

دانشگاه تهران – دانشکده مهندسی برق و کامپیوتر

درس علم داده: تمرین نهایی جامع (نسخه حرفه‌ای توسعه‌یافته)

تیم درس علم داده – بهار ۱۴۰۴

ویرایش ۰.۲

عنوان پروژه: تحلیل مهاجرت جهانی استعدادهای فنی با رویکرد داده‌محور
داده اصلی: GlobalTechTalent_50k.csv (۵۰,۰۰۰ رکورد)
مسئله اصلی: پیش‌بینی Migration_Status (۱=مهاجرت، ۰=عدم مهاجرت)

۱) هدف آموزشی و فلسفه ارزیابی

این تمرین نهایی برای سنجش هم‌زمان درک ریاضی، پیاده‌سازی مهندسی، تحلیل انتقادی، و گزارش‌نویسی حرفه‌ای طراحی شده است. انتظار می‌رود دانشجوی در این پروژه نشان دهد که می‌تواند:

- مسئله واقعی را به یک مسئله داده‌محور قابل ارزیابی تبدیل کند؛
- از داده خام تا استقرار پیشنهادی، یک خط لوله کامل و بازتولیدپذیر بسازد؛
- از مدل‌های ساده تا مدل‌های غیرخطی/پیشرفته را با استدلال علمی مقایسه کند؛
- با روش‌های XAI و تحلیل عدالت، خروجی مدل را قابل حسابرسی نماید.

۲) اقلام تحویلی الزامی

۱. نوبت‌بک اصلی پروژه با ساختار Q1 تا Q14 و بخش Capstone.
۲. گزارش نهایی (حداکثر ۲۵ صفحه بدون پیوست) شامل شکل‌ها، جدول‌ها، تفسیرها و محدودیت‌ها.
۳. بسته کد شامل اسکریپت‌های قابل اجرا، فایل وابستگی‌ها، و راهنمای اجرای کامل.
۴. پاسخ‌نامه تشریحی برای همه سوالات (فرمول، استدلال، خروجی).
۵. خلاصه مدیریتی ۱ تا ۲ صفحه برای مخاطب غیرتخصصی.

حداقل استاندارد فنی:

- ثبت random seed در تمام بخش‌های تصادفی.
- تفکیک شفاف train/validation/test.
- کنترل صریح نشت داده و نشت زمانی.
- ثبت نسخه کتابخانه‌ها و محیط اجرا.
- توان بازتولید نتایج صرفاً با اجرای دستورات اعلام‌شده.

۳) ساختار نمره‌دهی (۲۰۰ نمره)

نمره	حوزه مهارتی	بلوک
۲۰	مبانی: چرخه عمر علم داده، عملیات پایتونی، EDA	A
۲۰	استنباط آماری، طراحی بصری و روایت داده	B
۲۵	SQL پیشرفته و مهندسی داده در مقیاس	C
۴۵	مدل‌سازی نظارت‌شده و بهینه‌سازی	D
۲۰	یادگیری بدون نظارت و کاهش بُعد	E
۳۰	یادگیری عمیق، NLP، مدل‌های زبانی/عامل‌ها	F
۱۵	عدالت، سوگیری، حاکمیت و استقرار مسئولانه	G
۲۵	کپستون یکپارچه (پیا德سازی + تبیین + گزارش)	H
۲۰۰	جمع کل	

۴) صورت سوال توسعه‌یافته

بلوک A – مبانی (۲۰ نمره)

Q۱) چرخه عمر علم داده و صورت‌بندی مسئله (۱۰ نمره)

یک سند فنی کوتاه ارائه دهید که شامل موارد زیر باشد:

۱. تعریف مسئله، ذی‌نفعان، و تصمیم‌هایی که مدل قرار است پشتیبانی کند.
 ۲. تعریف KPIهای فنی (مثل AUC، Recall@K، Calibration) و عملیاتی.
 ۳. فهرست ریسک‌ها (نشت داده، drift، تغییر سیاست مهاجرت، خطای برچسب).
 ۴. برنامه پایش پس از استقرار (آستانه هشدار، فرکانس بازآموزی، مسیر ارجاع انسانی).
- خروجی مورد انتظار:** نمودار چرخه عمر + یک جدول ریسک.

Q۲) عملیات پایتونی، کنترل کیفیت و EDA (۱۰ نمره)

۱. پروفایل‌گیری داده: نوع ستون‌ها، مقادیر گمشده، تکراری، بازه‌های نامعتبر.
۲. شناسایی پرت‌ها با حداقل دو روش (مثلاً IQR و z-score) و تحلیل پیامد.
۳. حداقل ۸ نمودار اکتشافی با تفسیر کاربردی.
۴. پیاده‌سازی یک تابع پیش‌پردازش ماژولار و قابل آزمون.

بلوک B – استنباط آماری و مصورسازی (۲۰ نمره)

Q۳) طراحی مطالعه و استنباط (۱۰ نمره)

- تفکیک مطالعه مشاهده‌ای از ادعای علی برای این مسئله.
- ارائه حداقل یک بازه اطمینان معتبر و تفسیر دقیق آن.
- تعریف یک آزمون فرض با H_0/H_1 ، سطح معنی‌داری، کنترل خطای نوع اول.
- بررسی اعتبار پیش‌فرض‌ها (نرمال‌بودن تقریبی/حجم نمونه/استقلال).

(Q4 طراحی بصری و روایت برای تصمیم‌گیر (۱۰ نمره)

- طراحی یک بخش داشبوردی برای مخاطب غیر تخصصی.
- توجیه انتخاب رنگ، مقیاس، ترتیب، و annotation ها.
- ارائه حداقل یک مثال نمودار گمراه‌کننده + نسخه اصلاح‌شده.

بلوک C – SQL و مهندسی داده (۲۵ نمره)

(Q5 SQL پیشرفته (۱۵ نمره)

سه کوئری مستقل ارائه دهید:

۱. میانگین متحرک ۳ ساله Research_Citations با window frame.
۲. رتبه‌بندی/دهک‌بندی کشورها یا کاربران با RANK/DENSE_RANK/NTILE.
۳. یک تحلیل cohort با CTE (ماندگاری یا انتقال وضعیت در زمان).

(Q6 نشت داده و معماری داده در مقیاس (۱۰ نمره)

- ویژگی‌های نشت‌دهنده را شناسایی و با منطق زمانی حذف کنید.
- معماری Bronze/Silver/Gold یا معادل آن را برای داده آفلاین/آنلاین پیشنهاد دهید.
- راهکار train-serving consistency با feature store توضیح دهید.

بلوک D – مدل‌سازی نظارت‌شده و بهینه‌سازی (۴۵ نمره)

(Q7 مدل‌های خطی/لجستیک + Elastic Net (۱۵ نمره)

۱. مدل پایه خطی/لجستیک را اجرا کنید.
۲. تابع هزینه Elastic Net را بنویسید و گرادینت/زیرگرادیان را استخراج کنید.
۳. تفسیر ضریب، p-value، بازه اطمینان، و پایداری ضرایب را ارائه دهید.

(Q8 تحلیل بهینه‌سازی در Ravine (۱۰ نمره)

- رفتار SGD، Momentum، و Adam را روی یک تابع دره‌ای مقایسه کنید.
- مسیر همگرایی و نوسان را رسم و تحلیل کنید.
- برای ناهمگنی مقیاس ویژگی‌ها پیشنهاد عملی بدهید.

(Q9 مقایسه خانواده مدل‌های غیرخطی (۲۰ نمره)

- KNN و SVM
- Decision Tree و Random Forest
- یک مدل Boosting (ترجیحاً XGBoost)

الزامات:

۱. پروتکل cross-validation و تنظیم ابرپارامتر.
۲. جدول متریک‌ها (حداقل: Accuracy, ROC-AUC, F1, Precision, Recall).
۳. تحلیل خطا (الگوهای اشتباه، ماتریس درهم‌ریختگی، حساسیت تصمیم).

بلوک E – بدون نظارت (۲۰ نمره)

(Q۱۰ کاهش بُعد (۱۰ نمره)

- PCA: محاسبه و تفسیر Explained Variance Ratio.
- یک روش مکمل (UMAP/t-SNE/Random Projection).
- تفسیر هندسی ابعاد نهفته.

(Q۱۱ خوشه‌بندی (۱۰ نمره)

- Silhouette + Elbow + KMeans.
- DBSCAN یا روش چگالی‌محور معادل.
- تحلیل پایداری خوشه‌ها و معنای کاربردی آن‌ها.

بلوک F – یادگیری عمیق، NLP و مدل‌های زبانی (۳۰ نمره)

(Q۱۲ مدل عصبی جدولی و مدل توالی/متن (۱۵ نمره)

- یک MLP روی داده جدولی آموزش دهید.
- یک مدل توالی/متنی (RNN/LSTM/GRU/CNN) اجرا کنید.
- عملکرد را با baseline کلاسیک مقایسه کنید.

(Q۱۳ مدل‌های زبانی و LLM Agent (۱۵ نمره)

جریان عامل‌محور طراحی کنید:

plan -> retrieve -> reason -> verify

و برای آن:

- معیار ارزیابی صحت/وفاداری/ایمنی تعریف کنید.
- سیاست مهار hallucination و محدودیت دسترسی ابزارها مشخص کنید.

بلوک G – اخلاق، عدالت و حاکمیت (۱۵ نمره)

(Q۱۴ عدالت، سوگیری، و استقرار مسئولانه (۱۵ نمره)

- تحلیل زیرگروهی بر اساس کشور/تحصیلات/سابقه.
- بررسی proxy discrimination و سوگیری تاریخی سیاست‌ها.
- ارائه سیاست human-in-the-loop و مسیر اعتراض/بازبینی.

بلوک H – کیستون یکپارچه (۲۵ نمره)

خروجی کیستون باید شامل موارد زیر باشد:

۱. خط لوله کامل داده تا پیش‌بینی با کنترل نشت.
۲. یک گزارش model card شامل مفروضات، محدودیت‌ها، متریک‌ها.
۳. تبیین محلی و سراسری با SHAP.
۴. جدول عدالت زیرگروهی + توصیه استقرار + مانیتورینگ.

۵) قالب‌های تحویل و استاندارد مستندسازی

- همه شکل‌ها باید عنوان، واحد، و caption دقیق داشته باشند.
- تمام ادعاها باید با خروجی کد/جدول پشتیبانی شوند.
- کد باید ماژولار، دارای نام‌گذاری استاندارد، و قابل اجرا با دستور واحد باشد.
- گزارش باید شامل بخش «محدودیت‌ها» و «کارهای آینده» باشد.

۶) معیار کسر نمره

- وجود نشت داده بدون گزارش: تا ۵۰٪ کسر در بخش مرتبط.
- نبود بازتولیدپذیری: تا ۳۰٪ کسر در پروژه.
- تفسیر نادرست آماری (مثلاً برداشت اشتباه از p-value): کسر ۲۰٪ در سوال.
- فقدان تحلیل عدالت یا اخلاق: کسر کامل نمره بلوک G.

۷) امتیاز تشویقی (تا +۱۰ نمره)

- توسعه تحلیل علی با DAG و بحث شناسایی.
- افزودن temporal validation و تحلیل drift بین سال‌ها.
- اضافه‌کردن conformal prediction یا برآورد عدم قطعیت پیشرفته.

۸) اصول اخلاق علمی

- استفاده از منبع بیرونی بدون ارجاع، تخلف آموزشی محسوب می‌شود.
- نتایج ضعیف/منفی نیز باید صادقانه گزارش شوند.
- استفاده از مدل‌های زبانی باید شفاف و مستند باشد.

جمع‌بندی: این تمرین صرفاً «ساخت مدل» نیست؛ هدف، ارائه یک سامانه تصمیم‌یار قابل دفاع است که از نظر فنی، آماری، اخلاقی و مهندسی در سطح حرفه‌ای دانشگاهی قابل ارزیابی باشد.