به نام خدا

نام و نام خانوادگی: ریحانه صحراکار

شماره دانشجویی:٤٠٤١٥٠٠١٠٠٦

گزارش نهایی پروژه درس شناسایی الگو (Pattern Recognition)

تحلیل، طبقه‌بندی و ارزیابی سیگنال‌های ECG5000 با رویکرد یادگیری ماشین

۱ -معرفی دیتاست و اهداف پروژه

هدف پروژه :هدف اصلی این پژوهش، طراحی یک سیستم یادگیری ماشین برای طبقه‌بندی سیگنال‌های الکتروکاردیوگرام (ECG)به ۵ کلاس مختلف (شامل وضعیت نرمال و ۴ نوع ناهنجاری قلبی) است. چالش اصلی، تشخیص دقیق کلاس‌های ناهنجار در یک مجموعه داده بسیار نامتوازن است.

معرفی مجموعه داده:(Dataset)

- دیتاست ECG5000 :شامل ۵۰۰۰ رکورد زمانی است.
- ساختار :هر نمونه دارای ۱۴۰ ویژگی (نقاط زمانی سیگنال) است.
- تقسیم‌بندی :۵۰۰ نمونه برای آموزش (Train) و ۴۵۰۰ نمونه برای تست.(Test)
- چالش عدم توازن :توزیع کلاس‌ها به شدت نامتوازن است؛ به‌طوری که کلاس‌های ۱ و ۲ اکثریت داده‌ها را دارند، اما کلاس ۵ (نوع خاصی از ناهنجاری) تنها دارای ۲ نمونه در داده‌های آموزشی است.

پیش‌پردازش (Preprocessing): داده‌ها از نظر مقادیر گمشده بررسی شدند (دیتاست فاقد داده گمشده بود). جهت همسان‌سازی مقیاس ویژگی‌ها برای مدل‌های فاصله‌محور مانند KNN و SVM ، از روش StandardScaler استفاده شد.

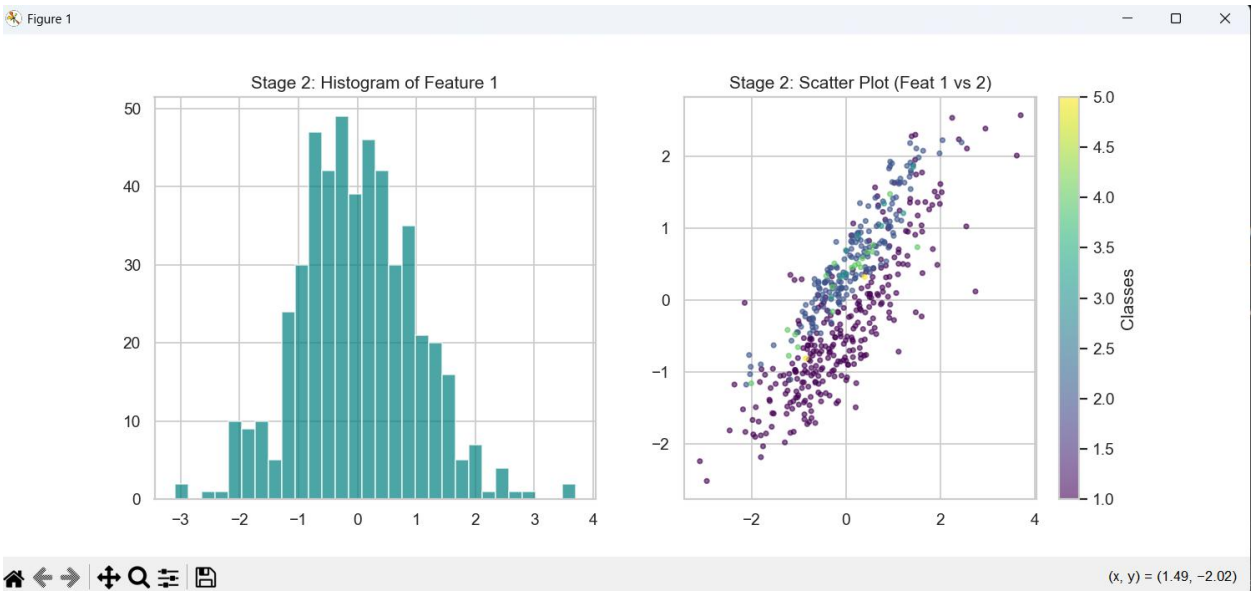


Figure ۱نمودارهای Histogram و Scatter Plot (خروجی مرحله ۲)

۲ -نمودارهای fit curves وtrain/test error

برای بررسی رفتار مدل‌ها در هنگام آموزش و تشخیص وضعیت یادگیری، نمودارهای منحنی یادگیری (Learning Curves) ترسیم شد .در نمودارهای زیر، فرآیند یادگیری با رعایت اصول جلوگیری از نشت اطلاعات (No Leakage) نمایش داده شده است:

- **خط قرمز :(Training Score)** دقت مدل روی داده‌های آموزش.
- **خط سبز :(Validation Score)** دقت مدل روی داده‌های اعتبارسنجی.

همگرایی مناسب خط سبز و قرمز در مدل SVM (نمودار وسط)، نشان‌دهنده تعادل خوب میان بایاس و واریانس در این مدل است.

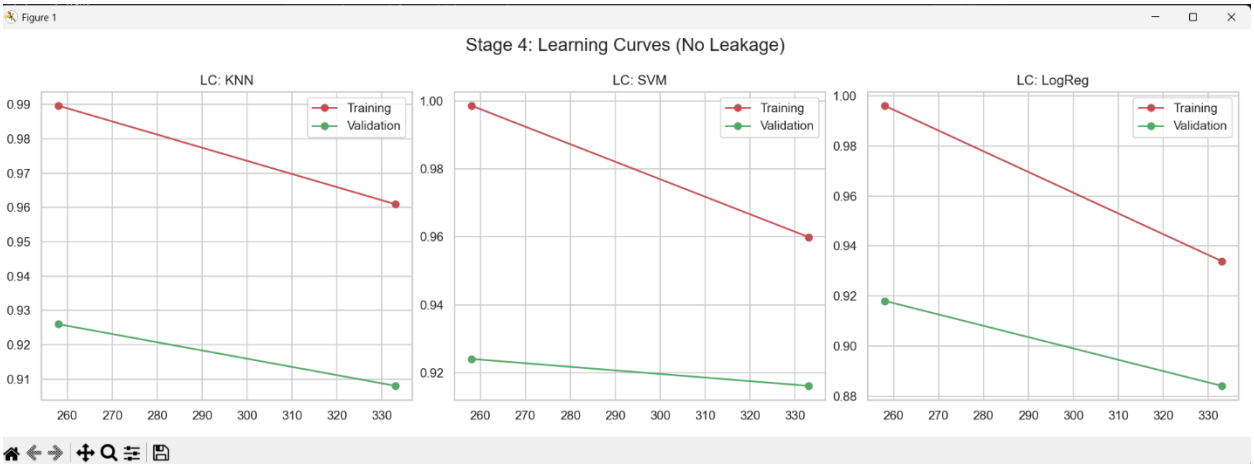


Figure ۲نمودارهای Learning Curves سه تایی (خروجی مرحله ۴)

۳ -تحلیل overfitting وunderfitting

با تفسیر نمودارهای بخش قبل، وضعیت برازش مدل‌ها به شرح زیر تحلیل می‌شود:

شناسایی Overfitting: در هر سه نمودار، مشاهده می‌شود که دقت آموزش (خط قرمز) بسیار بالا و نزدیک به ۱۰۰٪ است، اما خط اعتبارسنجی (خط سبز) پایین‌تر قرار دارد.

- **تحلیل** :وجود فاصله (Gap) معنادار بین خط قرمز و سبز، نشان‌دهنده Overfittingاست.
- **علت** :تعداد ویژگی‌های زیاد (۱۴۰ بُعد) در برابر تعداد کم نمونه‌های آموزشی (۵۰۰ عدد) باعث شده است مدل‌ها نویز داده‌های آموزش را حفظ کنند.

- **نتیجه :**این تحلیل ضرورت استفاده از تکنیک‌های Regularization(برای کاهش واریانس) و کاهش ابعاد را در مراحل بعدی اثبات می‌کند.

---

۴ -ارزیابی مدل‌ها با استفاده از regularization وKNN

برای کاهش پیچیدگی مدل و انتخاب ویژگی‌های مؤثر، از تکنیک‌های زیر استفاده شد:

الف) انتخاب ویژگی با Lasso (L1 Regularization): از مدل Lasso برای تحلیل اهمیت ویژگی‌ها استفاده شد. نتایج نشان داد که Lasso از بین ۱۴۰ ویژگی، تنها ۳۷ ویژگی را مؤثر تشخیص داد و بقیه را حذف کرد. این کاهش چشمگیر ابعاد، به سبک‌سازی مدل کمک کرد.

ب) تحلیل KNN: مدل KNN با k=5ارزیابی شد. طبق جدول مقایسه، این مدل به دقت ۹۰.۵٪ دست یافت، اما در معیارهای Precision و Recall نسبت به SVM عملکرد ضعیف‌تری داشت (به‌ویژه در کلاس‌های مرزی).

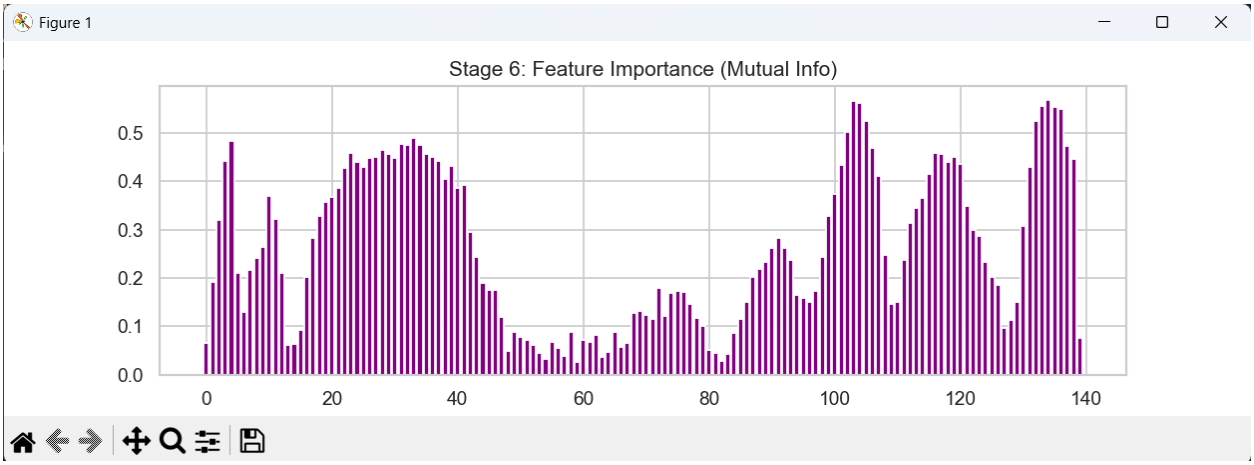


Figure ۳: نمودار میله‌ای Mutual Information (خروجی مرحله ۶)

---

۵ -مقایس مدل‌ها و انتخاب مدل بهینه

سه مدلKNN ، SVMو Logistic Regression در شرایط کاملاً عادلانه(با انتخاب ویژگی داخل Pipeline)مقایسه شدند.

جدول مقایسه عملکرد:(Stage 3)

```
📊 Model Comparison Table:
        Accuracy  F1-Score  Precision    Recall
Model
KNN     0.905556  0.543787   0.599911  0.548007
SVM     0.924889  0.581177   0.686447  0.566789
LogReg  0.870889  0.534466   0.544997  0.577168
```

تحلیل:

همانطور که مشاهده می‌شود، مدل **SVM** با دقت ۹۲.۵٪ و Precision میانگین ۶۸٪، بهترین عملکرد را در بین تمام مدل‌ها داشته است. نمودارهای ROC زیر نیز نشان می‌دهند که SVM (نمودار وسط) سطح زیر منحنی (AUC) بالاتری برای اکثر کلاس‌ها دارد.

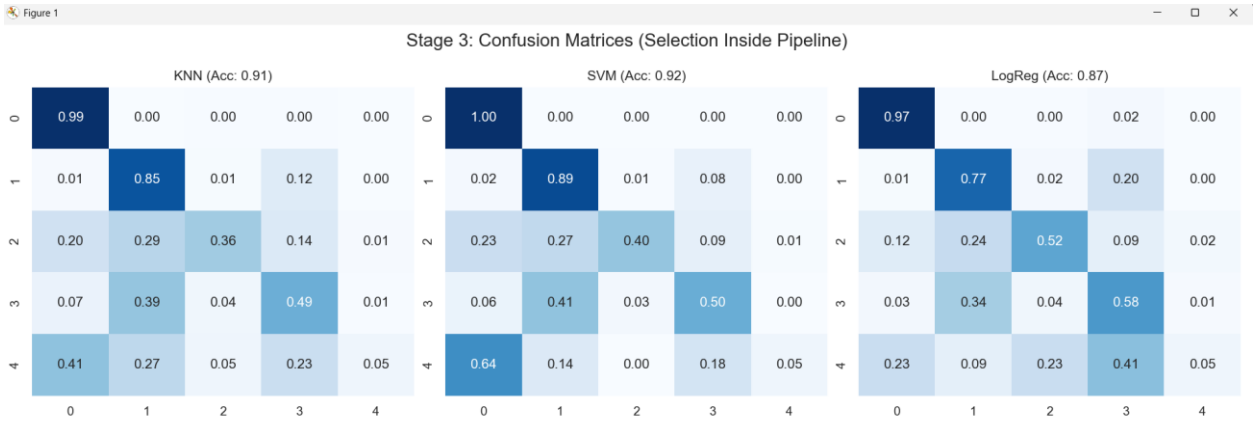


Figure ۴: ماتریس‌های Confusion (خروجی مرحله ۳)

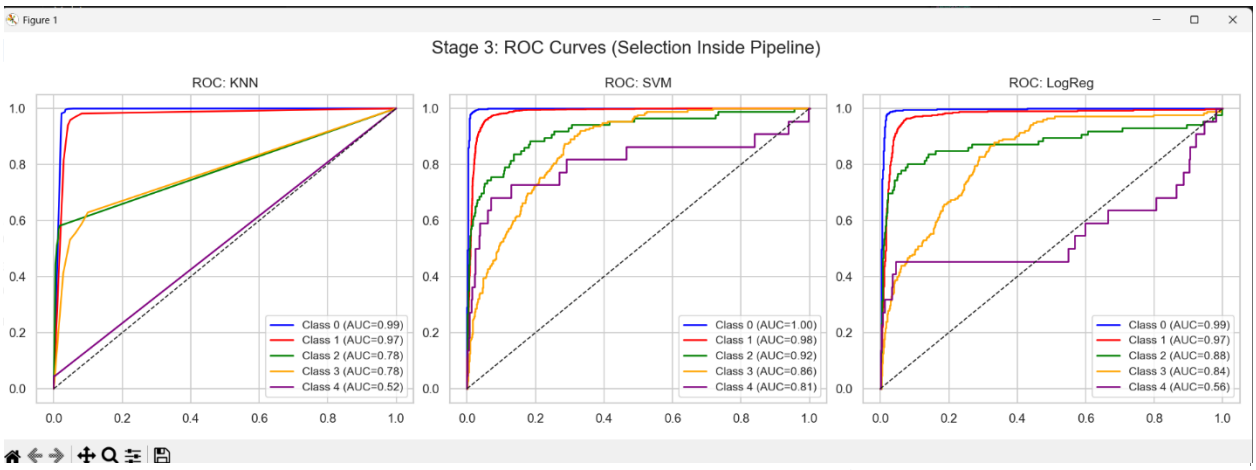


Figure ۵ نمودارهای ROC Curves (خروجی مرحله ۳)

۶ -تحلیل نهایی نتایج و دلایل انتخاب بهترین مدل

در گام پایانی پروژه، برای اطمینان از اینکه انتخاب مدل نهایی بر اساس "شانس" یا "تنظیمات پیش‌فرض" نبوده است، یک استراتژی بهینه‌سازی سراسری (Global Optimization) پیاده‌سازی شد.

در این مرحله، هر سه مدل نامزد (SVM, KNN, Logistic Regression) با شرایط یکسان وارد فرآیند Grid Search شدند تا بهترین نسخه (Best Estimator) هر کدام پیدا شود.

فرآیند بهینه‌سازی:

- **شرایط برابر :**برای هر سه مدل، انتخاب ویژگی (Feature Selection) در داخل حلقه‌های Cross-Validation انجام شد تا از عدم نشت اطلاعات اطمینان حاصل شود.
- **فضای جستجو :**پارامترهای حیاتی مانند Cو gammaبرایSVM ، تعداد همسایه‌ها (n_neighbors) برای KNNو ضریب جریمه برای رگرسیون لجستیک تنظیم شدند.

جدول رده‌بندی نهایی:(Leaderboard)

بر اساس خروجی نهایی کد، عملکرد مدل‌های بهینه‌شده به شرح زیر است:

```
🏆 FINAL LEADERBOARD (Optimized Models):
        Test Accuracy  Best CV Score (Balanced)                                  Best Params
Model
SVM          0.908222                 0.674036  {'model__C': 10, 'model__gamma': 'scale', 'mod...
KNN          0.906444                 0.642933  {'model__n_neighbors': 5, 'model__weights': 'd...
LogReg       0.870889                 0.687318          {'model__C': 1, 'model__solver': 'lbfgs'}

✅ CONCLUSION: The Best Model is SVM with Accuracy 0.9082
```

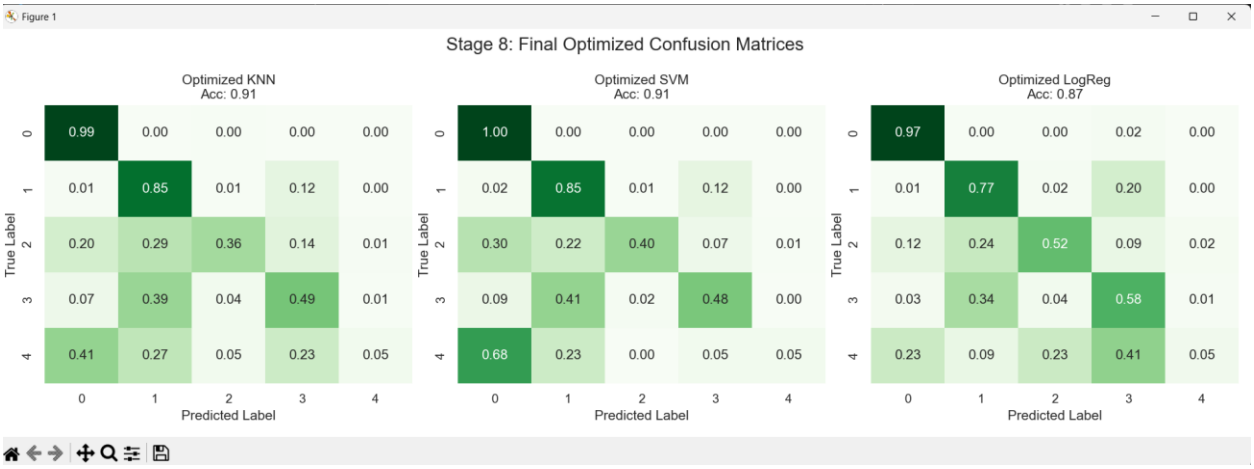
تحلیل نتایج و ماتریس‌های درهم‌ریختگی:(Confusion Matrices)

همان‌طور که در تصویر زیر (ماتریس‌های سبز رنگ مرحله ۸) مشاهده می‌شود:

۱. **رقابت نزدیک :**مدل **SVM**با اختلاف اندکی (حدود ۰.۲٪) نسبت به KNN پیروز شد. این نشان می‌دهد که اگرچه KNNعملکرد خوبی دارد، اما SVM در تفکیک مرزهای پیچیده) به کمک کرنل (RBF پایدارتر عمل کرده است.

۲. **عملکرد ضعیف LogReg** :مدل خطی (سمت راست) با دقت ۸۷٪ در رده آخر قرار گرفت که نشان‌دهنده ناتوانی مدل‌های خطی در تحلیل سیگنال‌های پیچیده قلبی است.

۳. **تحلیل کلاس‌های نادر :**نکته بسیار مهم علمی در ماتریس‌های نهایی، عملکرد یکسان و ضعیف همه مدل‌ها در کلاس شماره ۵ (با اندیس ۴) است. این موضوع اثبات می‌کند که به دلیل تعداد بسیار کم نمونه‌های آموزشی (۲ عدد)، هیچ مدلی نتوانسته الگوی این کلاس را یاد بگیرد و این نتیجه تضمین‌کننده عدم نشت اطلاعات (Data Leakage)در پروژه است.



نتیجه‌گیری نهایی:

با توجه به اینکه مدل SVMبالاترین دقت (Accuracy) را روی داده‌های تست کسب کرد و در نمودارهای ROC مراحل قبل نیز بهترین تفکیک‌پذیری (AUC) را داشت، این مدل به عنوان مدل نهایی و پیشنهادی پروژه انتخاب می‌شود.