مبانی مدل‌های یادگیری ماشین

دیتاست موردنیاز این جلسه:

student_performance_session2.csv

ستون‌های دیتاست:

- StudentID: شناسه دانشجو
- StudyHours: تعداد ساعت مطالعه
- AttendancePercent: درصد حضور در کلاس
- PracticeTests: تعداد آزمون‌های تمرینی
- PreviousScore: نمره قبلی دانشجو
- FinalScore: نمره نهایی
- Passed: قبولی یا عدم قبولی

سؤال ۱ — تشخیص نوع مسئله یادگیری ماشین

برای هر مسئله مشخص کنید از کدام نوع یادگیری استفاده می‌شود:

- یادگیری نظارت‌شده — رگرسیون
- یادگیری نظارت‌شده — طبقه‌بندی
- یادگیری بدون نظارت
- یادگیری تقویتی

۱. پیش‌بینی نمره نهایی یک دانشجو
۲. پیش‌بینی قبولی یا عدم قبولی دانشجو
۳. گروه‌بندی دانشجویان براساس الگوی مطالعه، بدون وجود برچسب قبلی
٤. آموزش یک عامل هوشمند برای انتخاب تمرین مناسب با دریافت پاداش و جریمه

سپس پاسخ دهید:

۱. تفاوت اصلی Regression و Classification چیست؟
۲. کدام ستون این دیتاست می‌تواند Target یک مسئله رگرسیون باشد؟
۳. کدام ستون می‌تواند Target یک مسئله طبقه‌بندی باشد؟

سؤال ۲ — مراحل پروژه یادگیری ماشین

مراحل زیر را به ترتیب صحیح مرتب کنید:

- ارزیابی مدل
- جمع‌آوری داده
- آموزش مدل
- استقرار مدل
- پاک‌سازی داده
- تقسیم داده به Training وTest
- مهندسی ویژگی

سپس پاسخ دهید:

۱. چرا داده باید قبل از آموزش مدل پاک‌سازی شود؟
۲. چرا مدل نباید با داده‌های Test آموزش ببیند؟
۳. اگر عملکرد مدل مناسب نبود، دو اقدام برای بهبود پروژه پیشنهاد دهید.
٤. آیا آموزش مدل آخرین مرحله پروژه یادگیری ماشین است؟ دلیل بنویسید.

سؤال ۳ — شناخت مسئله و انتخابFeature

دیتاست را با Pandas بخوانید و پنج سطر اول و اطلاعات کلی آن را نمایش دهید.

مسئله اول: پیش‌بینی FinalScore

۱. ستون Target را مشخص کنید.
۲. ستون‌های عددی دیتاست را انتخاب کنید.
۳. همبستگی ستون‌های عددی را با FinalScore محاسبه کنید.
٤. ستون‌ها را براساس شدت رابطه آن‌ها با FinalScore بررسی کنید.
٥. برای سه ستونی که به نظر می‌رسد رابطه بیشتری با FinalScore دارند، Scatter Plot رسم کنید.
٦. با توجه به موارد زیر، سه Feature مناسب انتخاب کنید:
   - مقدار همبستگی آن‌ها با FinalScore
   - شکل رابطه درScatter Plot
   - منطقی‌بودن ارتباط آن ستون با نمره نهایی
۷. برای هر Feature انتخاب‌شده، یک دلیل کوتاه بنویسید.
۸. مشخص کنید این مسئله Regression است یا Classification و دلیل خود را توضیح دهید.

- **نکات انتخاب Feature**

در انتخاب Feature به این موارد توجه کنید:

- StudentID فقط شناسه دانشجو است و معمولاً اطلاعات مفیدی برای پیش‌بینی ندارد.
- خود ستون FinalScoreنباید به‌عنوان Feature انتخاب شود، چون Target مسئله است.

- ستون Passed ممکن است از روی FinalScore ساخته شده باشد؛ بنابراین استفاده از آن می‌تواند باعث نشت اطلاعات شود.
- زیادبودن همبستگی به‌تنهایی کافی نیست؛ ارتباط ستون با مسئله نیز باید منطقی باشد.

سؤال ۴ — تقسیم داده‌های رگرسیون

برای پیش‌بینی :FinalScore

- ستون StudyHours را به‌عنوان Feature انتخاب کنید.
- ستون FinalScore را به‌عنوان Target انتخاب کنید.
- داده‌ها را با test_size=0.2 و random_state=42 به بخش‌های Training و Test تقسیم کنید.

سپس:

۱. X_train، X_test، y_train و y_test را نمایش دهید.
۲. چند درصد داده‌ها برای آموزش استفاده شده‌اند؟
۳. چند درصد داده‌ها برای ارزیابی باقی مانده‌اند؟
٤. مدل کدام بخش از داده‌ها را هنگام آموزش مشاهده می‌کند؟
٥. نقش random_state=42 چیست؟

سؤال ۵— بررسی رابطه قبل از ساخت مدل

بین ستون‌های StudyHours و FinalScore یک Scatter Plot رسم کنید.

سپس براساس نمودار پاسخ دهید:

۱. هر نقطه در نمودار نماینده چیست؟
۲. محور افقی و محور عمودی مربوط به کدام ستون‌ها هستند؟
۳. با افزایش ساعت مطالعه، نمره نهایی معمولاً افزایش می‌یابد یا کاهش؟
٤. آیا رابطه مشاهده‌شده تقریباً خطی است؟
٥. آیا تمام نقاط دقیقاً روی یک خط قرار گرفته‌اند؟
٦. آیا استفاده از Linear Regression برای شروع تحلیل این رابطه منطقی است؟ دلیل کوتاهی بنویسید.

سؤال ۶ — آموزش Linear Regression

یک مدل LinearRegression بسازید و آن را با داده‌های Training آموزش دهید.

سپس مقادیر زیر را نمایش دهید:

- coef_
- intercept_

براساس خروجی پاسخ دهید:

۱. علامت ضریب StudyHours مثبت است یا منفی؟
۲. علامت ضریب درباره رابطه ساعت مطالعه و نمره نهایی چه می‌گوید؟
۳. intercept_ در معادله خط چه نقشی دارد؟
٤. در معادله $y = wx + b$، کدام خروجی مدل نشان‌دهنده w است؟
٥. کدام خروجی نشان‌دهنده b است؟

سؤال ۷ — پیش‌بینی و ارزیابی Linear Regression

با استفاده از مدل آموزش‌دیده، برای داده‌های Test پیش‌بینی انجام دهید.

سپس معیارهای زیر را محاسبه کنید:

- MAE
- MSE
- RMSE
- R^2

براساس خروجی پاسخ دهید:

۱. MAE نشان‌دهنده چه چیزی است؟
۲. چرا مقدار کمتر MAE مطلوب‌تر است؟
۳. چرا MSE خطاهای بزرگ را شدیدتر جریمه می‌کند؟
٤. چرا تفسیر RMSE از MSE ساده‌تر است؟
٥. واحد RMSE با واحد کدام متغیر یکسان است؟
٦. مقدار R^2 به یک نزدیک‌تر است یا به صفر؟
۷. مقدار به‌دست‌آمده برای R^2 چه برداشتی از عملکرد مدل ایجاد می‌کند؟
۸. آیا فقط با مشاهده یک معیار می‌توان درباره کیفیت نهایی مدل تصمیم گرفت؟

سؤال ۸ — رسم خط رگرسیون

داده‌های Training را با Scatter Plot نمایش دهید و خط پیش‌بینی‌شده توسط مدل Linear Regression را روی همان نمودار رسم کنید.

سپس پاسخ دهید:

۱. نقاط نمودار چه چیزی را نشان می‌دهند؟
۲. خط رگرسیون چه چیزی را نشان می‌دهد؟
۳. فاصله عمودی هر نقطه از خط بیانگر چیست؟
٤. نقطه‌ای که فاصله بیشتری از خط دارد، خطای پیش‌بینی بزرگ‌تری دارد یا کوچک‌تر؟
٥. آیا عبور نکردن خط از تمام نقاط به معنی اشتباه‌بودن مدل است؟
٦. خط رگرسیون روند کلی داده‌ها را نشان می‌دهد یا تک‌تک مقادیر را حفظ می‌کند؟

سؤال ۹ — درک روش KNN Regression

فرض کنید برای یک دانشجوی جدید، سه دانشجوی نزدیک‌تر دارای نمرات زیر هستند:

- دانشجوی اول: ۴۸
- دانشجوی دوم: ۵۴
- دانشجوی سوم: ۵۷

اگر K=3 باشد:

۱. نمره پیش‌بینی‌شده را محاسبه کنید.
۲. منظور از «نزدیک‌ترین همسایه» چیست؟
۳. فاصله کمتر بین دو مشاهده نشان‌دهنده شباهت بیشتر است یا کمتر؟
٤. Kدر الگوریتم KNN چه چیزی را مشخص می‌کند؟
٥. آیا KNN مانند Linear Regression یک معادله خطی یاد می‌گیرد؟

۶. KNNاطلاعات داده‌های Training را ذخیره می‌کند یا فقط دو پارامتر wو bرا یاد می‌گیرد؟

سؤال ۱۰ — آموزش و مقایسهKNN Regression

با همان داده‌های Training و Test سؤال‌های قبلی، یک مدل KNeighborsRegressor با n_neighbors=5 آموزش دهید.

سپس:

- برای داده‌های Test پیش‌بینی انجام دهید.
- MAE، MSE، RMSEو R^2 را محاسبه کنید.
- نتایج را با Linear Regression مقایسه کنید.

در ادامه مقدار K را به‌ترتیب برابر موارد زیر قرار دهید:

- 1
- 3
- 10
- 20

برای هر مقدار، MAEو R^2 را محاسبه کنید.

سپس پاسخ دهید:

۱. کدام مقدار K کمترین MAE را ایجاد کرده است؟

۲. کدام مقدار K بیشترین R^2 را ایجاد کرده است؟

۳. چرا K=1 ممکن است مدل را به نویز حساس کند؟

٤. چرا K بسیار بزرگ ممکن است الگوهای محلی را از بین ببرد؟

٥. آیا یک مقدار K برای تمام دیتاست‌ها بهترین است؟

٦. Linear Regression و KNN از نظر روش پیش‌بینی چه تفاوتی دارند؟

سؤال ۱۱ Logistic Regression — و پیش‌بینی قبولی

قبل از آموزش مدل، داده‌ها را با test_size=0.2 و random_state=42 به دو بخش Training و Test تقسیم کنید. مدل را فقط با داده‌های Training آموزش دهید و پیش‌بینی کلاس و احتمال کلاس‌ها را برای داده‌های Test انجام دهید.

برای تبدیل ستون Passed به برچسب عددی، از روش زیر استفاده کنید:

- Yes برابر ۱
- No برابر ۰

ستون StudyHours را به‌عنوان Feature انتخاب کنید و مدل LogisticRegression را آموزش دهید.

سپس:

- کلاس‌ها را با predict() پیش‌بینی کنید.
- احتمال کلاس‌ها را با predict_proba() به دست آورید.

براساس خروجی پاسخ دهید:

۱. چرا این مسئله Classification است؟

۲. هر خروجی predict()نشان‌دهنده چیست؟

۳. دو ستون خروجی predict_proba() چه چیزی را نشان می‌دهند؟

٤. اگر احتمال کلاس ۱ برابر 0.72 باشد، کلاس نهایی با آستانه 0.5چیست؟

٥. اگر احتمال کلاس ۱ برابر 0.48 باشد، کلاس نهایی چیست؟

٦. Decision Boundary چه نقشی در طبقه‌بندی دارد؟

۷. چرا Linear Regression برای پیش‌بینی احتمال مناسب نیست؟

۸. چرا Logistic Regression با وجود نامRegression ، یک الگوریتم طبقه‌بندی است؟