

دانشگاه تهران – دانشکده مهندسی برق و کامپیوتر

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

تیم درس علم داده – بهار ۱۴۰۴

ویرایش ۰.

عنوان پژوهش: تحلیل مهاجرت جهانی استعدادهای فنی با رویکرد داده محور
داده اصلی: GlobalTechTalent_50k.csv (۵۰،۰۰۰ رکورد)
مسئله اصلی: پیش‌بینی Migration_Status (۰ = مهاجرت، ۱ = 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 دانشجو در این پژوهش نشان دهد که می‌تواند:

• مسئله واقعی را به یک مسئله داده محور قابل ارزیابی تبدیل کند؛

• از داده خام تا استقرار پیشنهادی، یک خط لوله کامل و بازنویسیدن 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 تحولی الزامی

۱. نوتبوک اصلی پژوهش 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.

• تقسیم شفاف train/validation/test.

• کنترل صریح نشت داده و نشت زمانی.

• ثبت نسخه کتابخانه‌ها و محیط اجرا.

• توان بازنویسیدن نتایج صرفاً با اجرای دستورات اعلام شده.

۳) ساختار نمره‌دهی (۲۰۰ نمره)

نمره	حوزه مهارتی	بلوک
۲۰	مبانی: چرخه عمر علم داده، عملیات پایتونی، EDA	A
۲۰	استتباط آماری، طراحی بصری و روایت داده	B
۲۵	SQL پیشرفت و مهندسی داده در مقیاس	C
۴۵	مدل‌سازی نظرات‌شده 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 مسنولانه	G
۲۵	کپسون یکپارچه (پیاده‌سازی + تبیین + گزارش)	H
۴۰۰	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 (مثلًا IQR و z-score) و تحلیل پیامد.
۳. حداقل ۸ نمودار اکتشافی با تفسیر کاربردی.
۴. پیاده‌سازی یکتابع پیش‌پردازش مازولار و قابل آزمون.

بلوک B – استتباط آماری و مصورسازی (۲۰ نمره)

(Q۳) طراحی مطالعه و استتباط (۱۰ نمره)

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

(Q۴) طراحی بصری و روایت برای تصمیمگیر (۱۰ نمره)

- طراحی یک بخش داشبوردی برای مخاطب غیرتخصصی.
- توجیه انتخاب رنگ، مقیاس، ترتیب، و annotation‌ها.
- ارائه حداقل یک مثال نموذار گمراهنده + نسخه اصلاح شده.

بلوک C – SQL و مهندسی داده (۲۵ نمره)

(SQL Q۵) 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 (ماندگاری یا انتقال وضعیت در زمان) با CTE .

(Q۶) 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 feature store با train-serving consistency توضیح دهید.

بلوک D – مدل‌سازی نظارت شده و بهینه‌سازی (۴۵ نمره)

(Q۷) مدل‌های خطی/لجستیک + Elastic Net (۱۵ نمره)

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

(Q۸) 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.

(Q۹) مقایسه خانواده مدل‌های غیرخطی (۲۰ نمره)

• KNN و SVM

• Random Forest و Decision Tree

• 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 خطای اشتباه، ماتریس در هم ریختگی، حساسیت تصمیم.

بلوک E – بدون نظارت (۲۰ نمره)

(Q1۰) کاهش بُعد (۱۰ نمره)

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

(Q1۱) خوشبندی (۱۰ نمره)

- Silhouette + Elbow + KMeans
- DBSCAN یا روش چگالی محور معادل.
- تحلیل پایداری خوشها و معنای کاربردی آنها.

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

(Q1۲) 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.

(Q1۳) مدل‌های زبانی و LLM Agent (۱۵ نمره)

جريان عامل محور 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)

(Q1۴) 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 – کپستون یکپارچه (۲۵ نمره)

خروجی کپستون باید شامل موارد زیر باشد:

۱. خط لوله کامل داده تا پیش‌بینی با کنترل نشت.
۲. یک گزارش model card شامل مفروضات، محدودیت‌ها، متريک‌ها.
۳. تبيين محلی و سراسری با SHAP.
۴. جدول عدالت زیرگروهی + توصيه استقرار + مانيتورينگ.

۵) قالب‌های تحویل و استاندارد مستندسازی

- همه شکل‌ها باید عنوان، واحد، و caption دقیق داشته باشند.
- تمام ادعاهای باید با خروجی کد/جدول پشتیبانی شوند.
- کد باید مازولار، 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 مرتبه.
- نبود بازتولیدپذیری: تا ۳۰٪ کسر در پروژه.
- تفسیر نادرست آماری (مثلًا 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.
- افروزندهای temporal validation و تحلیل drift بین سال‌ها.
- اضافه کردن conformal prediction یا برآورد عدم قطعیت پیشرفت.

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 تصمیم‌گیر قابل دفاع است که از نظر فنی، آماری، اخلاقی و مهندسی در سطح حرفه‌ای دانشگاهی قابل ارزیابی باشد.