearchCV مقادیر بهینهٔ پارامترها را (که دقت مدل را بیشینه می‌کند) بدست آورید.

ج-۸) به کمک شاخص‌های معرفی شده در درس، عملکرد مدل‌های مختلف را با یکدیگر مقایسه کنید.

د) دسته‌بندی چندگانه

هدف این بخش دسته‌بندی چندگانهٔ داده‌هاست به گونه‌ای که بتوان سالم بودن یا نوع آسیب ابزار را تشخیص داد. در این بخش هدف شما همان ستون Failure Types خواهد بود.

د-۱) دادگان پالایش شده را به صورت تصادفی به دو بخش آموزش (۸۰٪) و آزمون (۲۰٪) تقسیم کنید(Random_State = 42) و با استفاده از کتابخانه scikit-learn مدل‌های زیر را آموزش دهید:

K-Nearest-Neighbors
Decision Tree

Random Forest
Support Vector Machine

* برای مدل SVM، با استفاده از روش‌های "یکی در برابر یکی" یا "یکی در برابر همه" دسته‌بندی چندگانه را انجام دهید.

د-۲) با تشکیل ماتریس آشفتگی (confusion matrix) ، دقت (accuracy) و Classification Report هر مدل را بدست آورید و نتایج را در جدولی ارایه کنید.

د-۳) برای مدل‌های ذکر شده دو پارامتر را از میان هایپرپارامترها انتخاب کرده و آن‌ها را تغییر دهید (در مورد مدل KNN تنها مقدار K را تغییر دهید) و به کمک تابع GridSearchCV مقادیر بهینهٔ پارامترها را (که دقت مدل را بیشینه می‌کند) بدست آورید..

د-۴) به کمک شاخص‌های معرفی شده در درس، عملکرد مدل‌های مختلف را با یکدیگر مقایسه کنید.

چند توضیح :

- ✓ برای یادگیری مفاهیمی که در تمرین مطرح شده و در کلاس تدریس نشده‌اند از منابع موجود در اینترنت استفاده کنید.
- ✓ برای انجام بخش‌های مختلف تمرین می‌توانید از کتابخانه‌های آماده‌ای مانند pandas ، matplotlib، numpy ، sklearn و seaborn استفاده کنید .
- ✓ تحویل گزارش این تمرین ضروری است و به تمرین بدون گزارش نمره‌ای تعلق نمی‌گیرد. حجم گزارش معیاری برای ارزیابی نخواهد بود و لزومی به توضیح جزئیات کد نیست؛ اما از آنجا که برای این تمرین از کتابخانه‌های موجود استفاده می‌کنید لطفا تمامی پارامترهای تنظیم‌شده در هر قسمت از کد را گزارش کرده و فرض‌هایی را که برای پیاده‌سازی‌ها و محاسبات خود به کار برده‌اید ذکر کنید. از ارائهٔ توضیحات کلیشه‌ای و همانند بَرداری از منابع موجود بپرهیزید.
- ✓ در فرایند ارزیابی گزارش، کدهای شما لزوما اجرا نخواهد شد. بنابراین همهٔ نتایج و تحلیل‌های خود را به‌طور کامل ارائه کنید.

- ✓ شباهت بیش از حد گزارش و کدها باعث از دست دادن نمرهٔ تمرین خواهد شد. همچنین گزارش‌هایی که در آنها از کدهای آماده استفاده شده باشد پذیرفته نخواهند شد.
- ✓ گزارش شما باید به صورت تایپ شده و با فرمت pdf ارائه شود و کدهایی که به همراه گزارش تحویل می‌دهید باید قابل اجرا باشند. در انتها تمامی فایل‌های لازم را در یک فایل zip یا rar بارگذاری و ارسال کنید.
- ✓ در صورت استفاده از گیت هاب جهت ارائهٔ گزارش و کد، نمره امتیازی به شخص تعلق میگیرد.
- ✓ پرسش‌های خود را از طریق ایمیل یا تلگرام از دستیار آموزشی مربوطه بپرسید:

	ایمیل	تلگرام
امیرحسین کمیجانی	Ah1379.k@gmail.com	amirhossein_komi