لیگ علم داده

کاری از کارگروه کاوش علمی
انجمن هوش مصنوعی دانشگاه خوارزمی

تهیه شده در آذر 1404

مسیر پیش رو

لیگ علم داده با هدف یادگیری پروژه محور طراحی شده است و شامل دو فاز اصلی خواهد بود:

فاز اول (۴ هفته)

در این مرحله همه ما روی یک دیتاست مشترک (نمرات دانش‌آموزان) کار می‌کنیم. هدف این است که با الفبای استاندارد کار با داده آشنا شوید، نگاه تحلیلی پیدا کنید و پایه‌ای قوی برای مراحل پیشرفته بسازید.

فاز دوم (تخصصی)

پس از تسلط بر مفاهیم پایه، در فاز دوم گروه‌بندی انجام خواهد شد. در این مرحله قصد داریم روی دیتاست‌های چالشی و مسابقاتی (نظیر Kaggle) کار کنیم و مسیرهای تخصصی(متن، تصویر، یا داده‌های جدولی پیشرفته) را پیش ببریم.

راهنمای شروع کار

برای دریافت داده‌ها و ثبت نتایج، مراحل زیر را انجام دهید:

1. وارد سایت **دیتاکوییز (Dataqueez)** شوید و ثبت‌نام کنید.
2. به بخش **سوالات** رفته و سوال **"نمره ریاضی"** را انتخاب کنید.
3. در صفحه سوال، وارد تب **داده‌ها** شوید.
4. فایل‌های **داده آموزش (Train)** و **داده آزمایش (Test)** را دانلود کنید.
5. شما می‌توانید فایل پیش‌بینی خود را در بخش ارسال پاسخ آپلود کرده و جایگاه خود را در تب **رتبه‌بندی** مشاهده کنید.

نقشه راه ۴ هفته‌ای (The Big Picture)

1. **تحلیل اکتشافی (EDA)**
   Data Understanding & Stats
2. **پاکسازی و مهندسی**
   Preprocessing & Features
3. **مدل‌سازی مقدماتی**
   Basic Modeling
4. **مدل‌سازی پیشرفته**
   Ensemble & Optimization

* ما در این هفته روی مرحله اول (تحلیل اکتشافی و درک آماری) تمرکز داریم. هدف، کشف داستانِ پشتِ اعداد است.

🔥 چالش‌های هفته اول: کشف لایه‌های پنهان

فرض کنید شما به عنوان تحلیل‌گر ارشد در موسسه "InsightEdu" استخدام شده‌اید. هیئت مدیره مدرسه فایل‌هایی را برای شما ارسال کرده است، اما هیچ توضیحی نداده‌اند. وظیفه شما استخراج دانش از این فایل‌هاست.

نکته مهم: لطفاً یک ژوپیتر نوت‌بوک ایجاد کرده و پاسخ هر چالش را در آن مستند کنید. سلول‌هایی که ایجاد می‌کنید باید شامل **کد (Code)** و **سلول‌های متنی (Markdown)** باشند. راجع به نحوه نوشتن Markdown تحقیق کنید تا گزارش‌هایتان حرفه‌ای و خوانا باشد.

چالش ۱: اولین برخورد (FIRST CONTACT)

پرونده‌ها روی میز شماست. قبل از هر کاری باید ببینیم با چه چیزی طرف هستیم.

- داده‌های `train.csv` و `test.csv` را با استفاده از کتابخانه **Pandas** لود کنید.
- با استفاده از متدهای `describe`، `info()`، `head()` و `value_counts()` یک نمای کلی از داده‌ها بدست آورید.
- در سلول‌های Markdown، درک خود را از داده‌ها و از ستون‌های ویژگی (Attributes) بنویسید. به نظرتان هر ستون چه معنایی دارد؟
- کدام ستون‌ها عددی (Numerical) و کدام‌یک دسته‌ای (Categorical) هستند؟

📚 مطالعه مخصوص این چالش: **Pandas 10 minutes - Viewing Data**

چالش ۲: معمای ماشین زمان (DATA LEAKAGE)

یکی از کارآموزان پیشنهاد داده که: "بیایید اول تمام داده‌های Train و Test را با هم ترکیب کنیم، داده‌های گمشده (Missing Values) را با میانگین کل پر کنیم و سپس دوباره آن‌ها را جدا کنیم تا مدلمان دقیق‌تر شود."

- چرا این پیشنهاد خطرناک است؟ (راهنمایی: مفهوم **Data Leakage** یا نشت داده را بررسی کنید).
- اگر ما میانگین نمرات `test` را بدانیم و از آن برای پر کردن جاهای خالی `train` استفاده کنیم، آیا در حال تقلب از آینده نیستیم؟
- ترتیب صحیح "تقسیم داده (Splitting)" و "پیش‌پردازش (Preprocessing)" چگونه باید باشد؟

📚 مطالعه مخصوص این چالش: **Data Leakage in Machine Learning**

چالش ۳: شرط‌بندی روی ثبات (VARIANCE & STD)

یک سرمایه‌گذار می‌خواهد روی یکی از دو مدرسه "Gabriel Pereira" یا "Mousinho da Silveira" سرمایه‌گذاری کند. شرط او این است: "من به دنبال نوابغ نیستم، من به دنبال ثبات (Consistency) هستم. مدرسه‌ای را می‌خواهم که نمرات دانش‌آموزانش نوسان کمتری داشته باشد."

- به نظر شما کدام مفهوم آماری به سرمایه‌گذار کمک می‌کند؟ مدرسه مناسب را با توجه به آن پیدا کنید.
- برای هر مدرسه یک نمودار جعبه‌ای (Boxplot) از نمرات نهایی (G3) رسم کنید.
- اگر واریانس داده‌ها در یک مدرسه بسیار زیاد باشد، انتظار دارید شکل Boxplot آن چه ویژگی‌های خاصی داشته باشد؟ (عرض جعبه؟ طول شاخک‌ها؟)

مطالعه مخصوص این چالش: **Boxplots Interpretation**

چالش ۴: راز صفرها (DISTRIBUTIONS & OUTLIERS)

معاون آموزشی مدرسه گزارش داده که برخی دانش‌آموزان نمره نهایی (G3) صفر گرفته‌اند و نگران است که آیا سیستم نمره‌دهی مشکلی دارد یا خیر.

- نمودار توزیع (Histogram) نمرات G3 را رسم کنید. آیا توزیع نرمال است؟
- آیا نمرات "صفر" بخشی از روند طبیعی کلاس هستند یا داده‌های پرت (Outliers) محسوب می‌شوند؟
- با بررسی سایر ویژگی‌ها (مثلاً غیبت‌ها `absences` یا نمرات ترم‌های قبل) کارآگاه‌بازی درآورید! آیا این دانش‌آموزان درس نخوانده‌اند یا احتمالاً در جلسه امتحان غایب بوده‌اند؟

مطالعه مخصوص این چالش: **Identifying Outliers**

چالش ۵: جنگ آمارها (CORRELATION & HYPOTHESIS)

مدیریت مدرسه با هیجان می‌گوید: "طبق محاسبات ما، همبستگی (Correlation) مثبتی بین دسترسی به اینترنت و نمره نهایی وجود دارد. باید فوراً بودجه خرید مودم را تصویب کنیم!"

اما **مدیر مالی** که فردی دقیق و شکاک است مخالفت می‌کند: "میانگین نمراتشان فقط کمی بالاتر است، شاید شانسی باشد! نمی‌توانیم بر اساس شانس بودجه را هدر دهیم."

- **گام اول:** ماتریس همبستگی (Correlation Matrix) را رسم کنید. آیا ادعای مدیریت درباره همبستگی صحیح است؟
- **گام دوم (پاسخ به مدیر مالی):** شما باید حرف مدیر مالی را با علم آمار رد یا تایید کنید. دو گروه (اینترنت‌دار و بدون اینترنت) را جدا کنید و با استفاده از آزمون **T-Test** و مقدار **P-value** بررسی کنید که آیا اختلاف میانگین نمرات این دو گروه "معنادار" (Significant) است یا صرفاً تصادفی؟

✨ **بخش امتیازی (Optional):**
آیا می‌توانید کدی بنویسید که مثال نقضی برای داده‌های پرت پیدا کند؟
راهنمایی: چگونه می‌توان نقاطی را شناسایی کرد که در تک‌تک ابعاد "عادی" هستند (مثلاً سن نرمال، نمره نرمال)، اما "اجتماع" ویژگی‌هایشان غیرعادی است؟ (Multivariate Outliers).

📚 مطالعه مخصوص این چالش: **Understanding T-Tests and P-values**

چالش ۶: زبان مشترک با کامپیوتر (ENCODING)

مشکل: کامپیوترها کلمات را نمی‌فهمند. ستون شغل مادر (Mjob) شامل مقادیری مثل 'Teacher', 'Health',' Other'' است. چطور می‌توانیم این کلمات را به عدد تبدیل کنیم؟

- اگر به Teacher عدد 1 و به Health عدد 2 بدهیم، آیا مدل ریاضی ما به اشتباه فکر نمی‌کند که Health "بزرگتر" یا "بهتر" از Teacher است؟ (مشکل داده‌های Nominal).
- تحقیق کنید: برای داده‌هایی که ترتیب ندارند (مثل شغل) چه روش تبدیلی (Encoding) پیشنهاد می‌شود؟
- کدی بنویسید که داده‌های متنی ستون‌های شغلی را به شکل صحیح به عدد تبدیل کند.

📚 مطالعه مخصوص این چالش: **Encoding Categorical Features**

⛔ تمرین نهایی: شکار باگ نامرئی (Logic Error)

کد زیر توسط یک برنامه‌نویس تازه‌کار نوشته شده است. این کد ارور نمی‌دهد (Syntax Error ندارد)، اما از نظر "منطق علم داده" دارای یک اشتباه مرگبار است.

با توجه به **چالش ۳**، ایراد کد زیر را پیدا کنید و نسخه اصلاح شده (Corrected Version) را بنویسید.

```
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
import numpy as np

train_data = pd.read_csv('train.csv')
test_data = pd.read_csv('test.csv')

full_data = pd.concat([train_data, test_data])

print("Average Student Grade (Global):", full_data['G3'].mean())

full_data['absences'] = (full_data['absences'] - full_data['absences'].mean()) /
full_data['absences'].std()

df_encoded = pd.get_dummies(full_data, drop_first=True)

train_processed = df_encoded.iloc[:len(train_data)]
test_processed = df_encoded.iloc[len(train_data):]

model = LinearRegression()

X_train = train_processed
y_train = train_processed['G3']

model.fit(X_train, y_train)

predictions = model.predict(test_processed)
rmse = np.sqrt(mean_squared_error(test_processed['G3'], predictions))
print("Model RMSE:", rmse)
```

📚 منابع برای یادگیری بیشتر

کتاب آمار و احتمال شلدون راس

مرجع کلاسیک و ارزشمند برای درک عمیق مفاهیم پایه آماری.

دوره جامع پایتون (Pytopia)

مناسب برای یادگیری و مرور مفاهیم پایه پایتون.

مبانی یادگیری ماشین (Pytopia)

آشنایی با مفاهیم تئوری و الگوریتم‌های یادگیری ماشین.

دام‌های رایج در یادگیری ماشین

مستندات رسمی Scikit-Learn درباره اشتباهات رایج (مثل نشت داده).

StatQuest with Josh Starmer

بهترین کانال یوتیوب برای یادگیری بصری مفاهیم آماری و یادگیری ماشین.

برگه تقلب Pandas

خلاصه‌ای کاربردی از دستورات مهم کتابخانه پانداس برای دسترسی سریع.

نقشه‌راه جامع علم‌داده (GitHub)

مسیر یادگیری و منابع دسته‌بندی شده.

منابع رایگان هوش مصنوعی

مجموعه‌ای از کتاب‌ها و دوره‌های رایگان فارسی و انگلیسی.

جعبه‌ابزار دیتاساینس در پایتون

لیست بهترین کتابخانه‌های پایتون برای علم داده.

مستندات Scikit-Learn

مرجع اصلی الگوریتم‌ها و ابزارهای یادگیری ماشین.