گزارش تکلیف چهارم درس الگوریتم های علوم داده

نام و نام خانوادگی: فاطمه ترودی

شماره دانشجویی: ۴۰۳۴۲۲۰۴۸

نام استاد: دکتر سعیدرضا خردپیشه

نیمسال دوم ۰۴-۱۴۰۳

فهرست مطالب

۱. مقدمه

این پروژه به منظور توسعه یک مدل طبقه‌بندی برای پیش‌بینی یک دسته از میان ۱۱ کلاس ممکن (با برچسب‌های ۰ تا ۱۰) بر اساس مجموعه‌ای از ۶۴ ویژگی باینری سنتز شده انجام شده است. این ویژگی‌ها، نشانگرهای "حضور/عدم حضور" را شبیه‌سازی می‌کنند. اهداف اصلی این پروژه شامل ساخت یک مدل پیش‌بینی‌کننده با دقت قابل قبول، انتخاب بهترین مدل از میان چندین کاندید، تفسیر یافته‌های مدل برای درک بهتر ویژگی‌های تأثیرگذار، و ارائه پیشنهادهای عملی بر اساس این درک بود .این پروژه در چارچوب یک رقابت Kaggle تعریف شده و ارزیابی نهایی مدل‌ها بر اساس دقت آن‌ها بر روی یک مجموعه داده تست پنهان انجام گرفته است. چالش اصلی این مسئله، ماهیت پیچیده و ظریف آن با وجود تعداد نسبتاً کم نمونه‌ها و ویژگی‌های باینری بود.

۲. شرح داده‌ها

داده‌های مورد استفاده در این پروژه از طریق پلتفرم Kaggle در دسترس قرار گرفتند و شامل سه فایل اصلی بودند:

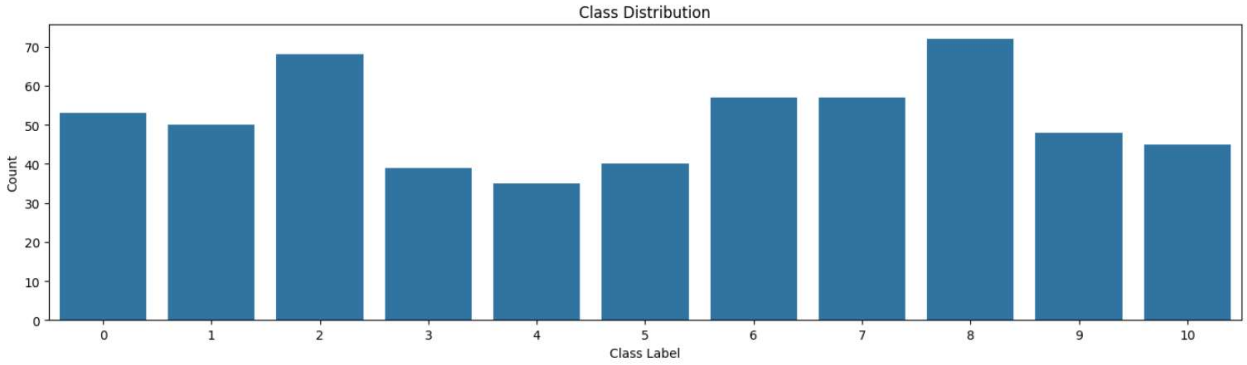
- **train.csv**: این فایل حاوی ۵۶۴ نمونه بود که برای آموزش و اعتبارسنجی مدل‌های یادگیری ماشین استفاده شد. هر نمونه شامل یک شناسه منحصربه‌فرد (ID)، ۶۴ ستون ویژگی باینری (با نام‌های feature0 تا feature63 که مقادیر ۰ یا ۱ را می‌پذیرفتند، و یک ستون هدف با نام label بود.
- **test.csv**: این فایل شامل ۱۴۳ نمونه بود که برای ارزیابی نهایی مدل در لیدربورد Kaggle مورد استفاده قرار گرفت. ساختار این فایل مشابه train.csv بود، با این تفاوت که فاقد ستون label بود.
- **sample_submission.csv**: این فایل یک نمونه از فرمت مورد انتظار برای فایل ارسالی به Kaggle را ارائه می‌داد که شامل دو ستون ID و label (با مقادیر پیش‌فرض) بود.

ویژگی‌ها و متغیر هدف: مجموعه داده شامل ۶۴ ویژگی ورودی بود که همگی از نوع باینری (۰ یا ۱) هستند. بر اساس توضیحات مسئله، این ویژگی‌ها شبیه‌ساز نشانگرهای "حضور/عدم حضور" هستند. متغیر هدف (label) یک متغیر طبقه‌ای با ۱۱ سطح (از ۰ تا ۱۰) است. هیچ مقدار گمشده‌ای در داده‌های آموزشی یا تست وجود نداشت که فرآیند پیش‌پردازش را ساده‌تر می‌کرد.

۳. تحلیل اکتشافی داده‌ها

پیش از شروع فرآیند مدل‌سازی، یک تحلیل اکتشافی جامع بر روی داده‌های آموزشی (train.csv) انجام شد تا درک بهتری از ویژگی‌ها و توزیع داده‌ها به دست آید.

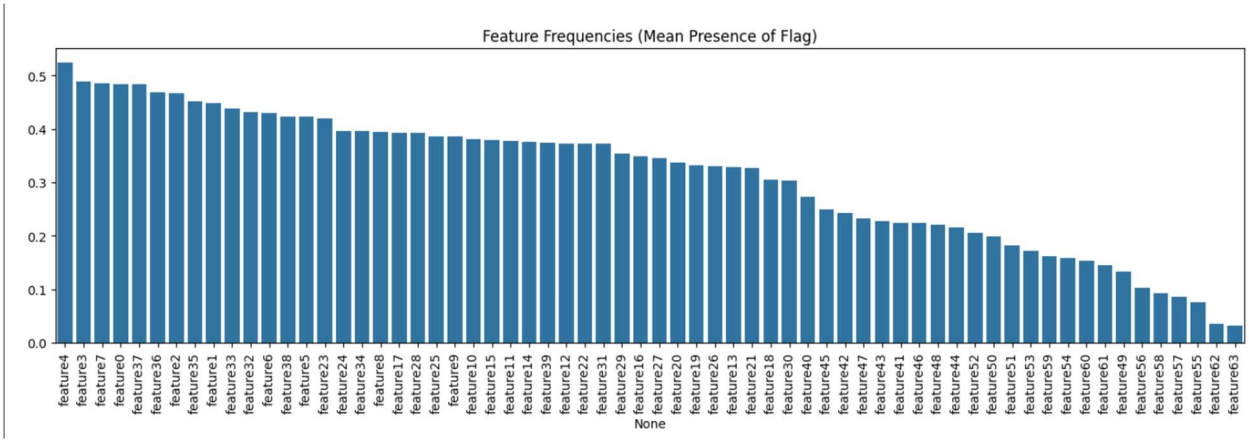
- **توزیع کلاس‌ها (متغیر هدف):** بررسی توزیع متغیر هدف (label) نشان داد که کلاس‌ها با عدم تعادل متوسطی مواجه هستند. از مجموع ۵۶۴ نمونه آموزشی:
  - کلاس ۸ با ۷۲ نمونه (۱۲.۷۷٪) بیشترین فراوانی را داشت.
  - کلاس ۴ با ۳۵ نمونه (۶.۲۱٪) کمترین فراوانی را داشت.
  - سایر کلاس‌ها نیز تعداد نمونه‌های متفاوتی داشتند (به عنوان مثال، کلاس ۰: ۵۳، کلاس ۱: ۵۰، کلاس ۲: ۶۸، کلاس ۳: ۳۹، کلاس ۵: ۴۰، کلاس ۶: ۵۷، کلاس ۷: ۵۷، کلاس ۹: ۴۸، کلاس ۱۰: ۴۵ نمونه) .این عدم تعادل در مراحل بعدی مانند تقسیم داده‌ها و انتخاب برخی پارامترهای مدل (مانند class_weight) مورد توجه قرار گرفت.



شکل ۱- نمودار میله‌ای توزیع متغیر کلاس‌ها (متغیر هدف)

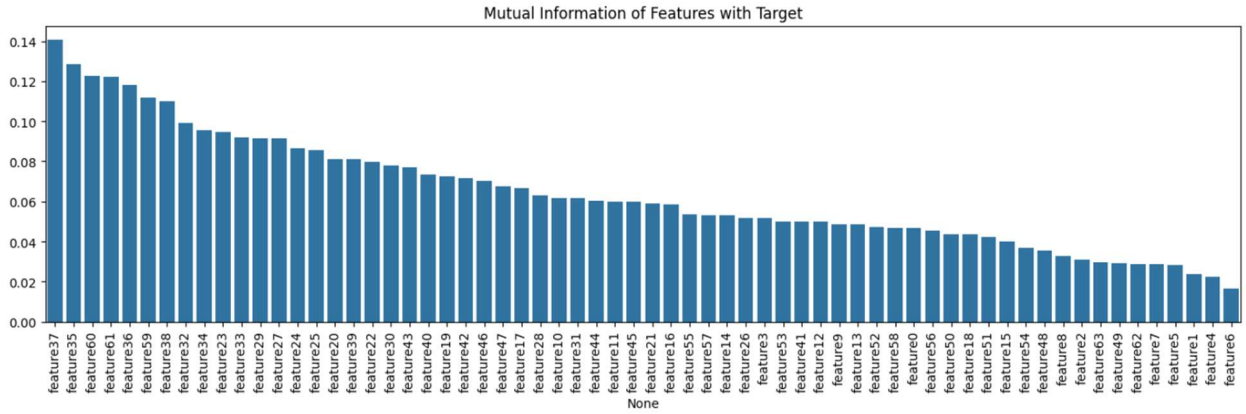
- **فراوانی حضور ویژگی‌ها:** تحلیل فراوانی مقدار ۱ برای هر یک از ۶۴ ویژگی باینری نشان داد که تنوع قابل توجهی در میزان رایج بودن ویژگی‌ها وجود دارد.
  - برخی ویژگی‌ها بسیار رایج بودند، به عنوان مثال feature4 در ۲۹۶ نمونه (حدود ۵۲.۵٪) از داده‌های آموزشی حضور داشت. ویژگی‌های دیگری مانند feature3 (۲۷۶ نمونه) و feature7 (۲۷۴ نمونه) نیز فراوانی بالایی داشتند.

- در مقابل، برخی ویژگی‌ها بسیار نادر بودند. به عنوان مثال، feature63 تنها در ۱۸ نمونه (حدود ۳.۲٪) و feature62 در ۲۰ نمونه (حدود ۳.۵٪) حضور داشتند .این توزیع متفاوت می‌توانست بر اهمیت و تأثیرگذاری هر ویژگی در مدل‌سازی تأثیر بگذارد.



شکل ۲- نمودار میله‌ای فراوانی حضور ویژگی‌ها

- **اطلاعات متقابل (Mutual Information) بین ویژگی‌ها و متغیر هدف:** برای ارزیابی اولیه میزان ارتباط و وابستگی هر ویژگی با متغیر هدف، از معیار اطلاعات متقابل استفاده شد. این معیار به ما کمک کرد تا ویژگی‌هایی را که به طور بالقوه اطلاعات بیشتری برای طبقه‌بندی ارائه می‌دهند، شناسایی کنیم.
  - ویژگی‌های feature37 (امتیاز $MI \approx 0.141$)، feature60 (امتیاز $MI \approx 0.123$)، feature35 (امتیاز $MI \approx 0.129$)، feature61 (امتیاز $MI \approx 0.122$) و feature36 (امتیاز $MI \approx 0.118$) بالاترین امتیاز اطلاعات متقابل را با لیبل کلاس داشتند.
  - یک یافته جالب این بود که برخی از ویژگی‌های با فراوانی پایین (مانند feature60 و feature61 که به ترتیب در ۸۶ و ۸۲ نمونه حضور داشتند) امتیاز اطلاعات متقابل بالایی کسب کردند. این نشان می‌داد که حتی نشانگرهای نادر نیز می‌توانند برای تشخیص کلاس‌ها بسیار مهم باشند.
  - در مقابل، برخی ویژگی‌های بسیار رایج (مانند feature4 که بیشترین فراوانی را داشت) امتیاز اطلاعات متقابل بالایی در بین ۱۰ ویژگی برتر نداشتند، که نشان می‌دهد فراوانی بالا لزوماً به معنای قدرت تفکیک بالا به تنهایی نیست.



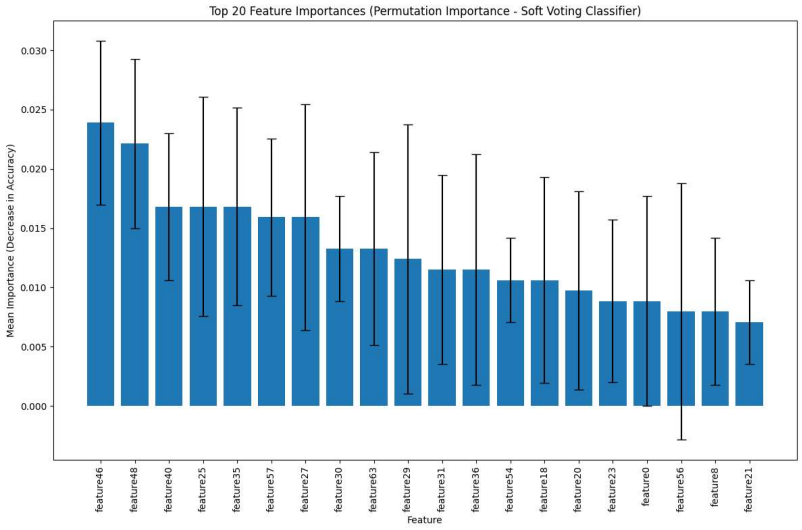
شکل ۳- نمودار میله‌ای امتیازات اطلاعات متقابل ویژگی‌ها با متغیر هد

۴. پیش‌پردازش داده‌ها و آماده‌سازی مدل

مراحل اصلی پیش‌پردازش و آماده‌سازی داده‌ها برای مدل‌سازی به شرح زیر بود:

- **جداسازی ویژگی‌ها و متغیر هدف:** از مجموعه داده آموزشی، ستون ID و label جدا شدند و مابقی ۶۴ ستون به عنوان ماتریس ویژگی‌ها (X) و ستون label به عنوان متغیر هدف (y) در نظر گرفته شدند.
- **تقسیم داده‌ها به مجموعه‌های آموزشی و اعتبارسنجی (Train/Validation Split)**: مجموعه داده آموزشی به نسبت ۸۰٪ برای آموزش (X_train, y_train) و ۲۰٪ برای اعتبارسنجی (X_val, y_val) تقسیم شد. برای اطمینان از اینکه توزیع کلاس‌ها در هر دو مجموعه آموزشی و اعتبارسنجی مشابه توزیع کلی باشد (با توجه به عدم تعادل مشاهده شده)، از روش نمونه‌گیری طبقه‌بندی شده (stratify=y) استفاده شد. مقدار random_state=42 برای تمام مراحل تقسیم داده و مدل‌سازی (در صورت وجود پارامتر مشابه) به منظور تضمین تکرارپذیری نتایج تنظیم گردید. پس از تقسیم، مجموعه آموزشی شامل ۴۵۱ نمونه و مجموعه اعتبارسنجی شامل ۱۱۳ نمونه بود.
- **اسکیل کردن ویژگی‌ها:** از آنجایی که تمام ویژگی‌های ورودی باینری (۰ یا ۱) بودند، برای بسیاری از مدل‌ها مانند Naive Bayes، Decision Tree و Random Forest، اسکیل کردن ویژگی‌ها ضرورت نداشت. با این حال، برای مدل‌هایی که به مقیاس ویژگی‌ها حساس هستند، مانند Support Vector Machines (SVM) با کرنل RBF و همچنین قبل از اعمال PCA، از StandardScaler استفاده شد. این scaler بر روی داده‌های آموزشی فیت شده و سپس برای تبدیل همان داده‌های آموزشی و همچنین داده‌های اعتبارسنجی و در نهایت داده‌های تست (X_test) استفاده شد.

- **انتخاب ویژگی‌ها (Feature Selection)**: پس از بررسی اولیه مدل‌ها، یک مرحله انتخاب ویژگی انجام شد. بر اساس نتایج Permutation Importance که برای مدل Soft Voting Classifier اولیه (آموزش دیده روی تمام ۶۴ ویژگی) محاسبه گردید، ۳۰ ویژگی برتر که بیشترین تأثیر را بر دقت مدل داشتند، انتخاب شدند. مدل Soft Voting Classifier بر روی این ۳۰ ویژگی منتخب نیز آموزش داده شد. ویژگی‌های منتخب عبارت بودند از:



شکل ۴- نمودار میله‌ای ۳۰ ویژگی برتر در مدل Soft Voting Classifier

۵. انتخاب، آموزش و تنظیم هایپرپارامترهای مدل‌ها

در این تمرین، چندین خانواده مختلف از مدل‌های طبقه‌بندی مورد بررسی و آزمایش قرار گرفتند. برای مدل‌های کلیدی، فرآیند تنظیم هایپرپارامترها با استفاده از GridSearchCV و اعتبارسنجی متقابل (Cross Validation) انجام شد تا بهترین ترکیب پارامترها برای دستیابی به بالاترین دقت بر روی مجموعه داده اعتبارسنجی شناسایی شود. معیار اصلی برای انتخاب بهترین پارامترها، دقت کلی (accuracy) بود.

- **مدل‌های پایه (Baseline Models)**:
  - **Logistic Regression**:
    - عملکرد اولیه (بدون تیونینگ): دقت اعتبارسنجی ۰.۳۲۷۴. این مدل در تفکیک بسیاری از کلاس‌ها با مشکل مواجه بود.

- پس از تیونینگ هایپرپارامترها شامل C، penalty، solver و class_weight با GridSearchCV بهترین پارامترهای یافت شده عبارت بودند از:

{'C': 0.1, 'class_weight': None, 'max_iter': 1000, 'penalty': 'l2', 'solver': 'liblinear'}

- دقت اعتبارسنجی مدل تیون شده به ۰.۳۵۴۰ رسید. با این حال، این مدل در پیش‌بینی کلاس‌های ۱ و ۵ به طور کامل ناموفق بود (F1-Score =0).

- **Bernoulli Naive Bayes**:
  - این مدل به دلیل ماهیت باینری ویژگی‌ها، یک کاندید مناسب بود. عملکرد اولیه آن (با پارامترهای پیش‌فرض) بسیار خوب بود و به دقت اعتبارسنجی ۰.۳۸۰۵ دست یافت، اما در پیش‌بینی کلاس‌های ۱ و ۸ ناموفق بود.
  - پس از تیونینگ هایپرپارامتر alpha با GridSearchCV بهترین alpha یافت شده حدود ۰.۰۰۲۶ بود. این مدل تیون شده به دقت اعتبارسنجی ۰.۳۹۸۲ رسید که بالاترین دقت برای یک مدل تکی بود. با این حال، این مدل تیون شده نیز در پیش‌بینی کلاس ۸ به طور کامل ناموفق بود (F1-score=0) اما توانست کلاس ۱ را تا حدی پیش‌بینی کند.

- **Decision Tree Classifier**:
  - عملکرد اولیه با پارامترهای پیش‌فرض یا max_depth=10 ضعیف بود (دقت اعتبارسنجی ۰.۲۳۸۹) و در پیش‌بینی کلاس‌های ۱ و ۵ ناموفق بود.
  - پس از یک جستجوی گسترده برای هایپرپارامترها شامل max_depth، min_samples_split، min_samples_leaf و class_weight با هدف یافتن یک درخت خوب و تعمیم‌پذیر، بهترین پارامترهای یافت شده عبارت بودند از:

{'ccp_alpha': 0.025, 'class_weight': 'balanced', 'criterion': 'entropy', 'max_depth': None, 'min_samples_leaf': 5, 'min_samples_split': 2}

  - این مدل تیون شده به دقت اعتبارسنجی ۰.۳۱۸۶ رسید. اگرچه نسبت به حالت اولیه بهبود داشت، اما در پیش‌بینی کلاس ۲ کاملاً ناموفق بود و عملکرد کلی آن پایین‌تر از سایر مدل‌های برتر بود.

- **Support Vector Machine (SVM)**:
  - **Linear SVM (LinearSVC)**: با استفاده از داده‌های اسکیل شده، این مدل به دقت اعتبارسنجی حدود ۰.۳۱۸۶ رسید که مشابه Logistic Regression بود.
  - **RBF SVM**: پس از تیونینگ هایپرپارامترها (C, gamma, class_weight) با GridSearchCV روی داده‌های اسکیل شده، بهترین پارامترهای یافت شده عبارت بودند از:

{'C': 0.5, 'class_weight': 'balanced', 'gamma': 'scale', 'kernel': 'rbf'}

  - این مدل تیون شده به دقت اعتبارسنجی ۰.۳۸۰۵ دست یافت. با این حال، این مدل در پیش‌بینی کلاس ۵ کاملاً ناموفق بود.

- **مدل‌های ترکیبی (Ensemble Models)**:
  - **Random Forest Classifier**:
    - عملکرد اولیه با ۱۰۰ درخت ضعیف بود (دقت اعتبارسنجی ۰.۳۰۰۹).
    - پس از یک جستجوی گسترده برای هایپرپارامترها شامل n_estimators، max_depth، min_samples_split، min_samples_leaf، max_features و class_weight با GridSearchCV بهترین پارامترهای یافت شده عبارت بودند از:

{'class_weight': None, 'max_depth': 5, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 150}

    - این مدل تیون شده به دقت اعتبارسنجی ۰.۳۸۰۵ رسید. با این حال، این مدل در پیش‌بینی کلاس‌های ۱ و ۵ کاملاً ناموفق بود.

  - **Soft Voting Classifier با تمام ۶۴ ویژگی**: این مدل از ترکیب سه مدل تیون شده برتر که روی تمام ۶۴ ویژگی آموزش دیده بودند، ساخته شد: Tuned Bernoulli Naive Bayes، Tuned RBF SVM و Tuned Random Forest. از روش رای‌گیری نرم (Soft Voting) استفاده شد.
    - دقت اعتبارسنجی این مدل ۰.۳۹۸۲ بود.

- امتیاز کسب شده در Kaggle برای این مدل: ۰.۳۶۸۴۲. گزارش طبقه‌بندی این مدل روی داده اعتبارسنجی نشان داد که در پیش‌بینی کلاس ۸ ناموفق است

Classification Report on Validation Set (Voting Classifier - Soft):

	precision	recall	f1-score	support
0	0.77	0.91	0.83	11
1	0.27	0.30	0.29	10
2	0.33	0.14	0.20	14
3	0.56	0.62	0.59	8
4	0.30	0.43	0.35	7
5	1.00	0.12	0.22	8
6	0.33	0.27	0.30	11
7	0.58	0.64	0.61	11
8	0.00	0.00	0.00	14
9	0.27	0.70	0.39	10
10	0.36	0.44	0.40	9
accuracy			0.40	113
macro avg	0.43	0.42	0.38	113
weighted avg	0.41	0.40	0.37	113

شکل ۵- گزارش classification مدل soft voting بر روی داده‌های اعتبارسنجی

- **Soft Voting Classifier با ۳۰ ویژگی منتخب:** این مدل مشابه مدل قبلی بود، اما مولفه‌های آن بر روی ۳۰ ویژگی برتر) انتخاب شده بر اساس (Permutation Importance آموزش دیدن
  - دقت اعتبارسنجی این مدل به ۰.۴۱۵۹ رسید که بالاترین دقت اعتبارسنجی در بین تمام مدل‌های آزمایش شده بود.
  - امتیاز کسب شده در Kaggle برای این مدل: ۰.۳۲۸۹۴. گزارش طبقه‌بندی این مدل روی داده اعتبارسنجی نشان داد که اگرچه هیچ کلاسی را به طور کامل از دست نمی‌دهد، اما عملکرد آن برای برخی کلاس‌ها (مانند ۵، ۶ و ۸) همچنان ضعیف است.

Classification Report on Validation Set (Soft Voting - Feature Selection):

	precision	recall	f1-score	support
0	0.67	0.91	0.77	11
1	0.38	0.30	0.33	10
2	0.36	0.29	0.32	14
3	0.60	0.75	0.67	8
4	0.43	0.43	0.43	7
5	0.33	0.12	0.18	8
6	0.25	0.18	0.21	11
7	0.57	0.73	0.64	11
8	0.20	0.07	0.11	14
9	0.35	0.60	0.44	10
10	0.20	0.33	0.25	9
accuracy			0.42	113
macro avg	0.39	0.43	0.40	113
weighted avg	0.39	0.42	0.39	113

شکل ۶- گزارش classification مدل soft voting با ۳۰ ویژگی منتخب بر روی داده‌های اعتبارسنجی

- **PCA (Principal Component Analysis)**: این روش بر روی داده‌های اسکیل شده با ۶۴ ویژگی امتحان شد و تعداد اجزا برای حفظ ۹۵٪ واریانس انتخاب گردید. سپس Soft Voting Classifierبا همان ساختار قبلی بر روی این داده‌های تبدیل‌شده با PCA آموزش داده شد. نتیجه بسیار ضعیف بود و دقت اعتبارسنجی به ۰.۲۷ کاهش یافت. دلایل احتمالی شامل از دست رفتن اطلاعات مهم و عدم سازگاری برخی مدل‌ها با داده‌های PCA بود.

۶. نتایج و مقایسه مدل‌ها

جدول زیر خلاصه‌ای از دقت اعتبارسنجی (Validation Accuracy) و امتیاز Kaggle (Test Accuracy) برای مدل‌های کلیدی را نشان می‌دهد:

مدل	دقت اعتبارسنجی	امتیاز Kaggle	نکات کلیدی عملکرد (روی اعتبارسنجی)
Logistic Regression (Tuned)	0.3540	ارسال نشده	ناموفق در کلاس‌های ۱ و ۵
Bernoulli Naive Bayes (Tuned)	0.3982	0.3552	ناموفق در کلاس ۸، بهترین مدل تکی در اعتبارسنجی
RBF SVM (Tuned)	0.3805	0.3421	ناموفق در کلاس ۵
Random Forest (Tuned)	0.3805	0.3026	ناموفق در کلاس‌های ۱ و ۵
Soft Voting (64 feats)	0.3982	0.36842	بهترین امتیازKaggle ، ناموفق در کلاس ۸ (روی اعتبارسنجی)
Soft Voting (30 feats)	0.4159	0.32894	بالاترین دقت اعتبارسنجی، اما امتیاز Kaggleپایین‌تر، کلاس‌های ضعیف دارد

همانطور که مشاهده می‌شود، مدل Soft Voting با ۳۰ ویژگی بالاترین دقت را روی مجموعه اعتبارسنجی داخلی ما کسب کرد، اما مدل Soft Voting با ۶۴ ویژگی عملکرد بهتری روی داده تست پنهان Kaggle داشت. این پدیده نشان‌دهنده اهمیت ارزیابی نهایی بر روی داده‌های تست کاملاً مستقل و همچنین احتمال بیش‌برازش مدل ۳۰ ویژگی بر روی مجموعه اعتبارسنجی ما است.

۷. انتخاب مدل نهایی و دلیل آن

با توجه به هدف اصلی که کسب بهترین عملکرد بر روی داده‌های تست نهایی Kaggle است، مدل Soft Voting Classifierبا استفاده از تمام ۶۴ ویژگی به عنوان مدل نهایی این پروژه انتخاب شد. این مدل توانست امتیاز ۰.۳۶۸۴۲ را در لیدربورد Kaggle کسب کند که بالاترین امتیاز در بین مدل‌ها بود.

دلایل انتخاب این مدل:

۱. **بهترین عملکرد روی داده تست Kaggle**: این مهم‌ترین معیار برای انتخاب مدل در چارچوب این رقابت است.
۲. **ماهیت ترکیبی (Ensemble)**: این مدل از ترکیب سه مدل مختلف با روش رأی‌گیری نرم بهره می‌برد که به طور بالقوه می‌تواند با ترکیب نقاط قوت هر یک از مدل‌های پایه، به نتایج قوی‌تر و پایدارتری منجر شود.
۳. **استفاده از تمام ویژگی‌ها**: در حالی که انتخاب ویژگی منجر به دقت اعتبارسنجی بالاتری شد، عملکرد ضعیف‌تر آن مدل روی داده تست Kaggle نشان می‌دهد که شاید تمام ۶۴ ویژگی حاوی اطلاعات مفیدی بودند که حذف بخشی از آن‌ها منجر به کاهش قدرت تعمیم مدل روی داده‌های کاملاً جدید شده است.

محدودیت‌های مدل نهایی: بر اساس ارزیابی روی مجموعه داده اعتبارسنجی، این مدل (با دقت ۰.۳۹۸۲) در پیش‌بینی کلاس ۸ به طور کامل ناموفق بود. اگرچه امتیاز Kaggle آن (۰.۳۶۸۴۲) از آستانه قبولی (۰.۳۵) بالاتر است، اما این ضعف در تشخیص یک کلاس خاص باید در نظر گرفته شود. عملکرد آن در سایر کلاس‌ها نیز متفاوت بود

۸. تفسیر مدل (Model Interpretation)

برای تفسیر مدل نهایی Soft Voting Classifier که با تمام ۶۴ ویژگی آموزش دیده، از تحلیل اهمیت ویژگی با استفاده از آزمون جایگشت (Permutation Importance) که مستقیماً روی این مدل ترکیبی اجرا شد، بهره می‌بریم. این روش، اهمیت هر ویژگی را با اندازه‌گیری میزان کاهش دقت مدل در صورت به هم ریختن تصادفی مقادیر آن ویژگی، ارزیابی می‌کند.

- **مهم‌ترین ویژگی‌ها بر اساس Permutation Importance**: نتایج آزمون جایگشت نشان داد که:

- مهم‌ترین ویژگی‌ها عبارتند از: feature46 (با کاهش دقت میانگین حدود ۰.۰۲۳۹)، feature48 (۰.۰۲۲۱) و سپس feature25، feature40 و feature35 (هر کدام با کاهش دقت حدود ۰.۰۱۶۸).
- در مجموع ۴۸ ویژگی از ۶۴ ویژگی اهمیت مثبت داشتند، که نشان می‌دهد مدل از اکثر ویژگی‌ها تا حدی استفاده می‌کند.
- حداکثر تأثیر یک ویژگی به تنهایی حدود ۲.۴٪ کاهش دقت بود که ماهیت ظریف مسئله و اتکای مدل به ترکیبی از ویژگی‌ها را تأیید می‌کند.

- **دو پیشنهاد مشخص:**

- با توجه به اینکه ویژگی‌های feature46، feature48، feature40، feature25 و feature35 بیشترین تأثیر را بر دقت مدل نهایی دارند، پیشنهاد می‌شود در فرآیندهای آتی جمع‌آوری داده یا در صورت امکان، بر صحت و کامل بودن ثبت این نشانگرهای خاص تمرکز ویژه‌ای شود.
- دانش به‌دست‌آمده از اهمیت ویژگی‌ها می‌تواند در توسعه سیستم‌های پشتیبانی از تصمیم‌گیری اولیه مورد استفاده قرار گیرد. به عنوان مثال، یک سیستم امتیازدهی وزنی بر اساس میزان اهمیت ویژگی‌های برتر (مانند موارد ذکر شده در پیشنهاد اول) می‌تواند برای شناسایی نمونه‌هایی با الگوی غیرطبیعی قوی ایجاد شود. این نمونه‌ها می‌توانند برای بررسی‌های جامع‌تر و تخصصی‌تر در اولویت قرار گیرند تا وضعیت دقیق آن‌ها از میان ۱۱ کلاس ممکن مشخص شود.

۹. نتیجه‌گیری

در این پروژه، هدف توسعه یک مدل Classification دقیق برای یک مسئله ۱۱ کلاسه با ۶۴ ویژگی باینری بود. پس از انجام تحلیل داده‌های اکتشافی، پیش‌پردازش، آزمایش چندین خانواده از مدل‌های یادگیری ماشین و تنظیم هایپرپارامترهای آن‌ها، و همچنین بررسی روش‌های انتخاب ویژگی، مدل Soft Voting Classifier با استفاده از تمام ۶۴ ویژگی به عنوان بهترین مدل انتخاب شد. این مدل توانست به امتیاز ۰.۳۶۸۴۲ در لیدربورد Kaggle دست یابد که از حداقل امتیاز قبولی فراتر رفته و نشان‌دهنده عملکرد قابل قبول مدل است. تحلیل اهمیت ویژگی‌ها با استفاده از آزمون جایگشت نشان داد که ویژگی‌هایی feature46، feature48، feature40، feature25 و feature35 مانند بیشترین تأثیر را بر عملکرد این مدل دارند. اگرچه مدل نهایی در پیش‌بینی تمام کلاس‌ها به طور یکسان موفق نبود (و بر اساس ارزیابی روی داده اعتبارسنجی، در پیش‌بینی کلاس ۸ ضعف داشت)، اما رویکرد ترکیبی توانست عملکردی بهتر از اکثر مدل‌های تکی در شرایط تست واقعی

ارائه دهد. چالش اصلی این مسئله، ماهیت ظریف و پیچیده آن با وجود تعداد نمونه‌های محدود برای هر کلاس بود که حتی با روش‌های پیشرفته نیز بهبود بیشتر دقت را دشوار می‌کرد.