



دانشگاه صنعتی خواجه نصیرالدین طوسی  
دانشکده مهندسی برق

یادگیری ماشین

پاسخ مینی پروژه‌ی شماره ۳

نام و نام خانوادگی	مهدی خداپنده لو
شماره دانشجویی	۴۰۱۰۵۳۱۴
تاریخ	بهار ۱۴۰۳



## فهرست مطالب

۵	۱ پرسش یک
۵	۱.۱ در مرحله اول دیتاست را فراخوانی کنید و اطلاعاتی نظیر ابعاد، تعداد نمونه ها، میانگین، واریانس و همبستگی ویژگی ها را به دست آورید و نمونه های دیتاست را به تصویر بکشید سپس، با توجه به اطلاعات عددی، آماری و بصری بدست آمده، تحلیل کنید که آیا کاهش ابعاد می تواند در این دیتاست قابل استفاده باشد یا خیر . . . . .
۵	۲.۱ داده ها را با استفاده از الگوریتم SVM با هسته ی خطی دسته بندی کنید و ماتریس درهم ریختگی آن را بدست آورید و مرزهای تصمیم گیری را ترسیم کنید . . . . .
۶	۳.۱ بخش قبلی را با استفاده از هسته های چند جمله ای و با استفاده از کتابخانه از scikit-learn درجه یک تا ۱۰ پیاده سازی کنید و نتایج را با معیارهای مناسب گزارش کرده و مقایسه و تحلیل کنید. در نهایت، با استفاده از کتابخانه imageio . . . . .
۷	۲ پرسش سه
۱۰	۱.۲ بزرگ ترین چالش ها در توسعه مدل های تشخیص تقلب چیست؟ این مقاله برای حل این چالش ها از چه روش هایی استفاده کرده است؟ . . . . .
۱۰	۲.۲ در مورد معماری شبکه ارائه شده در مقاله به صورت مختصر توضیح دهید . . . . .
۱۱	۳.۲ مدل ارائه شده را پیاده سازی کرده و با استفاده از این دیتاست آموزش دهید. برای جلوگیری از بیش برازش، آموزش مدل را طوری تنظیم کنید که در انتهای آموزش، بهترین وزن های مدل بر اساس خطای قسمت اعتبارسنجی بازگردانده شود . . . . .
۱۴	۴.۲ ماتریس درهم ریختگی را روی قسمت آزمون داده ها رسم کنید و مقادیر Accuracy، Precision Recall، را گزارش کنید. فکر می کنید در مسائلی که توزیع برچسب ها نامتوازن است، استفاده از معیاری مانند به Accuracy تنهایی عمل کرد مدل را به درستی نمایش می دهد؟ چرا؟ اگر نه، کدام معیار می تواند به عنوان مکمل استفاده شود؟ . . . . .
۱۴	۵.۲ مدل را با استفاده از داده های نامتوازن و بدون حذف نویز، آموزش داده و موارد بخش قبلی را گزارش کنید و نتایج دو مدل را با هم مقایسه کنید. . . . .
۱۶	



## فهرست تصاویر

۵	نمودار هیت مپ دیتاست قبل از کاهش ابعاد	۱
۶	نمودار pairplot قبل از کاهش ابعاد	۲
۷	کاهش ابعاد به یک بعد با استفاده از TSNE	۳
۸	کاهش ابعاد به دو بعد با استفاده از TSNE	۴
۹	کاهش ابعاد به سه بعد با استفاده از TSNE	۵
۱۰	کاهش ابعاد به یک بعد با استفاده از LDA	۶
۱۱	کاهش ابعاد به دو بعد با استفاده از LDA	۷
۱۲	نواحی تصمیم برای SVM با هسته‌ی خطی	۸
۱۳	ماتریس درهم ریختگی برای SVM با هسته‌ی خطی	۹
۱۴	نواحی تصمیم گیری برای SVM با هسته‌ی چند جمله‌ای با درجه‌ی ۱	۱۰
۱۵	نواحی تصمیم گیری برای SVM با هسته‌ی چند جمله‌ای با درجه‌ی ۲	۱۱
۱۶	نواحی تصمیم گیری برای SVM با هسته‌ی چند جمله‌ای با درجه‌ی ۳	۱۲
۱۷	نواحی تصمیم گیری برای SVM با هسته‌ی چند جمله‌ای با درجه‌ی ۴	۱۳
۱۸	نواحی تصمیم گیری برای SVM با هسته‌ی چند جمله‌ای با درجه‌ی ۵	۱۴
۱۹	نواحی تصمیم گیری برای SVM با هسته‌ی چند جمله‌ای با درجه‌ی ۶	۱۵
۲۰	نواحی تصمیم گیری برای SVM با هسته‌ی چند جمله‌ای با درجه‌ی ۷	۱۶
۲۱	نواحی تصمیم گیری برای SVM با هسته‌ی چند جمله‌ای با درجه‌ی ۸	۱۷
۲۲	نواحی تصمیم گیری برای SVM با هسته‌ی چند جمله‌ای با درجه‌ی ۹	۱۸
۲۳	نواحی تصمیم گیری برای SVM با هسته‌ی چند جمله‌ای با درجه‌ی ۱۰	۱۹
۲۴	ماتریس درهم ریختگی برای حالتی که از اتوانکدر برای رفع نویز استفاده شده	۲۰
۲۵	ماتریس درهم ریختگی نرمالایز شده برای حالتی که از اتوانکدر برای رفع نویز استفاده شده	۲۱
۲۶	ماتریس درهم ریختگی برای حالتی که از اتوانکدر برای رفع نویز استفاده نشده	۲۲
۲۷	ماتریس درهم ریختگی نرمالایز شده برای حالتی که از اتوانکدر برای رفع نویز استفاده نشده	۲۳



## فهرست جداول

۷	svm linear for report classification	۱
۸	(degree۱) core polynomial with SVM for report classification	۲
۹	(degree۲) core polynomial with SVM for report classification	۳
۱۰	(degree۳) core polynomial with SVM for report classification	۴
۱۱	(degree۴) core polynomial with SVM for report classification	۵
۱۲	(degree۵) core polynomial with SVM for report classification	۶
۱۳	(degree۶) core polynomial with SVM for report classification	۷
۱۴	(degree۷) core polynomial with SVM for report classification	۸
۱۵	(degree۸) core polynomial with SVM for report classification	۹
۱۶	(degree۹) core polynomial with SVM for report classification	۱۰
۱۷	(degree۱۰) core polynomial with SVM for report classification	۱۱
۱۷	data denoised with classifier for report classification	۱۲
۱۸	denoising without classifier for report classification	۱۳



## فهرست برنامه‌ها

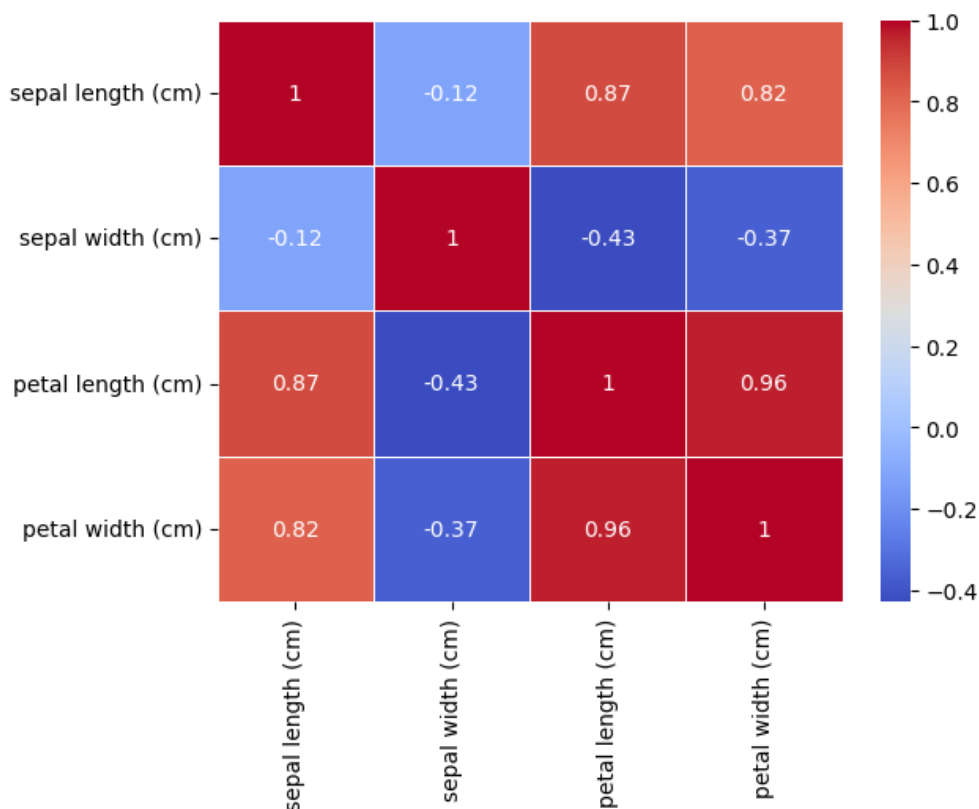


Notebook Colab  
github

## ۱ پرسش یک

۱.۱ در مرحله اول دیتاست را فراخوانی کنید و اطلاعاتی نظیر ابعاد، تعداد نمونه ها، میانگین، واریانس و همبستگی ویژگی ها را به دست آورید و نمونه های دیتاست را به تصویر بکشید سپس، با توجه به اطلاعات عددی، آماری و بصری بدست آمده، تحلیل کنید که آیا کاهش ابعاد می تواند در این دیتاست قابل استفاده باشد یا خیر

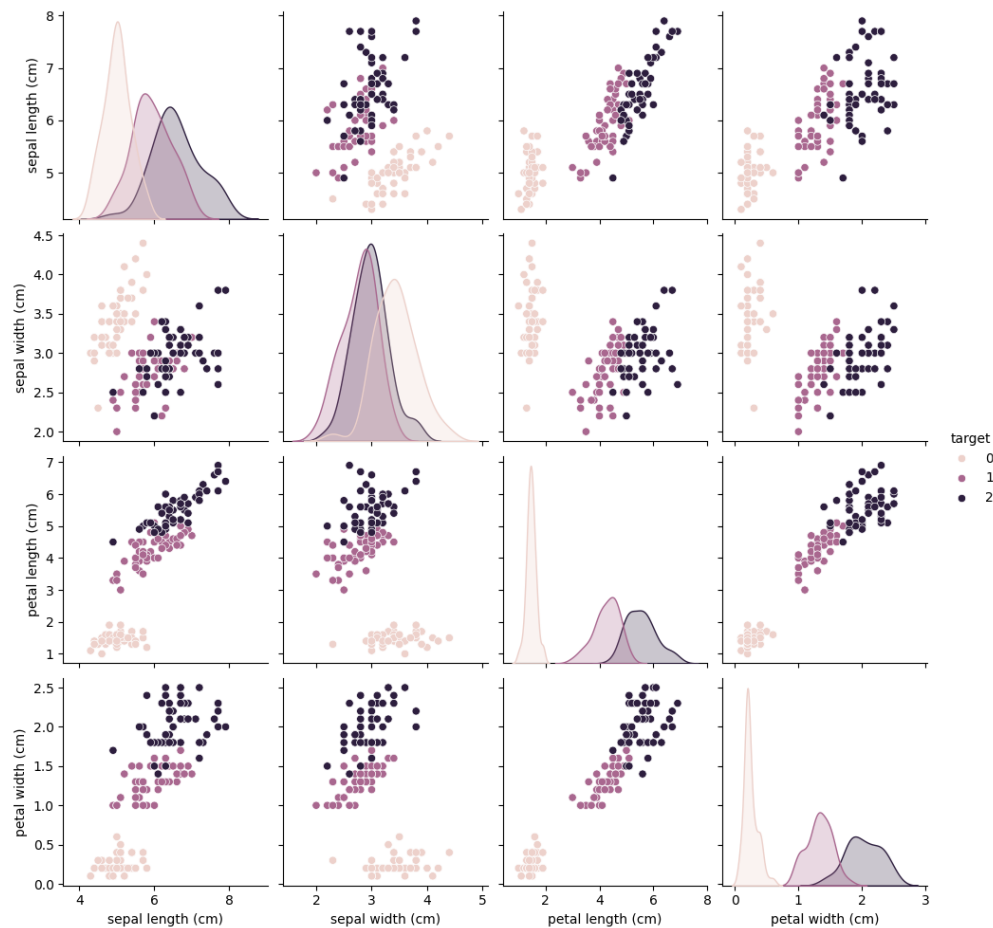
شکل ۱ نمودار هیت مپ همبستگی ویژگی ها را نمایش می دهد. همان طول که مشخص است ویژگی های sepal length، petal length و petal width همبستگی بسیار بالایی دارند و این به این معنی است که این ویژگی ها عملاً یک چیز را بیان می کنند.



شکل ۱: نمودار هیت مپ دیتاست قبل از کاهش ابعاد

کاهش ابعاد با استفاده از TSNE:

شکل های ۳، ۴ و ۵ به ترتیب کاهش ابعاد به یک، دو و سه بعد را نمایش می دهند. همان طور که قابل مشاهده است سه کلاس پس از کاهش بعد به حتی یک بعد به خوبی قابل جدا سازی اند.



شکل ۲: نمودار pairplot قبل از کاهش ابعاد

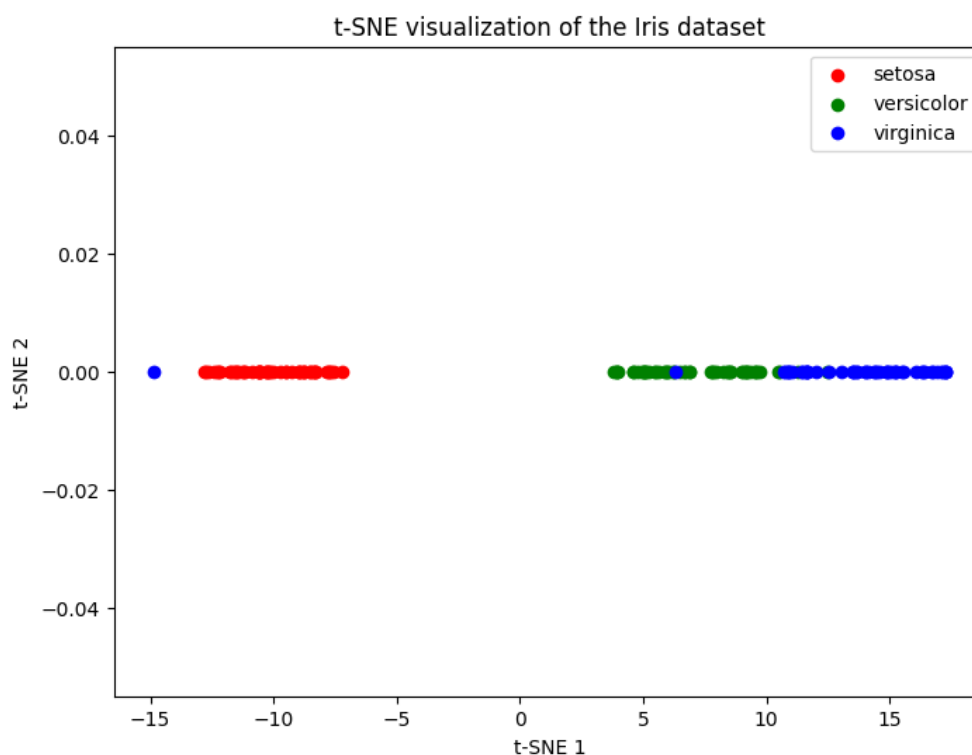
کاهش ابعاد با استفاده از: LDA

شکل‌های ۶ و ۷ به ترتیب کاهش ابعاد به یک و دو بعد را نمایش می‌دهند. همان طور که قابل مشاهده است در این حالت نیز سه کلاس پس از کاهش بعد به حتی یک بعد به خوبی قابل جدا سازی اند.

۲.۱ داده‌ها را با استفاده از الگوریتم SVM با هسته‌ی خطی دسته‌بندی کنید و ماتریس درهم ریختگی آن را بدست آورید و مرزهای تصمیم‌گیری را ترسیم کنید

از آنجایی که قرار است دسته بندی داشته باشیم و آموزش با ناظر می‌باشد LDA برای کاهش ابعاد مناسب تر است از این رو برای کاهش ابعاد از LDA استفاده شده است.

شکل ۸ نواحی تصمیم را برای SVM با هسته‌ی خطی نمایش می‌دهد و جدول ۱ گزارش دسته بندی است. همان طور که قابل مشاهده است خطایی در جداسازی نداریم.



شکل ۳: کاهش ابعاد به یک بعد با استفاده از TSNE

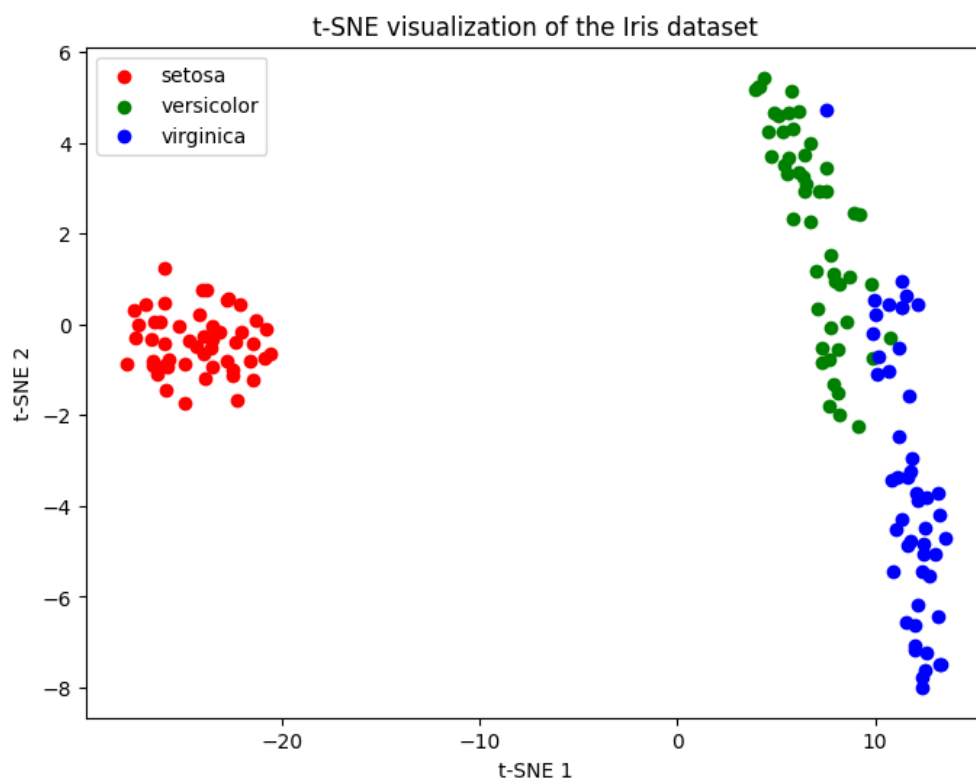
جدول ۱: svm linear for report classification

	precision	recall	f1-score	support
class 0	1.00	1.00	1.00	14
class 1	1.00	1.00	1.00	9
class 2	1.00	1.00	1.00	7
accuracy			1.00	30
macro avg	1.00	1.00	1.00	30
weighted avg	1.00	1.00	1.00	30

۳.۱ بخش قبلی را با استفاده از هسته های چند جمله ای و با استفاده از کتابخانهٔ scikit-learn درجه یک تا ۱۰ پیاده سازی کنید و نتایج را با معیارهای مناسب گزارش کرده و مقایسه و تحلیل کنید. در نهایت، با استفاده از کتابخانهٔ imageio ...

شکل های ۱۰ تا ۱۹ نواحی تصمیم گیری را برای SVM با هسته های چند جمله ای با درجه ای یک تا ده نمایش می دهند. همان طور که از تصاویر مشخص است نواحی تصمیم گیری برای درجات زوج اصلاً مطلوب نیست. همچنین برای درجات فرد نیز درجات پایین تر خروجی مطلوب تری دارند. از این تصاویر نتیجه می گیریم که پیچیدگی همیشه مطلوب نیست. جدول های ۲ تا ۱۱ گزارش دسته بندی را برای SVM با هسته های چند جمله ای با درجه ای یک تا ده را نمایش می دهند.



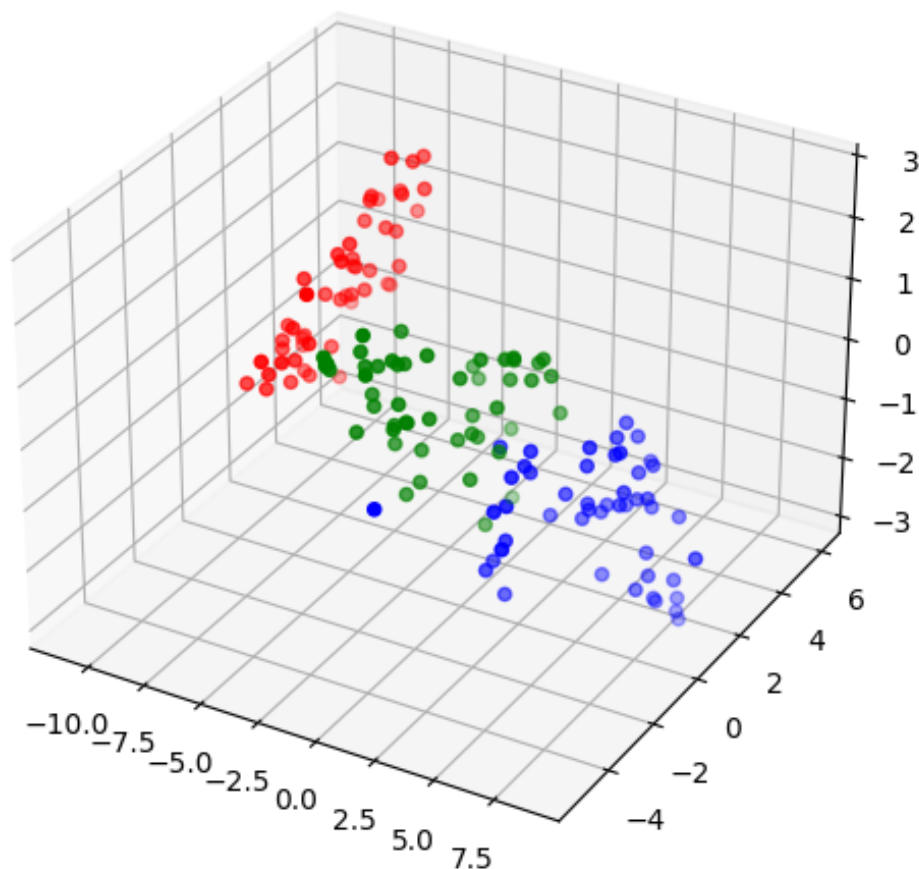


شکل ۴: کاهش ابعاد به دو بعد با استفاده از TSNE

جدول ۲: core polynomial with SVM for report classification (degree\)

	precision	recall	f1-score	support
class 0	1.00	1.00	1.00	14
class 1	1.00	1.00	1.00	9
class 2	1.00	1.00	1.00	7
accuracy			1.00	30
macro avg	1.00	1.00	1.00	30
weighted avg	1.00	1.00	1.00	30

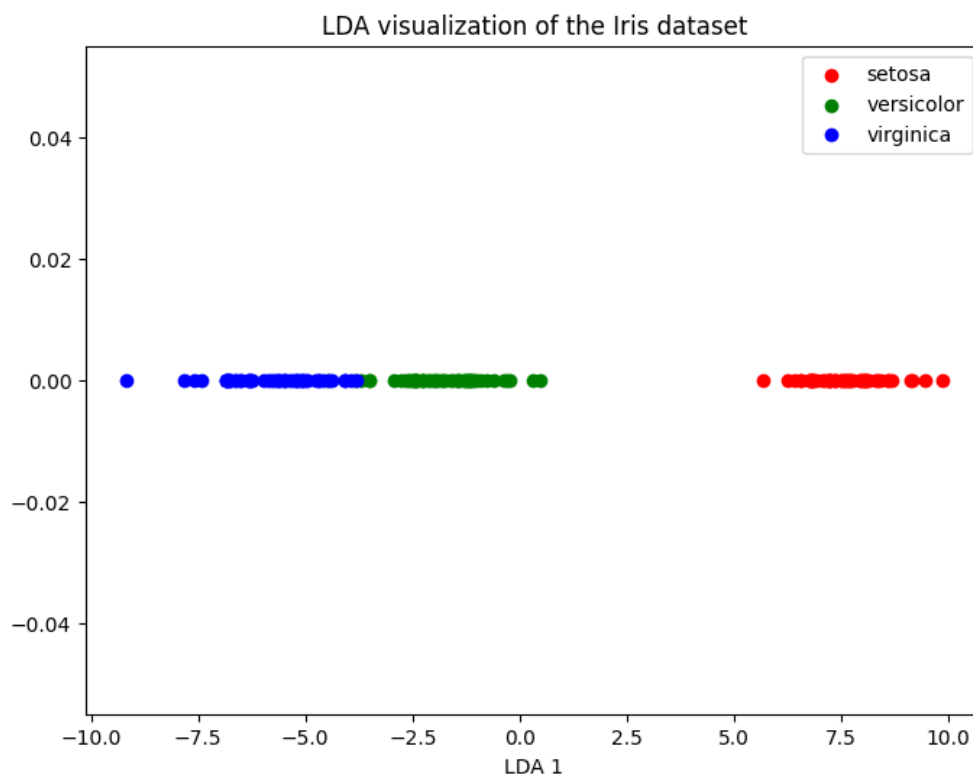
Decision boundary gif



شکل ۵: کاهش ابعاد به سه بعد با استفاده از TSNE

جدول ۳: core polynomial with SVM for report classification (degree ۲)

	precision	recall	f1-score	support
class 0	1.00	0.79	0.88	14
class 1	1.00	1.00	1.00	9
class 2	0.70	1.00	0.82	7
accuracy			0.90	30
macro avg	0.90	0.93	0.90	30
weighted avg	0.93	0.90	0.90	30



شکل ۶: کاهش ابعاد به یک بعد با استفاده از LDA

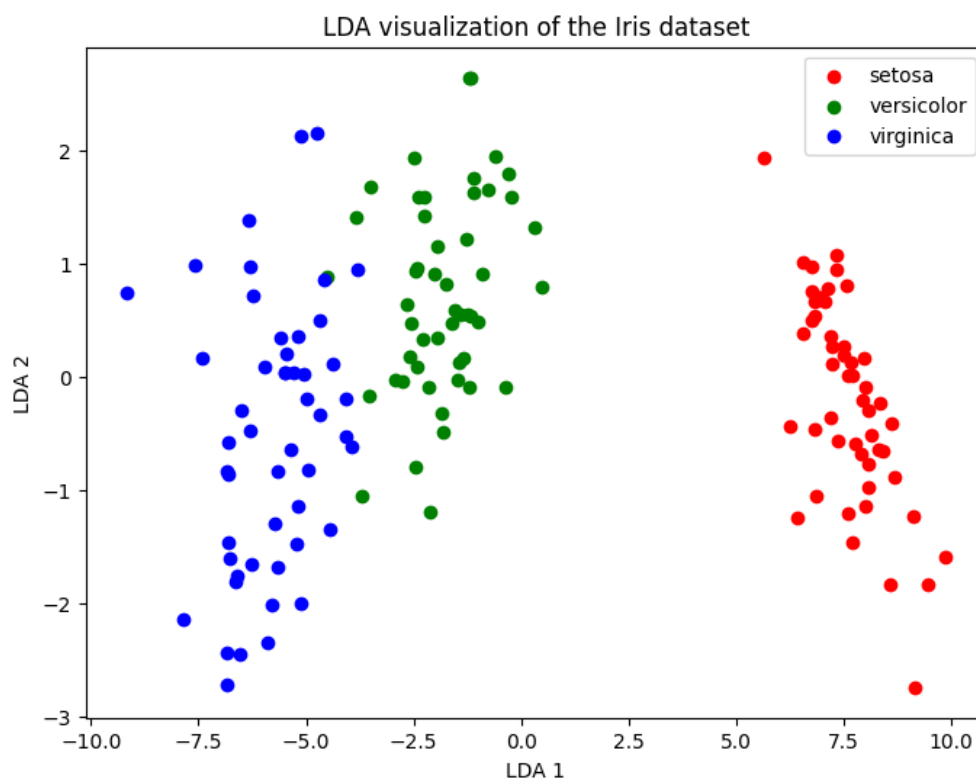
جدول ۴: core polynomial with SVM for report classification (degree ۳)

	precision	recall	f1-score	support
class 0	1.00	1.00	1.00	14
class 1	1.00	1.00	1.00	9
class 2	1.00	1.00	1.00	7
accuracy			1.00	30
macro avg	1.00	1.00	1.00	30
weighted avg	1.00	1.00	1.00	30

## ۲ پرسش سه

۱.۲ بزرگ ترین چالش ها در توسعه مدل های تشخیص تقلب چیست؟ این مقاله برای حل این چالش ها از چه روش هایی استفاده کرده است؟

- نامتوازن بودن داده: برای حل این مشکل داده ها را با روش oversample SMOTE می کنند
- نویزی شدن داده ها در اثر oversample: برای حل این مشکل از یک اتوانکدر استفاده شده است.
- نادر بودن داده های تقلب: از آنجایی که تشخیص داده های مرتبط با تقلب دشوار است بنابر این تعداد نمونه های کمی از این دسته



شکل ۷: کاهش ابعاد به دو بعد با استفاده از LDA

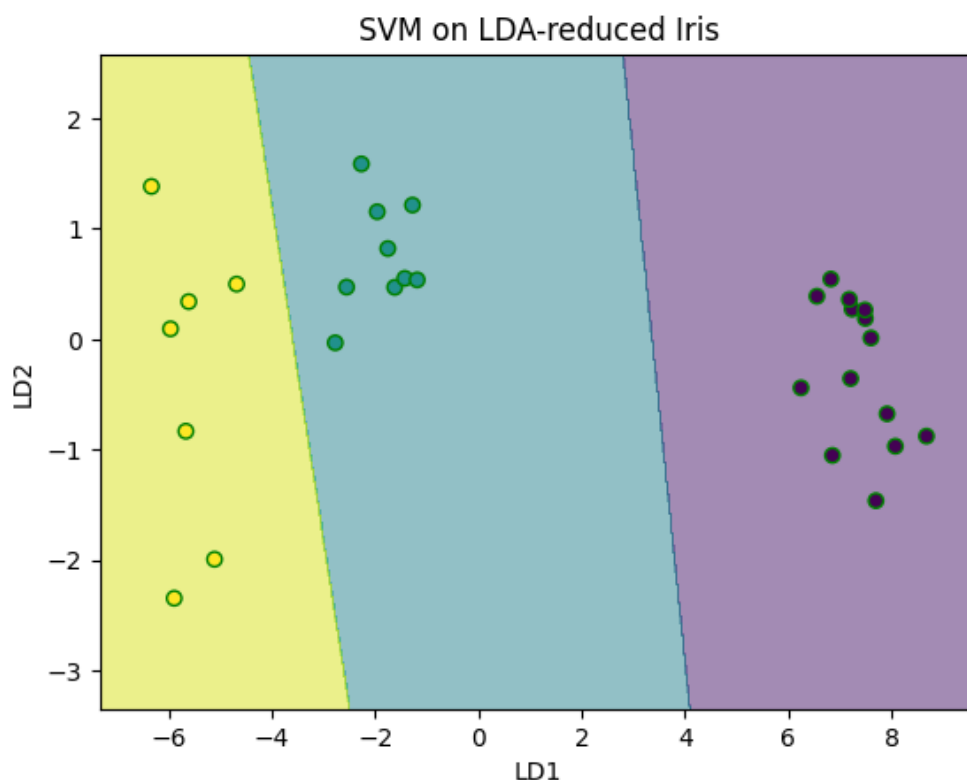
جدول ۵: core polynomial with SVM for report classification (degree ۴)

	precision	recall	f1-score	support
class 0	1.00	0.79	0.88	14
class 1	1.00	1.00	1.00	9
class 2	0.70	1.00	0.82	7
accuracy			0.90	30
macro avg	0.90	0.93	0.90	30
weighted avg	0.93	0.90	0.90	30

وجود دارد بنابراین در این مقاله با oversampling سعی در حل این مشکل شده است ولی این کار نیز تا حدی می تواند موثر باشد چون در واقع با این کار داده‌ی غیر واقعی تولید کرده‌ایم.

## ۲.۲ در مورد معماری شبکه‌ی ارائه شده در مقاله به صورت مختصر توضیح دهید

معماری اتوانکدر: انکدر شامل چهار لایه تمام متصل است که مرحله به مرحله ابعاد را کاهش می‌دهد تا به ۱۰ بعد برسد. دیکدر نیز شامل ۴ لایه تمام متصل است که ابعاد را مرحله به مرحله افزایش می‌دهد تا به ۲۹ بعد برسد. نکته‌ای که وجود دارد این است که با این که در مقاله در انتهای لایه‌ها تابع فعالساز استفاده نشده در این جا ما برای بهبود عملکرد و آموزش شبکه در انتهای لایه‌ها از فعال ساز ReLU



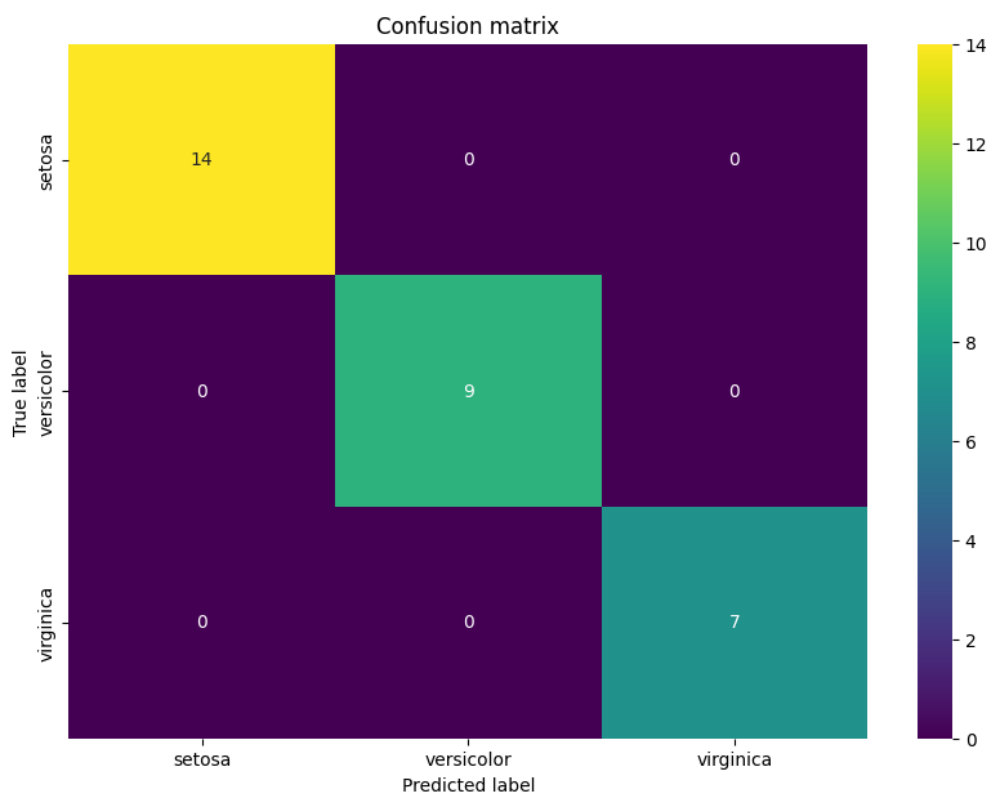
شکل ۸: نواحی تصمیم برای SVM با هسته‌ی خطی

جدول ۶: core polynomial with SVM for report classification (degree ۵)

	precision	recall	f1-score	support
class 0	1.00	1.00	1.00	14
class 1	1.00	1.00	1.00	9
class 2	1.00	1.00	1.00	7
accuracy			1.00	30
macro avg	1.00	1.00	1.00	30
weighted avg	1.00	1.00	1.00	30

استفاده شده است.

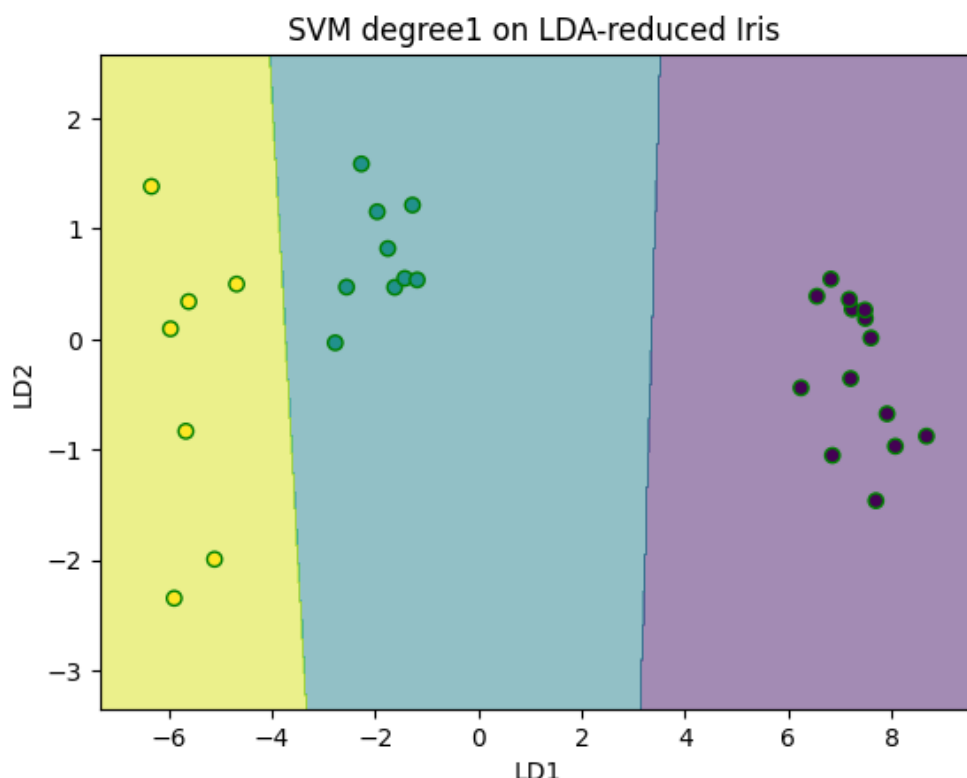
معماری طبقه بند: طبقه بند شامل پنج لایه‌ی تمام متصل است. با این که در مقاله ذکر نشده در اینجا برای بهبود شبکه در انتهای لایه‌ها (به جز لایه‌ی آخر) از فعالساز ReLU استفاده شده است. فعالساز لایه‌ی آخر Softmax است.



شکل ۹: ماتریس درهم ریختگی برای SVM با هسته ی خطی

جدول ۷: core polynomial with SVM for report classification (degree ۶)

	precision	recall	f1-score	support
class 0	1.00	0.71	0.83	14
class 1	1.00	1.00	1.00	9
class 2	0.64	1.00	0.78	7
accuracy			0.87	30
macro avg	0.88	0.90	0.87	30
weighted avg	0.92	0.87	0.87	30



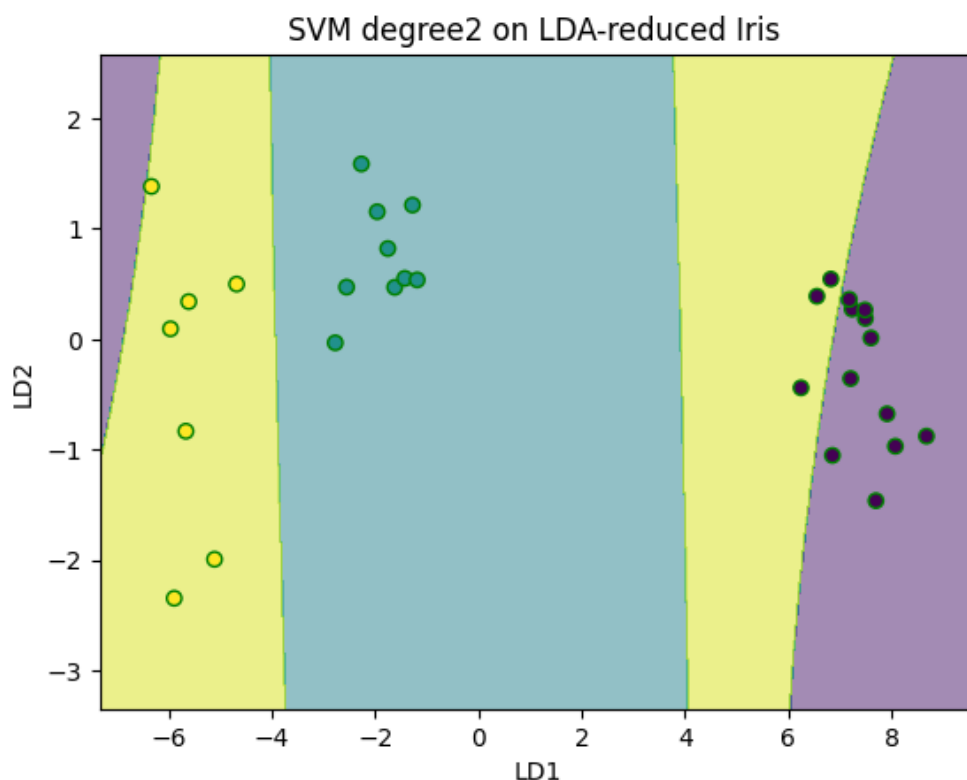
شکل ۱۰: نواحی تصمیم گیری برای SVM با هسته‌ای چند جمله‌ای با درجه‌ی ۱

جدول ۸: core polynomial with SVM for report classification (degree=۷)

	precision	recall	f1-score	support
class 0	1.00	1.00	1.00	14
class 1	0.90	1.00	0.95	9
class 2	1.00	0.86	0.92	7
accuracy			0.97	30
macro avg	0.97	0.95	0.96	30
weighted avg	0.97	0.97	0.97	30

۳.۲ مدل ارائه شده را پیاده سازی کرده و با استفاده از این دیتاست آموزش دهید. برای جلوگیری از بیش برآزش، آموزش مدل را طوری تنظیم کنید که در انتهای آموزش، بهترین وزن های مدل بر اساس خطای قسمت اعتبارسنجی بازگردانده شود

۴.۲ ماتریس درهم ریختگی را روی قسمت آزمون داده ها رسم کنید و مقادیر Precision Recall، Accuracy، و f1score را گزارش کنید. فکر می کنید در مسائلی که توزیع برچسب ها نامتوازن است، استفاده از معیاری مانند به Accuracy عمل کرد مدل را به درستی نمایش می دهد؟ چرا؟ اگر نه، کدام معیار می تواند به عنوان مکمل استفاده شود؟



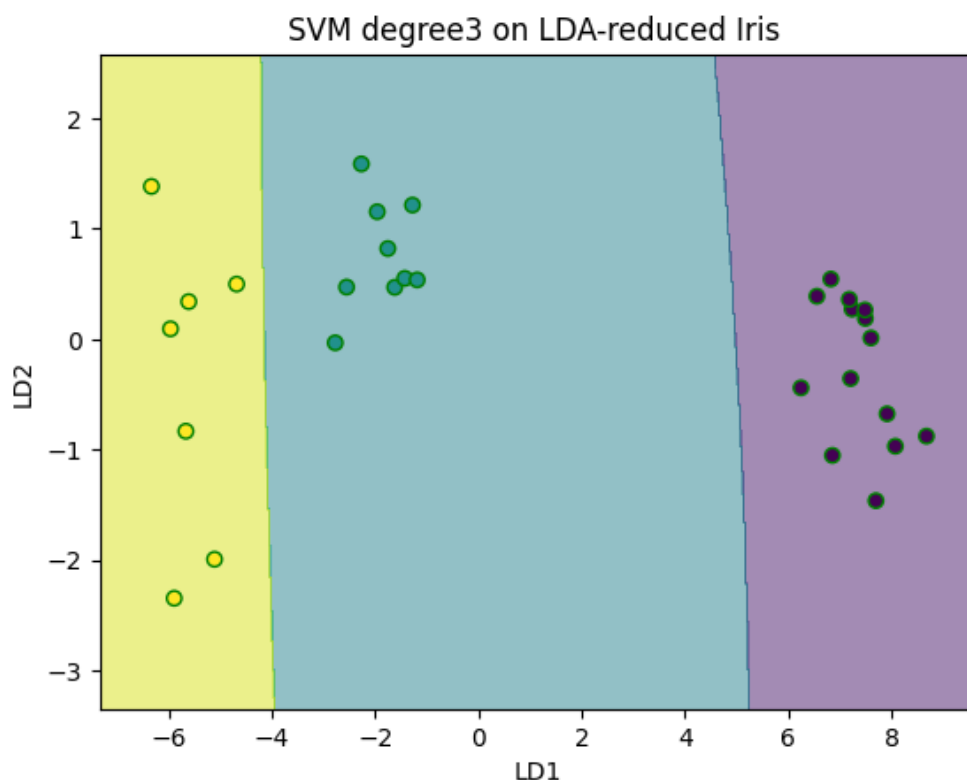
شکل ۱۱: نواحی تصمیم گیری برای SVM با هسته‌ی چند جمله‌ای با درجه‌ی ۲

جدول ۹: core polynomial with SVM for report classification (degree=۸)

	precision	recall	f1-score	support
class 0	1.00	0.71	0.83	14
class 1	0.90	1.00	0.95	9
class 2	0.60	0.86	0.71	7
accuracy			0.83	30
macro avg	0.83	0.86	0.83	30
weighted avg	0.88	0.83	0.84	30

در صورتی که داده‌ها نا متقارن باشند Accuracy نمی‌تواند معیار درستی برای ارزیابی باشد، دلیل این را با یک مثال بیان می‌کنیم. فرض کنید ۱۰۰۰۰۰ داده داریم که ۱۰۰ عدد از آن‌ها متعلق به کلاس ۱ و ۹۹۹۰۰ متعلق به کلاس ۲ باشند. در این صورت اگر حتی بدون به کارگیری یک طبقه بند خطی صرفاً کل داد‌ها را جزو کلاس ۲ در نظر بگیریم Accuracy برابر با ۹۹.۹۹ درصد می‌شود در حالی که هیچکدام از داد‌های کلاس ۱ را تشخیص نداده ایم و این اصلاً مطلوب نیست. به خصوص در کاربردهایی مانند تشخیص بیماری که تعداد نمونه‌های بیمار خیلی کم‌تر از نمونه‌های سالم است این معیار اصلاً مناسب نیست.





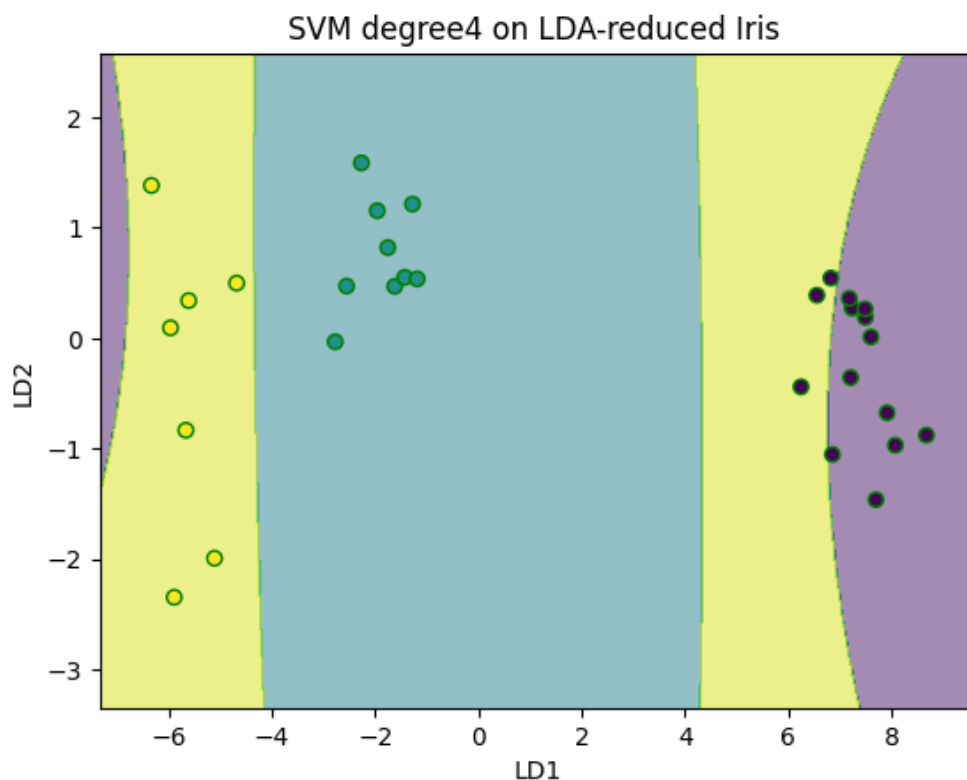
شکل ۱۲: نواحی تصمیم گیری برای SVM با هسته‌ی چند جمله‌ای با درجه‌ی ۳

جدول ۱۰: core polynomial with SVM for report classification (degree ۹)

	precision	recall	f1-score	support
class 0	1.00	1.00	1.00	14
class 1	0.90	1.00	0.95	9
class 2	0.92	0.86	1.00	7
accuracy			0.97	30
macro avg	0.97	0.95	0.96	30
weighted avg	0.97	0.97	0.97	30

۵.۲ مدل را با استفاده از داده‌های نامتوازن و بدون حذف نویز، آموزش داده و موارد بخش قبلی را گزارش کنید و نتایج مدل را با هم مقایسه کنید.

شکل‌های ۲۲ و ۲۳ به ترتیب ماتریس درهم‌ریختگی و ماتریس درهم‌ریختگی نرمالیزه شده را برای حالتی که داده‌ها denoise نشده‌اند نمایش می‌دهند، همچنین جدول ۱۳ معیارهایی مانند f1score و recall را نمایش می‌دهد. اگر جدول و ماتریس درهم‌ریختگی را با حالت قبلی مقایسه کنیم متوجه می‌شویم که در حالتی که داده‌ها denoise شده بودند خروجی بهتری داشتیم.



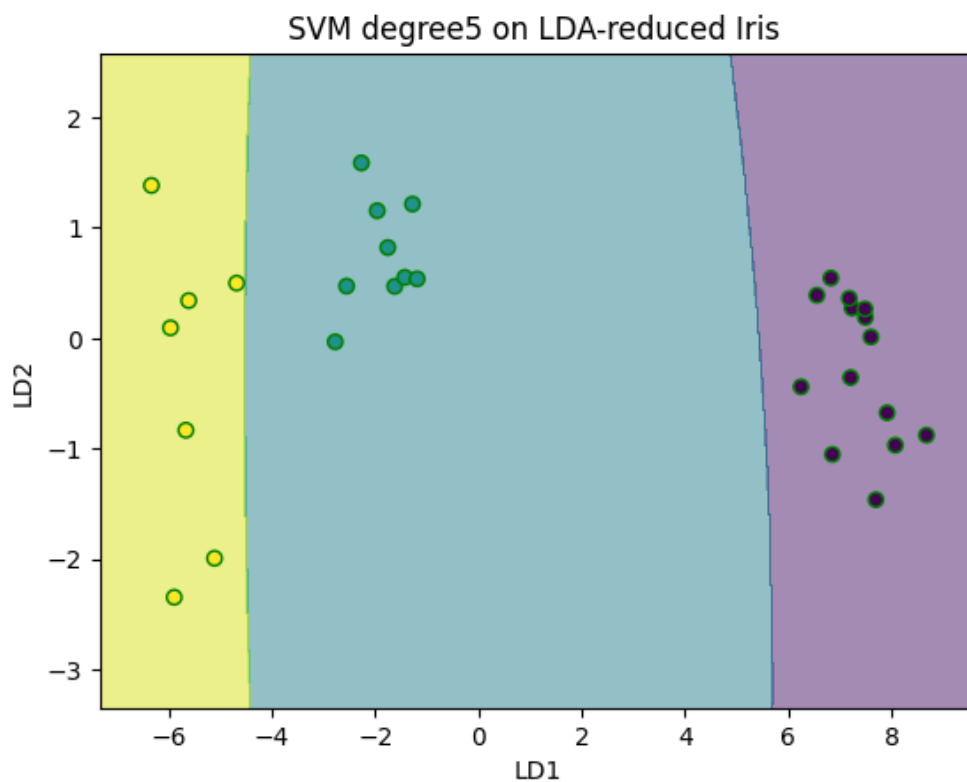
شکل ۱۳: نواحی تصمیم گیری برای SVM با هسته‌ی چند جمله‌ای با درجه‌ی ۴

جدول ۱۱: core polynomial with SVM for report classification (degree ۱۰)

	precision	recall	f1-score	support
class 0	1.00	0.64	0.78	14
class 1	0.90	1.00	0.95	9
class 2	0.55	0.86	0.67	7
accuracy			0.80	30
macro avg	0.82	0.83	0.80	30
weighted avg	0.86	0.80	0.80	30

جدول ۱۲: data denoised with classifier for report classification

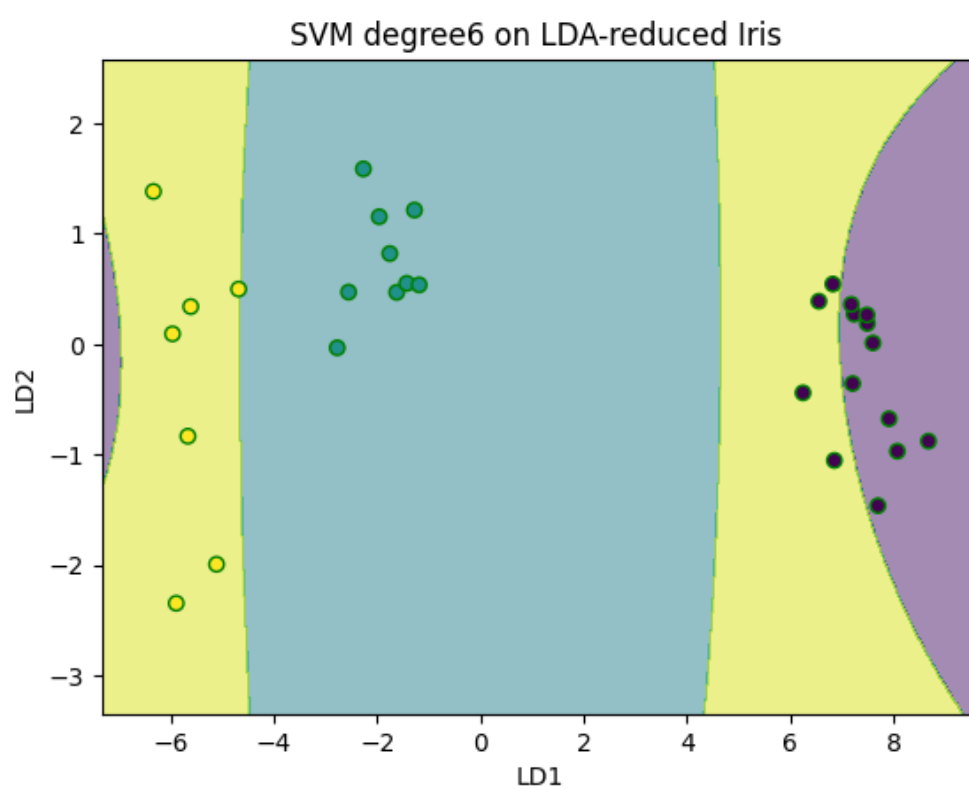
	precision	recall	f1-score	support
class 0	1.00	1.00	1.00	56887
class 1	0.63	0.79	0.70	84
accuracy			1.00	56961
macro avg	0.81	0.89	0.85	56961
weighted avg	1.00	1.00	1.00	56961



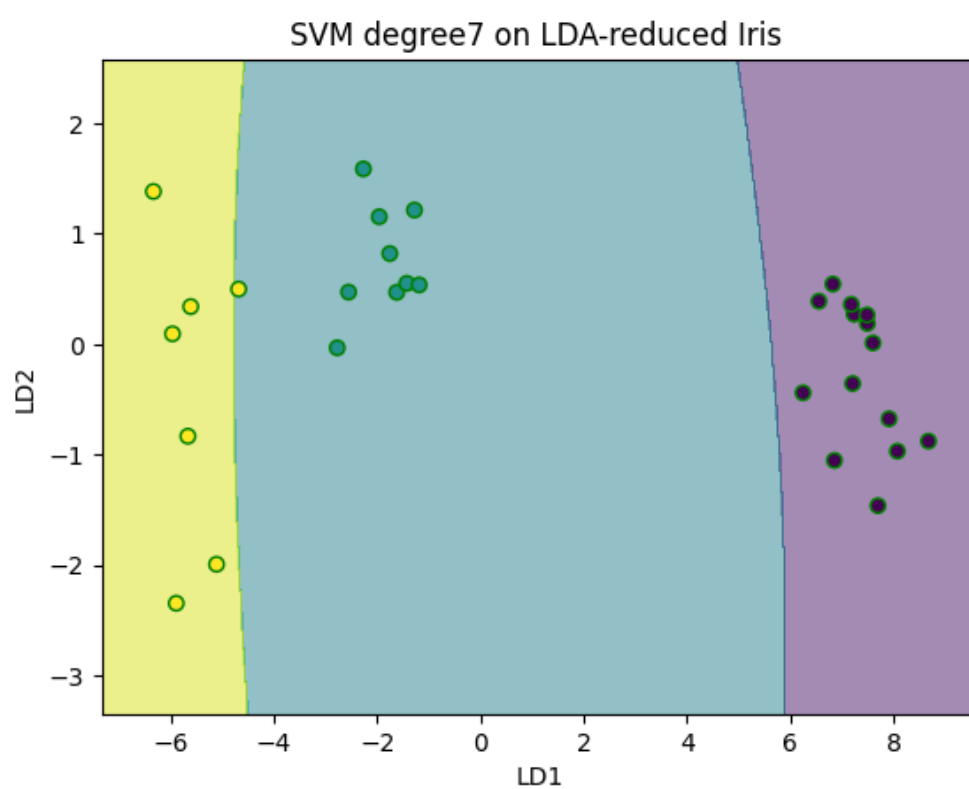
شکل ۱۴: نواحی تصمیم گیری برای SVM با هسته‌ی چند جمله‌ای با درجه‌ی ۵

جدول ۱۳: denoising without classifier for report classification

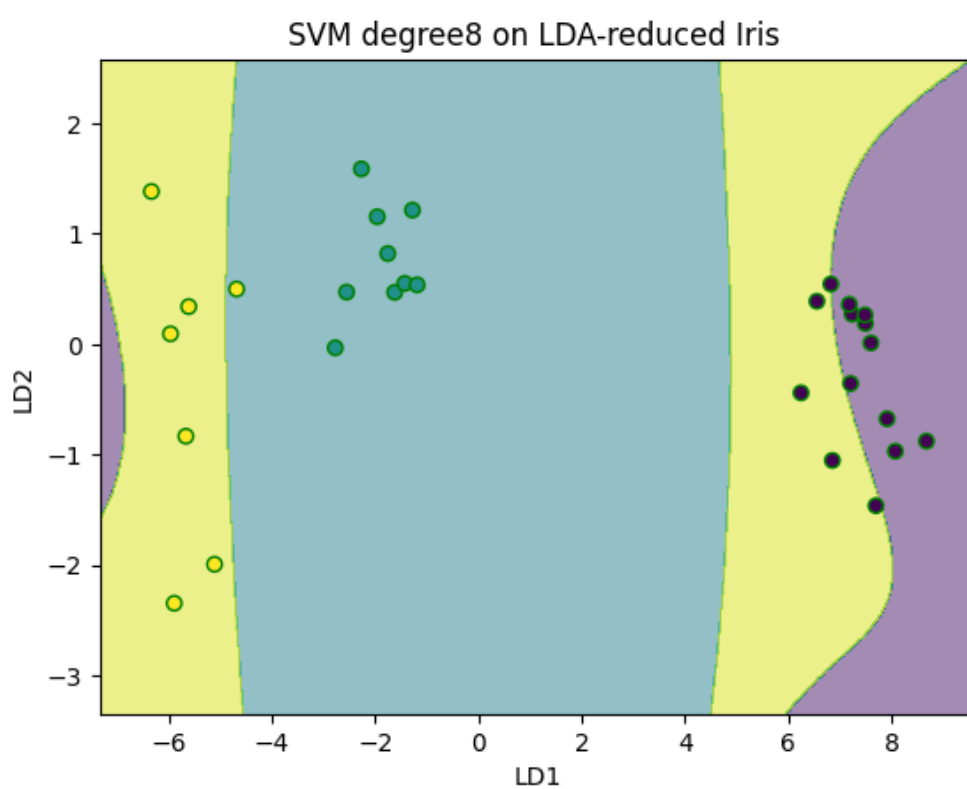
	precision	recall	f1-score	support
class 0	1.00	1.00	1.00	14
class 1	0.49	0.79	0.60	9
class 2	0.55	0.86	0.67	7
accuracy			1.00	30
macro avg	0.74	0.89	0.80	30
weighted avg	1.00	1.00	1.00	30



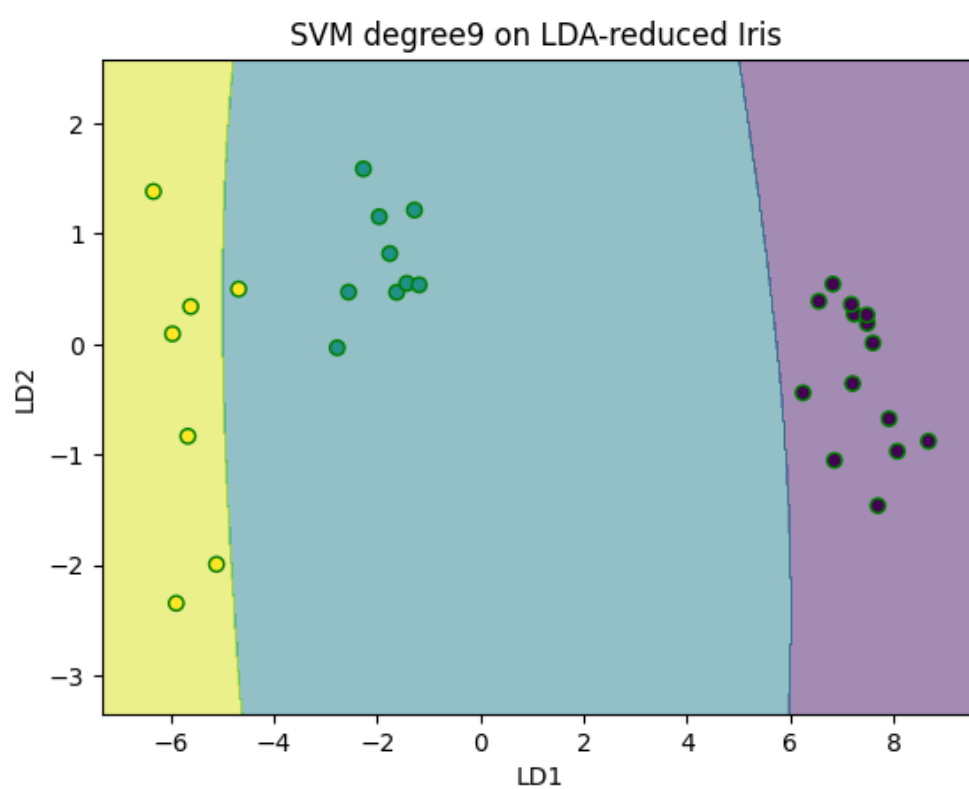
شکل ۱۵: نواحی تصمیم گیری برای SVM با هسته‌ی چند جمله‌ای با درجه‌ی ۶



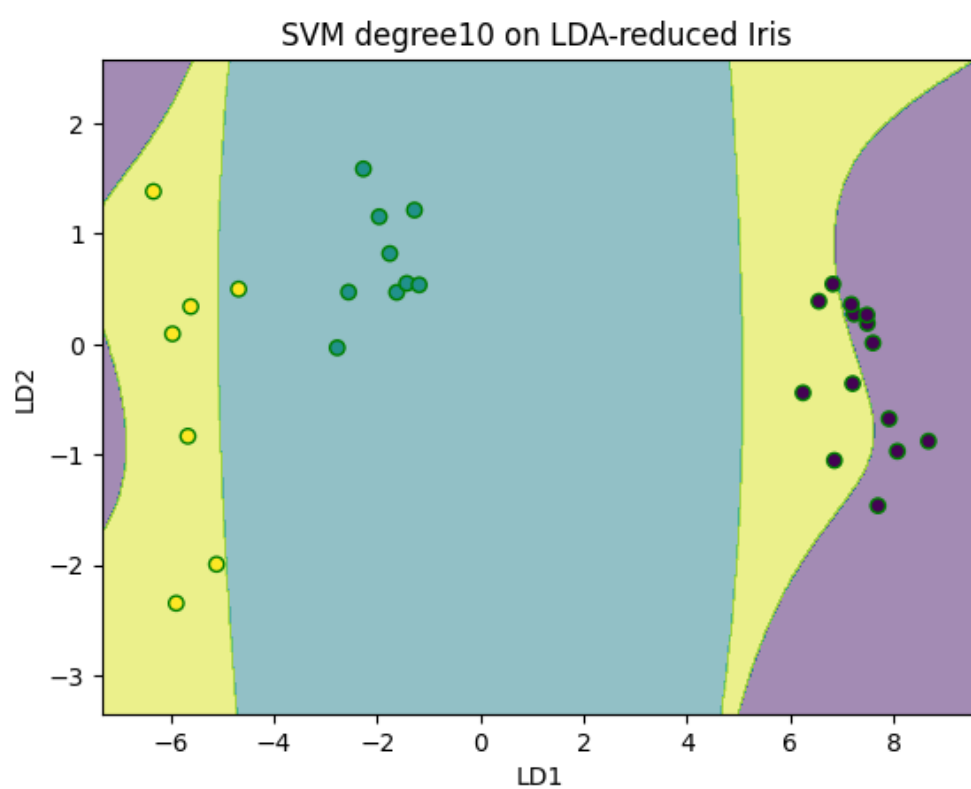
شکل ۱۶: نواحی تصمیم گیری برای SVM با هسته‌ی چند جمله‌ای با درجه‌ی ۷



شکل ۱۷: نواحی تصمیم گیری برای SVM با هسته‌ی چند جمله‌ای با درجه‌ی ۸

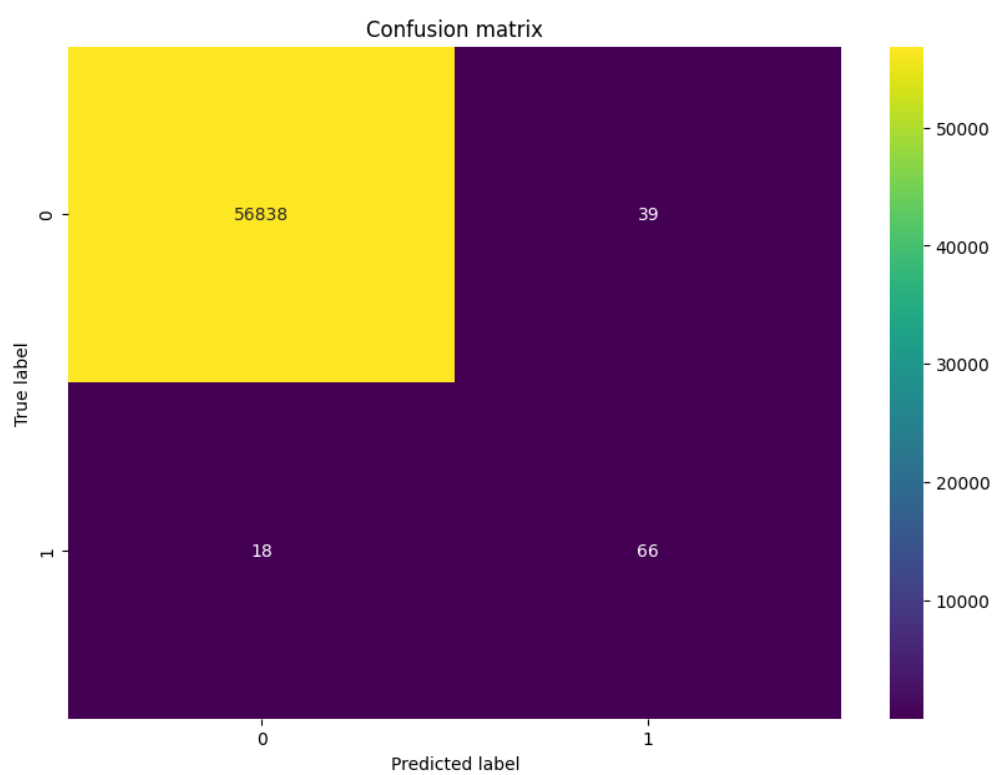


شکل ۱۸: نواحی تصمیم گیری برای SVM با هسته‌ی چند جمله‌ای با درجه‌ی ۹

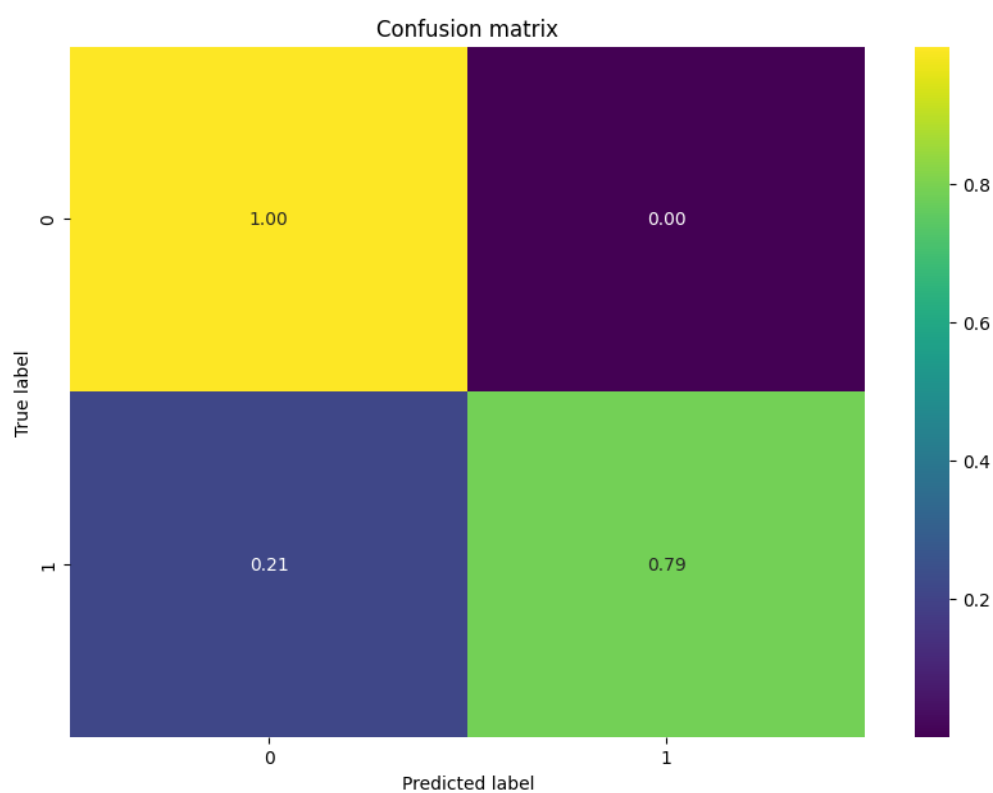


شکل ۱۹: نواحی تصمیم گیری برای SVM با هسته‌ی چند جمله‌ای با درجه‌ی ۱۰

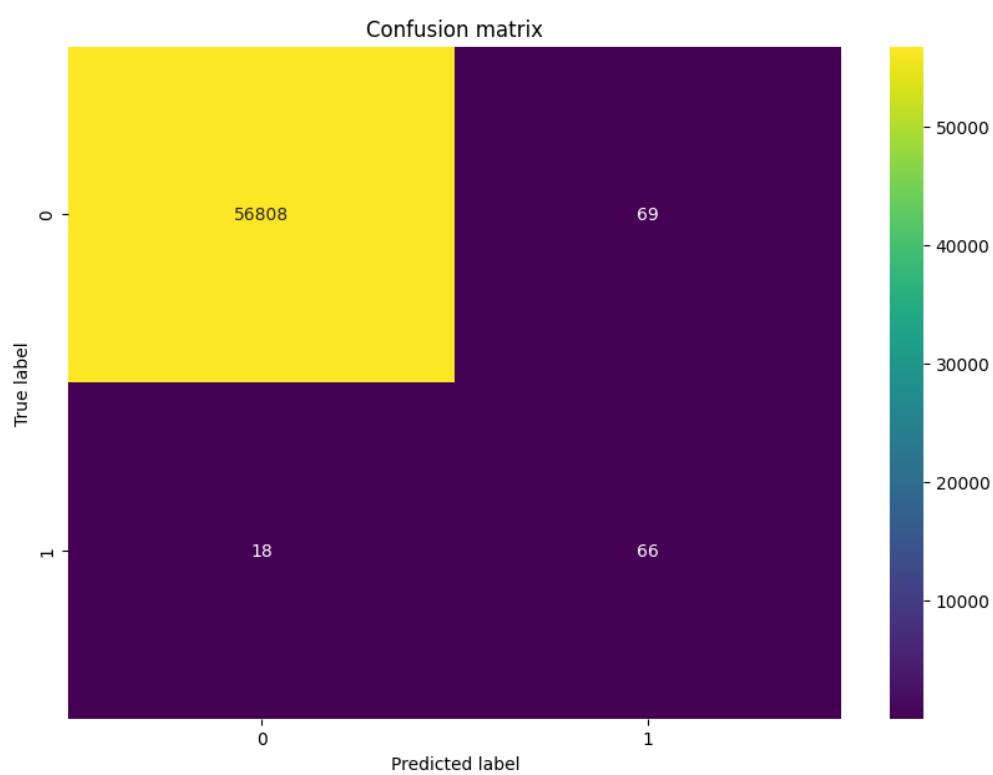




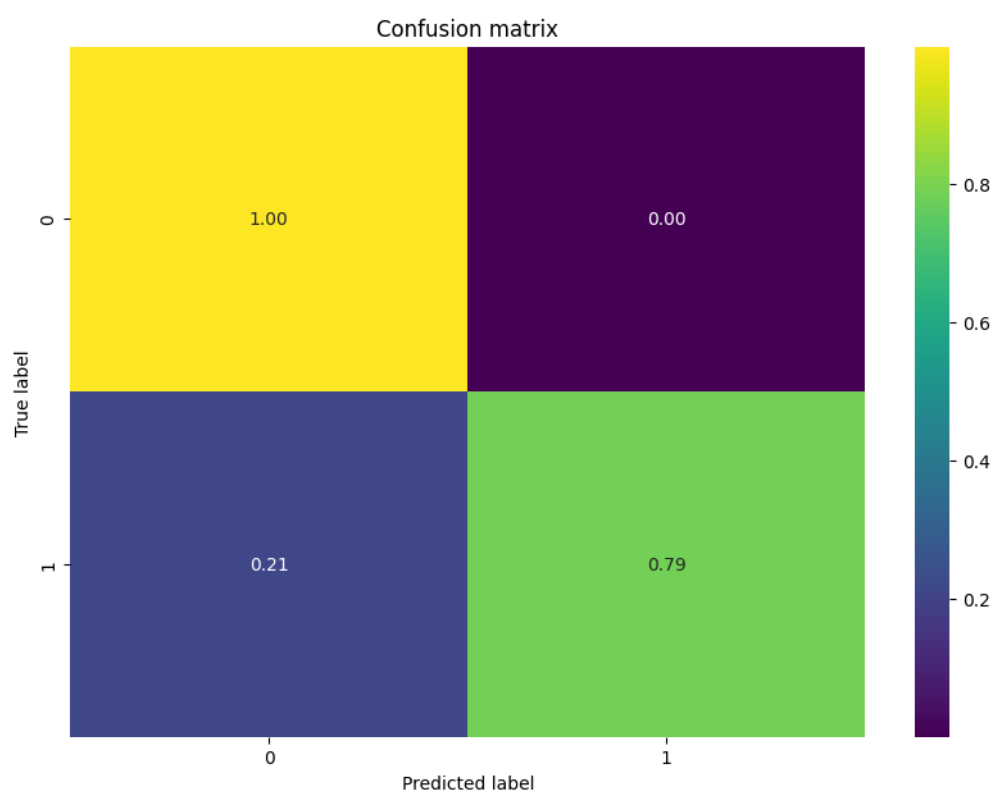
شکل ۲۰: ماتریس درهم‌ریختگی برای حالتی که از اتوانکدر برای رفع نویز استفاده شده



شکل ۲۱: ماتریس درهم‌ریختگی نرمالایز شده برای حالتی که از اتوانکدر برای رفع نویز استفاده شده



شکل ۲۲: ماتریس درهمریختگی برای حالتی که از اتوانکدر برای رفع نویز استفاده نشده



شکل ۲۳: ماتریس درهمریختگی نرمالایز شده برای حالتی که از اتوانکدر برای رفع نویز استفاده نشده