دانشگاه صنعتی خواجه نصیرالدین طوسی تهران
مبانی سیستم های هوشمند (پاییز ۱۴۰۴)، مهدی علیاری شوره‌دلی

ارزشیابی پایانی

دستورالعمل‌های اجرایی و شیوه تحویل ارزشیابی پایانی

دانشجویان گرامی، رعایت **کمال دقت** و **اصالت علمی** در تمام مراحل انجام پروژه الزامی است. خروجی نهایی باید دقیقاً مطابق ساختار زیر دسته‌بندی شده و در قالب یک فایل فشرده (.zip) با نام‌گذاری FIS4041_F_StudentNumber#1_StudentNumber#2 ارسال گردد:

- **۱. گزارش فنی (PDF):** یک فایل پی‌دی‌اف جامع و کامل شامل شرح روند کار، پاسخ به تمامی سوالات، و ارائه دقیق تحلیل‌ها. متریک‌ها و نمودارها باید با کیفیت بالا ترسیم و تفسیر شوند (صرفاً قراردادن نمودار کافی نیست).
- **۲. فایل ارائه (PowerPoint):** اسلایدهایی که تمام بخش‌های پروژه را به صورت ساختاریافته، دقیق و کامل برای ارائه پوشش دهد.
- **۳. ویدیوی توضیحات (اختیاری):** یک ویدیوی کوتاه از اجرای شبیه‌سازی و توضیح نتایج.
- **۴. زیرپوشهٔ کدهای پروژه (Code):** بر اساس روش تحویل کد خود، تمام فایل‌های اجرایی پایتون (.py)، دفترچه‌کدها (.ipynb)، متلب (.m)، مدل‌های سیمولینک (.slx) و داده‌ها باید در این پوشه قرار گیرند و قابل اجرا باشند.
- **۵. زیرپوشهٔ فایل‌های منبع گزارش (Report):** فایل‌های سورس گزارش (فایل Word یا فایل‌های LaTeX) در این قسمت قرار گیرند.

قوانین الزامی و امتیازات ویژه:

۱. **ارائهٔ شفاهی:** این پروژه دارای جلسهٔ ارائه شفاهی است و جدول زمان‌بندی دقیق آن متعاقباً اطلاع‌رسانی خواهد شد.

۲. **کار گروهی:** این پروژه در گروه‌های **حداکثر دو نفره** قابل تعریف است. در صورت انجام گروهی، مشارکت فعال و **تسلط کامل هر دو نفر** بر تمام اجزای پروژه (شامل کدنویسی، تئوری و تحلیل‌ها) الزامی است و مورد سنجش قرار می‌گیرد.

۳. **نمره امتیازی (تا ۲۰ درصد نمره اضافه):** موارد زیر به عنوان فعالیت‌های فراتر از انتظار، دارای امتیاز ویژه هستند:

- مدیریت ورژن و قرار دادن پروژه در **GitHub** (لینک ریپازیتوری در گزارش قید شود).
- ارائه خروجی‌ها (گزارش و اسلاید) در قالب **LaTeX** و **Beamer** با طراحی جذاب، زیبا و حرفه‌ای.
- انجام تحلیل‌های بیشتر و عمیق‌تر روی سوالات و ارائه خروجی‌های کامل‌تر و فراتر از صورت سوال.

این آزمون جهت سنجش نهایی آموخته‌های شما طراحی شده است. با توجه به شرایط حال حاضر و محدودیت‌های احتمالی در دسترسی به اینترنت بین‌الملل، لطفاً **آرامش خود را حفظ کنید**. تمامی تلاش‌های شما ارزشمند است و محدودیت‌های موجود در فرایند نمره‌دهی و بررسی پروژه‌ها کاملاً **درک خواهد شد**. نگران قطعی‌ها یا کندی سرعت نباشید و تنها تا حداکثر جای ممکن تلاش خود را به کار گیرید. ما در کنار شما هستیم.

پرسش یک

لینک مجموعه‌داده: Loan Dataset (در فایل فشرده نیز موجود است).

توضیح داده: مجموعه‌دادهٔ مورد استفاده مربوط به وضعیت وام‌های اعطایی یک مؤسسهٔ مالی است. هر ردیف نمایانگر متقاضی یک وام بوده و شامل اطلاعاتی همچون جنسیت، وضعیت تأهل، تعداد افراد تحت تکفل، وضعیت شغلی، میزان درآمد، سابقهٔ تحصیلی، نوع منطقهٔ محل سکونت، مبلغ وام و در نهایت برچسب خروجی «وضعیت وام» (تأیید یا رد) می‌باشد. هدف، شناسایی ویژگی‌هایی است که بیشترین نقش را در تصمیم نهایی (تأیید یا عدم تأیید وام) دارند.

با توجه به تعداد بالای ویژگی‌ها و احتمال وجود ویژگی‌های غیرمؤثر یا هم‌بسته، انتخاب زیرمجموعهٔ بهینه از آن‌ها می‌تواند منجر به بهبود عملکرد مدل یادگیری گردد. در این تمرین، شما باید از **الگوریتم‌های تکاملی** برای انجام فرآیند Feature Selection استفاده کنید و نتایج دو روش را مقایسه نمایید.

آ) داده را بارگذاری کرده و تمامی مقادیر NaN را حذف کنید. سپس ستون‌های متغیرهای اسمی مانند Gender، Married، Education، Self_Employed، Area و Status را با استفاده از LabelEncoder به مقادیر عددی تبدیل نمایید.

ب) الگوریتم Particle Swarm Optimization (PSO) را برای انتخاب ویژگی‌ها در فضای باینری پیاده‌سازی کنید (با استفاده از کتابخانهٔ pyswarms). هدف تابع برازندگی باید بیشینه کردن دقت مدل Random Forest و همزمان حداقل‌سازی تعداد ویژگی‌های انتخاب‌شده باشد. می‌توانید از ترکیب زیر برای تابع برازندگی استفاده کنید:

$$J = \alpha(1 - Acc) + (1 - \alpha)\left(1 - \frac{\text{تعداد ویژگی‌های انتخاب‌شده}}{\text{تعداد کل ویژگی‌ها}}\right)$$

که در آن α مقدار وزن بین دو عامل است.

ج) الگوریتم Genetic Algorithm (GA) را با استفاده از کتابخانهٔ DEAP برای همین مسئله پیاده‌سازی کنید. در آن، هر فرد یک رشتهٔ باینری است که مقدار ۱ نشان‌دهندهٔ انتخاب آن ویژگی و مقدار ۰ نشان‌دهندهٔ عدم انتخاب است. از عملگرهای Two-Point Crossover و Bit Flip Mutation برای ایجاد نسل‌های جدید استفاده کنید.

د) برای هر الگوریتم، دقت مدل Random Forest را پس از انتخاب ویژگی‌ها محاسبه کرده و میانگین دقت و تعداد ویژگی‌های انتخاب‌شده را گزارش دهید. نتایج دو الگوریتم را از نظر عملکرد (دقت و سادگی مدل) مقایسه نمایید.

ه) ویژگی‌هایی را که هر الگوریتم «مهم‌تر» تشخیص داده است چاپ کنید و در چند خط تحلیل کنید که چرا انتخاب‌ها ممکن است متفاوت باشند.

و) **(بخش مقایسه و تحلیل)** نمودار تغییرات مقدار برازندگی یا دقت را برای هر دو الگوریتم (PSO و GA) در طول تکرارها رسم کنید. محور افقی را «تعداد تکرارها / نسل‌ها» و محور عمودی را «مقدار برازندگی یا دقت» در نظر بگیرید. در یک نمودار واحد، منحنی هر الگوریتم را با رنگ متفاوت نمایش دهید و نتایج را از نظر سرعت همگرایی، پایداری و دقت نهایی مقایسه کنید. در پایان، تحلیلی کوتاه از مشاهدهٔ نمودارها ارائه دهید.

راهنمای پارامترها و نکات اجرایی:

- پارامترهای پیشنهادی برای PSO: $w = 0.9,\ c_2 = 0.3,\ c_1 = 0.5$، تعداد ذرات 50، تکرارها 100.
- پارامترهای پیشنهادی برای GA: اندازهٔ جمعیت 50، تعداد نسل‌ها 50، احتمال تقاطع 0.9، احتمال جهش 0.1.
- از train_test_split با نسبت 70%/30% برای آموزش و ارزیابی استفاده کنید.
- در خروجی گزارش، شامل نمودار مقایسه‌ای، فهرست ویژگی‌های انتخاب‌شده، مقادیر دقت نهایی و تحلیل کوتاه از عملکرد دو الگوریتم باشید.

۱ پرسش دو

داده‌ها: Mall Customer Segmentation (در فایل فشرده نیز موجود است).

آ. داده را بارگذاری، پاک‌سازی و ویژگی‌های عددی را انتخاب کنید. فقط روی ویژگی‌ها StandardScaler اعمال کنید. یک PCA دوبعدی صرفاً برای نمایش بسازید (مدل‌ها در فضای اصلی آموزش ببینند).

ب. برای $K \in \{2, \dots, 10\}$ KMeans را آموزش دهید. برای هر K: inertia و silhouette را گزارش کنید و K نهایی را با استدلال انتخاب کنید.

ج. AgglomerativeClustering را با linkageهای single/complete/average/Ward و تعداد خوشهٔ انتخابی اجرا کنید؛ سیلوئت را مقایسه و بهترین linkage را گزارش کنید.

د. DBSCAN را با یک شبکهٔ کوچک برای $\varepsilon \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ و $\text{min\_samples} \in \{3, 5, 10\}$ اجرا کنید. تعداد خوشه‌ها، نسبت نویز و سیلوئت روی نقاط غیرنویز را گزارش و بهترین تنظیم را انتخاب کنید.

ه. برای بهترین تنظیم هر خانواده (مرکزی/سلسله‌مراتبی/چگالی)، برچسب خوشه‌ها را روی فضای PCA دوبعدی رسم و تحلیل کنید.

اصل تکرارپذیری: در همهٔ بخش‌های تصادفی (مثل KMeans یا PCA) مقدار random_state را برابر **دو رقم آخر شمارهٔ دانشجویی** تنظیم کنید.

پرسش سه

آ) معادله‌ی به‌روزرسانی Q-learning را بنویسید و توضیح دهید هر کدام از پارامترهای α (نرخ یادگیری)، γ (ضریب تنزیل) و ϵ در سیاست ϵ-greedy چه نقشی در یادگیری دارند. همچنین توضیح دهید چرا Q-learning یک روش off-policy محسوب می‌شود.

$$Q_{t+1}(s_t, a_t) = (1 - \alpha)\, Q_t(s_t, a_t) + \alpha\Big(r_t + \gamma \max_a Q_t(s_{t+1}, a)\Big)$$

ب) فرض کنید جدول Q در ابتدا صفر است و گذار زیر مشاهده می‌شود:

$$(s_t = s_0,\ a_t = a_1,\ r_t = +2,\ s_{t+1} = s_1)$$

همچنین در همان لحظه داریم $\max_a Q(s_1, a) = 1.5$. با $\alpha = 0.2$ و $\gamma = 0.9$، مقدار جدید $Q(s_0, a_1)$ را محاسبه کنید (همه‌ی مراحل را شفاف بنویسید).

ج) دو عامل که می‌توانند باعث کندی همگرایی یا ناپایداری یادگیری شوند را نام ببرید و برای هرکدام یک راهکار عملی پیشنهاد کنید. برای مثال می‌توانید به مواردی مثل «طراحی/مقیاس‌دهی پاداش»، «تنظیم و کاهش تدریجی ϵ»، «افزایش تعداد بازدید از حالت‌ها»، «کلیپ کردن پاداش»، یا «گسسته‌سازی مناسب حالت‌ها» اشاره کنید.

پروژه پایانی: تحلیل جامع و پیاده‌سازی خط لوله یادگیری ماشین (سوال آزاد)

صورت مسئله:

به عنوان آخرین بخش از ارزیابی این درس، دانشجویان موظف هستند یک مسئله «آزاد» را انتخاب کرده و یک خط لوله کامل یادگیری ماشین (Machine Learning Pipeline) را برای حل آن پیاده‌سازی کنند. هدف از این تمرین، ارزیابی توانایی شما در تحلیل داده، پیش‌پردازش، انتخاب مدل، تنظیم دقیق پارامترها و تفسیر علمی نتایج است. شما می‌توانید یک مجموعه داده (Dataset) استاندارد انتخاب کنید و یا بر روی هوشمندسازی یک سیستم خاص تمرکز نمایید.

الزامات و گام‌های انجام پروژه:

این پروژه باید شامل مراحل زیر باشد و گزارش نهایی باید تمامی این بخش‌ها را با جزئیات دقیق پوشش دهد:

۱. **معرفی مسئله و داده‌ها:**

- معرفی دقیق مجموعه داده انتخاب شده (منبع، تعداد ویژگی‌ها، تعداد نمونه‌ها، توزیع کلاس‌ها و ...).
- در صورت انتخاب «هوشمندسازی سیستم»، توصیف دقیق عملکرد سیستم و نحوه جمع‌آوری داده‌ها.

۲. **پیش‌پردازش داده‌ها (Data Preprocessing):**

- اعمال روش‌های مختلف پیش‌پردازش (مانند مدیریت داده‌های گم‌شده، نرمال‌سازی، استانداردسازی، مدیریت داده‌های پرت و ...).
- **تحلیل اثر:** بررسی و نمایش تاثیر هر یک از روش‌های پیش‌پردازش بر روی توزیع داده‌ها و نتایج نهایی.

۳. **کاهش ابعاد (Dimensionality Reduction):**

- پیاده‌سازی تکنیک‌های کاهش ابعاد (مانند PCA، LDA، t-SNE و ...).
- تحلیل اثر کاهش ابعاد بر روی دو فاکتور کلیدی: **دقت مدل** (Accuracy) و **سرعت آموزش/استنتاج** (Speed/Computational Cost).

۴. **انتخاب و آموزش مدل‌ها:**

- انتخاب حداقل **دو روش** یادگیری (روش‌های کلاسیک مطرح شده در درس یا روش‌های پیشرفته‌تر مانند Deep Learning).
- توجیه انتخاب مدل‌ها با توجه به نوع داده و مسئله.

۵. **تنظیم هایپرپارامترها (Hyperparameter Tuning):**

- استفاده از روش‌های جستجوی نظام‌مند مانند Grid Search، Random Search یا کتابخانه‌های بهینه‌سازی مانند Optuna.
- گزارش دقیق تمام هایپرپارامترهای جستجو شده و مقادیر بهینه نهایی.

۶. **ارزیابی و تکرارپذیری (Evaluation & Reproducibility):**

- تنظیم دقیق Random State (یا Seed) در تمام بخش‌های کد (تقسیم داده، وزن‌دهی اولیه مدل‌ها و ...) برای تضمین تکرارپذیری کامل نتایج.
- توضیح شفاف نحوه تقسیم داده‌ها (Train/Validation/Test Split) و منطق پشت آن (مثلاً Stratified K-Fold).

۷. **تحلیل نتایج و مصورسازی (Visualization & Analysis):**

- مقایسه نتایج حداقل دو روش انتخاب شده با معیارهای مختلف (مانند Precision، Recall، F1-Score، ROC-AUC).
- رسم تمامی نمودارهای ممکن و مرتبط (مانند Confusion Matrix، Learning Curves، Loss Curves، ROC Curve).
- تحلیل و تفسیر عمیق نتایج (چرا یک مدل بهتر عمل کرد؟ چرا مدل دچار Overfitting شد یا نشد؟).

نکته مهم در نمره‌دهی:

نمره نهایی بر اساس «عمق تحلیل»، «تنوع آزمایش‌ها» (مانند مقایسه سناریوهای مختلف پیش‌پردازش)، «کامل بودن نمودارها» و «دقت در پیاده‌سازی و گزارش‌دهی» تعیین می‌شود. هرچقدر مقایسه‌ها جامع‌تر و تحلیل‌ها علمی‌تر باشند، نمره بالاتری تعلق خواهد گرفت.