

به نام خدا



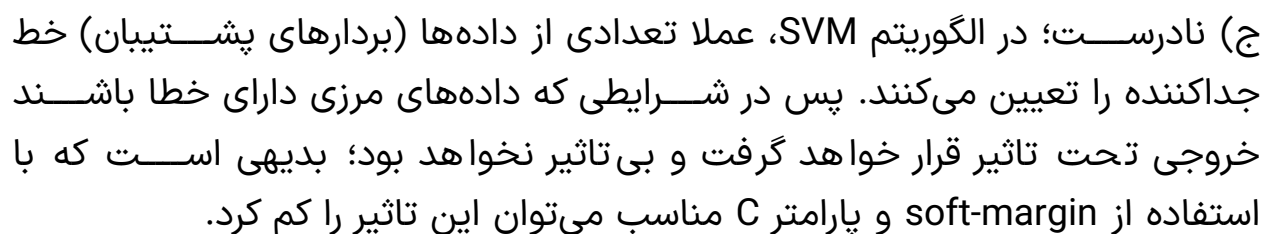
دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

درس یادگیری ماشین
استاد ناظر فرد

تمرین سوم

علیرضا مازوچی

۴۰۰۱۳۱۰۷۵



د) نادرست؛ خطای صفر برای بعضی از مجموعه‌های آموزشی و با هر الگوریتمی (هرچند بیش‌برازش شده!) امکان‌پذیر نیست. چراکه ممکن است دو داده با کلاس متفاوت به دلیل وجود خطا یا سایر دلایل دقیقاً در یک نقطه قرار گرفته باشند. طبیعتاً دسته‌بند برای آن نقطه اگر هر تصمیمی بگیرد، خطا صفر نخواهد شد!

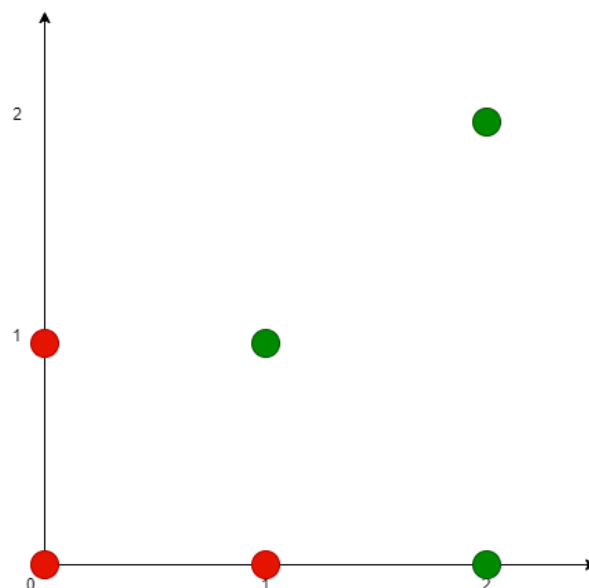
ه) نادرست؛ امکان تخصیص وزن منفی به یک دسته‌بند وجود دارد. باتوجه به آنکه وزن یک دسته‌بند از رابطه زیر حاصل می‌شود، اگر خطای دسته‌بندی بالای ۵۰٪ باشد، این دسته‌بند وزن منفی می‌گیرد:

$$\alpha^t = \frac{1}{2} \ln \left(\frac{1 - \epsilon^t}{\epsilon^t} \right)$$

و) درست؛ در الگوریتم آدابوست داده‌ای که در یک گام به اشتباه پیش‌بینی شود، وزنی بیش از پیش می‌گیرد؛ در نتیجه اگر داده نویزی در محل کلاس اشتباهی قرار بگیرد، وزن زیادی را در طی گام‌های متوالی خواهد گرفت و عملکرد الگوریتم را دچار مشکل می‌کند.

سوال ۲

الف) بله به وضوح جداپذیرند.



ب) باتوجه به تصویر قسمت الف، چهار نقطه $(0,1)$ ، $(1,0)$ ، $(1,1)$ و $(2,0)$ بردارهای پشتیبان هستند و داریم:

$$\begin{cases} w \cdot x_+ + b = 1 \\ w \cdot x_- + b = 1 \end{cases} \rightarrow \begin{cases} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} + b = 1 \\ \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 0 \end{bmatrix} + b = 1 \\ \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} + b = -1 \\ \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix} + b = -1 \end{cases} \rightarrow \begin{cases} w_1 + w_2 + b = 1 \\ 2w_1 + b = 1 \\ w_1 + b = -1 \\ w_2 + b = -1 \end{cases}$$

$$\rightarrow b = -3, w_1 = 2, w_2 = 2$$

معادله خط جداکننده:

$$\begin{bmatrix} 2 \\ 2 \end{bmatrix} \cdot x - 3 = 0$$

ج) خیر؛ اندازه حاشیه ثابت می‌ماند. چراکه دو بردار پشتیبان برای کلاس مثبت و دو بردار پشتیبان برای کلاس منفی وجود دارد که اگر یکی از داده‌ها حذف شود، یک کلاس دو داده‌اش باقی می‌ماند که به واسطه آن و تک داده باقی‌مانده همچنان خط جداکننده بدون تغییر می‌ماند.

د) این گزاره صحیح است. فرض کنید این گزاره غلط باشد؛ یعنی مثالی وجود داشته باشد که در آن با حذف یکی از بردارهای پشتیبان اندازه حاشیه کاهش پیدا کند. حالتی که این بردار پشتیبان حذف شده است را در نظر بگیرید. در این حالت اگر بردار پشتیبان را اضافه کنیم، طبیعتاً باید اندازه حاشیه افزایش یابد ولی چنین چیزی امکان ندارد. چراکه داده‌های فعلی در بیشترین فاصله ممکن از هم قرار گرفته‌اند و وجود یک داده‌ی جدید نه تنها نمی‌تواند این حاشیه را بیشتر کند بلکه ممکن است خود در داخل این حاشیه قرار بگیرد و لازم شود تا حاشیه کوچکتری در نظر گرفته شود. پس به تناقض می‌خوریم و فرض اولیه اثبات می‌شود.

سوال ۳

الف) خیر. با دیدن تصویر زیر مشخص است.



ب)

x^2	$\sqrt{2}x$	1	x	class
0	0	1	0	+
1	$-\sqrt{2}$	1	-1	-
1	$\sqrt{2}$	1	+1	-

در این حالت می‌توان داده‌ها را به صورت خطی از هم جدا کرد. تنها بعدی که باعث جدا شدن داده‌های مثبت و منفی شده است بعد x^2 است. در این بعد مقدار ۰/۵ بهترین حاشیه را برای جداسازی فراهم می‌کند. پس $x^2 = \frac{1}{2}$ یک ابرصفحه جداکننده در فضای جدید خواهد بود. لذا می‌توان معادله زیر را برای خط جداکننده پیشنهاد کرد:

$$\begin{bmatrix} 0 \\ 0 \\ -2 \end{bmatrix} \cdot x + 1 = 0$$

سوال ۴

ردیف	تنظیمات	جواب
الف	SVM خطی، soft-margin، $C=0.1$	۴
ب	SVM خطی، soft-margin، $C=10$	۳
ج	SVM، hard-margin، $k(u, v) = u \cdot v + (u \cdot v)^2$	۲
د	SVM، hard-margin، $k(u, v) = \exp\left(-\frac{1}{4} \ u - v\ ^2\right)$	۱
ه	SVM، hard-margin، $k(u, v) = \exp\left(-4 \ u - v\ ^2\right)$	۶

ابتدا الف و ب را بررسی می‌کنیم. طبیعی است که جواب این دو تصویر ۳ و ۴ خواهد بود. در تصویر ۳ خط در جایی قرار دارد که داده‌ها را به طور کامل از هم جدا کند ولی در تصویر ۴، بیشینه‌کردن حاشیه نیز اهمیت پیدا کرده است به گونه‌ای که دو داده در حوالی مرز جداکننده قرار گرفته‌اند. باتوجه به اینکه C ابرپارامتری است که زیادبودن آن اهمیت خطای slack را بیشتر می‌کند پس تنظیم ب با C بیشتر مربوط به ۳ و تنظیم الف مربوط به ۴ است.

تنظیم ج دارای یک کرنل چندجمله‌ای است؛ در نتیجه نمی‌تواند خروجی که در تصاویر ۱ و ۵ و ۶ نشان داده شده است را داشته باشد. پس تصویر ۲ بهترین تطبیق را با آن خواهد داشت.

نهایتاً نوبت به دو تنظیم د و ه می‌رسد. این دو تنظیم مانند یکدیگر هستند با این تفاوت که تنظیم ه دارای مقدار گامای بیشتری است. در میان تصاویر هم تنها تصویر ۱ و ۶ از تعدادی توزیع گاوسین تشکیل شده است. وقتی مقدار گاما خیلی زیاد باشد شعاع ناحیه یک کلاس تنها شامل خود بردار پشتیبان می‌شود و وقتی این مقدار خیلی کم باشد، شعاع ناحیه بخش زیادی از داده‌های آموزشی یا همه‌ی آن را در می‌گیرد. با این توضیح به نظر می‌رسد تصویر ۶ متعلق به یک تنظیم با گامای بالا یعنی تنظیم ه و تصویر ۱ متعلق به یک تصویر با گامای پایین یعنی تنظیم د خواهد بود.

سوال ۵

از هشت داده موجود یک داده به اشتباه دسته‌بندی شده است پس:

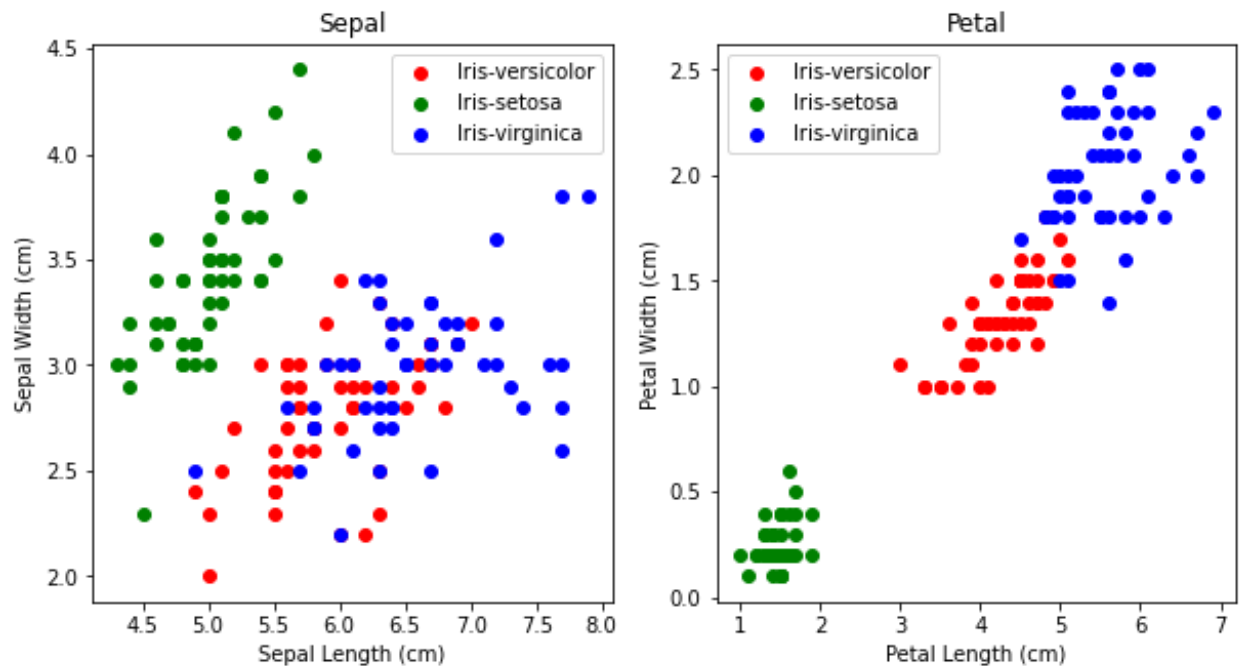
$$\epsilon^1 = \frac{1}{8}$$

$$\alpha^1 = \frac{1}{2} \ln \left(\frac{\frac{7}{8}}{\frac{1}{8}} \right) = \frac{1}{2} \ln(7) \approx 0.97$$

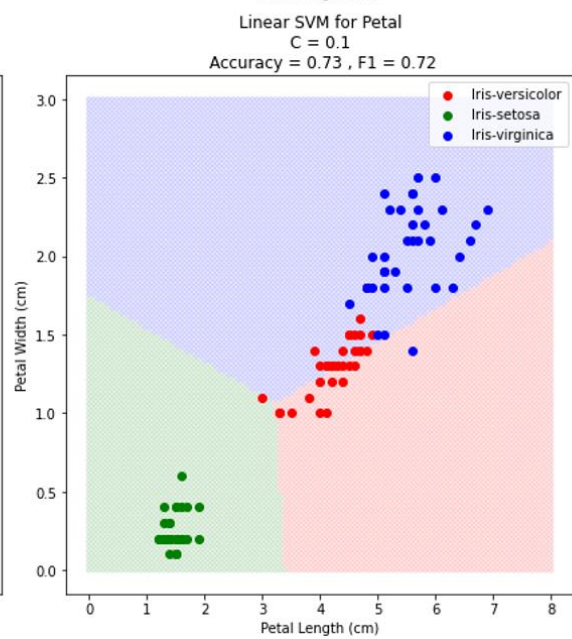
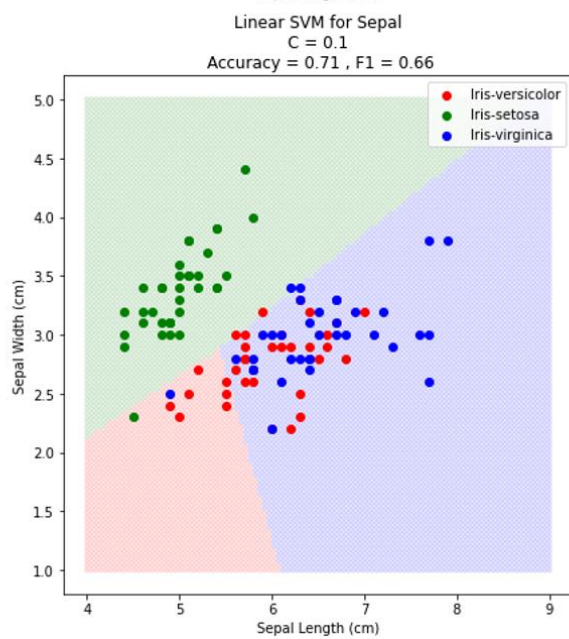
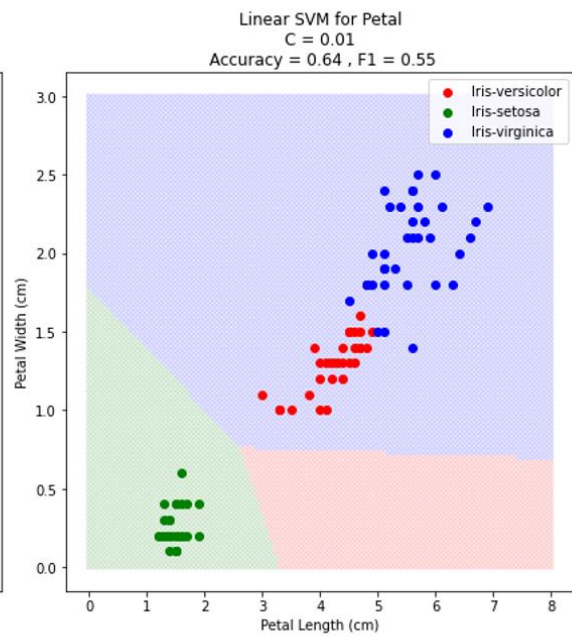
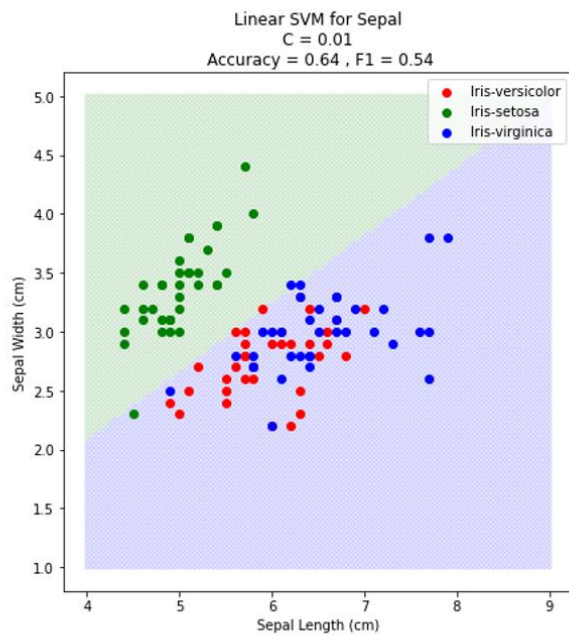
بخش اول: پیاده‌سازی

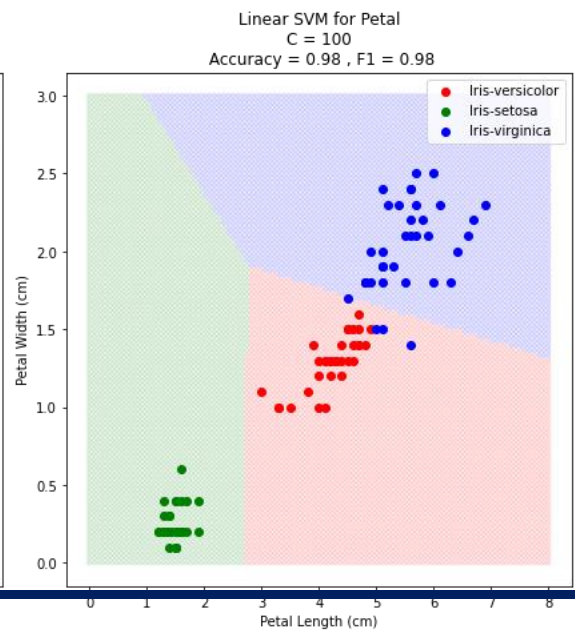
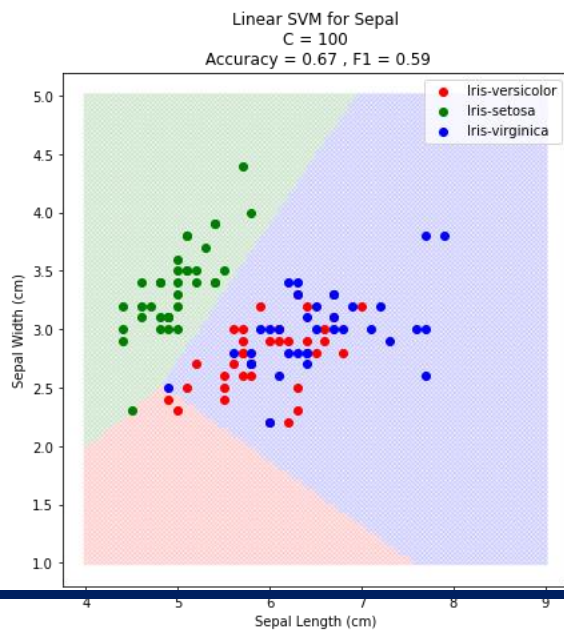
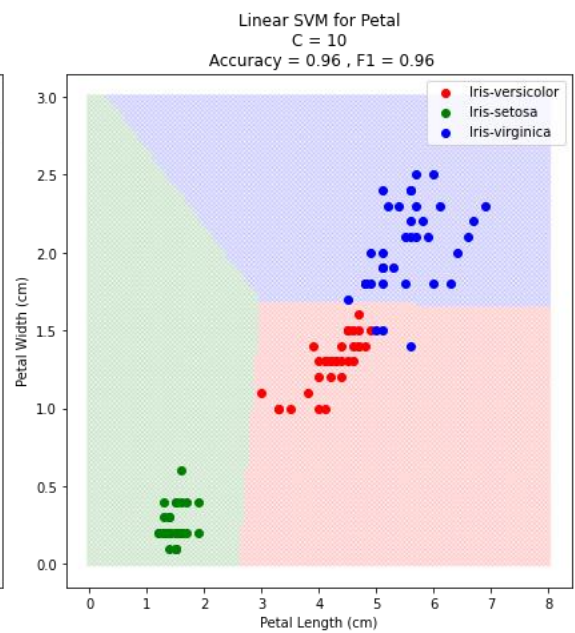
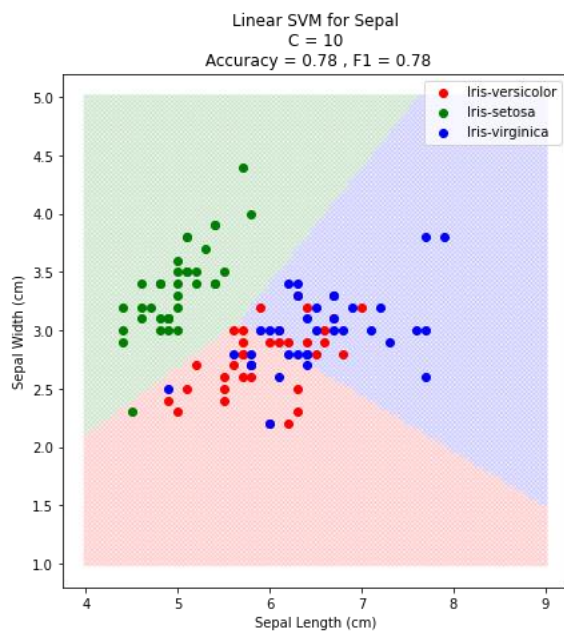
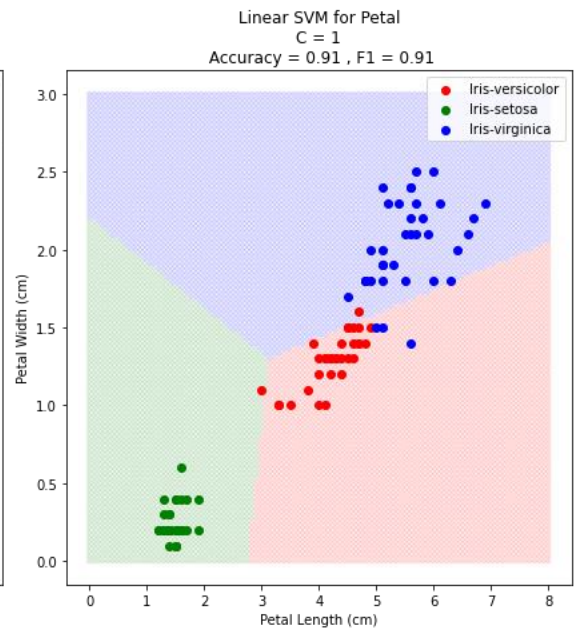
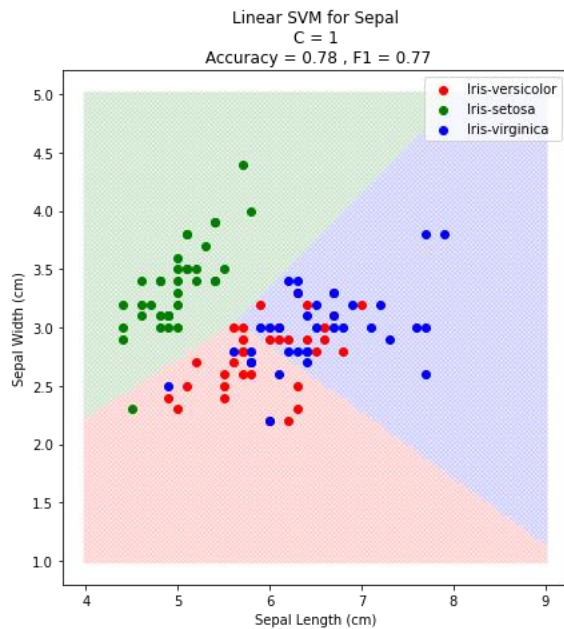
سوال ۱

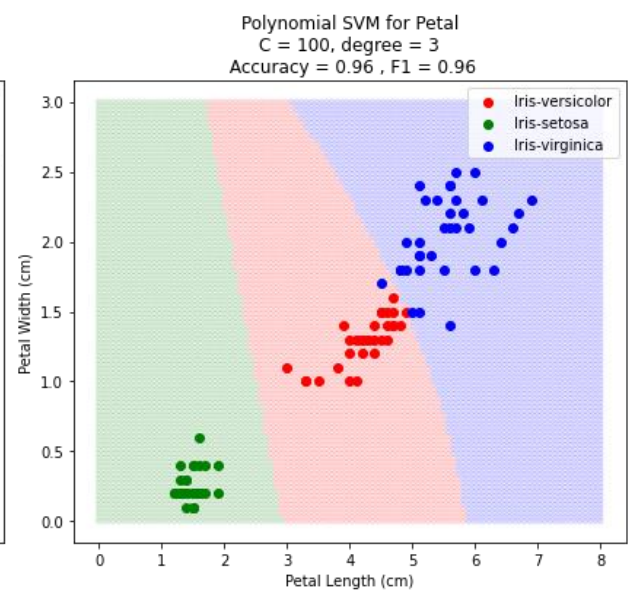
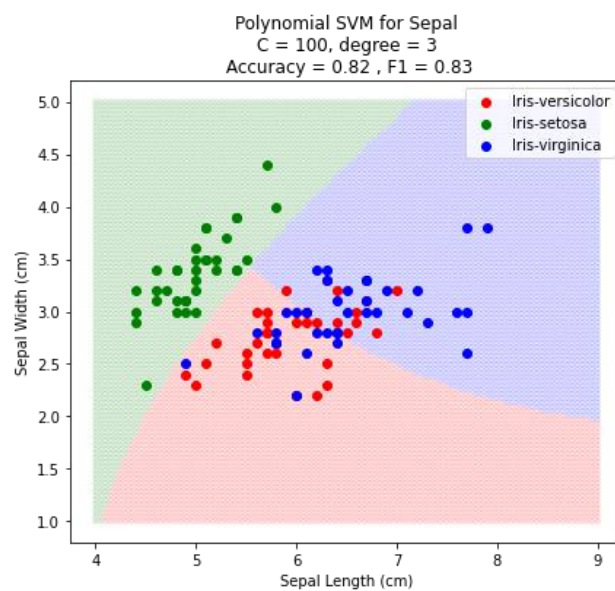
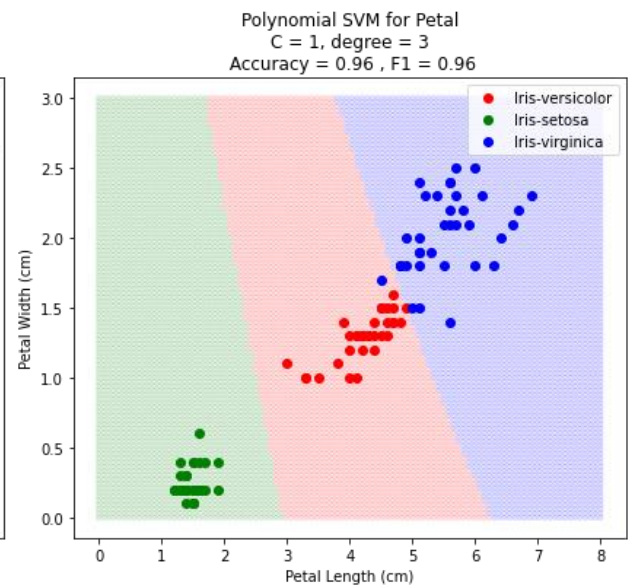
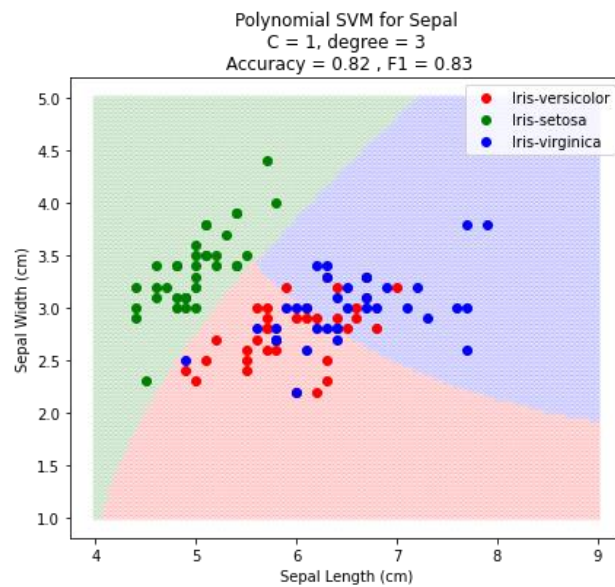
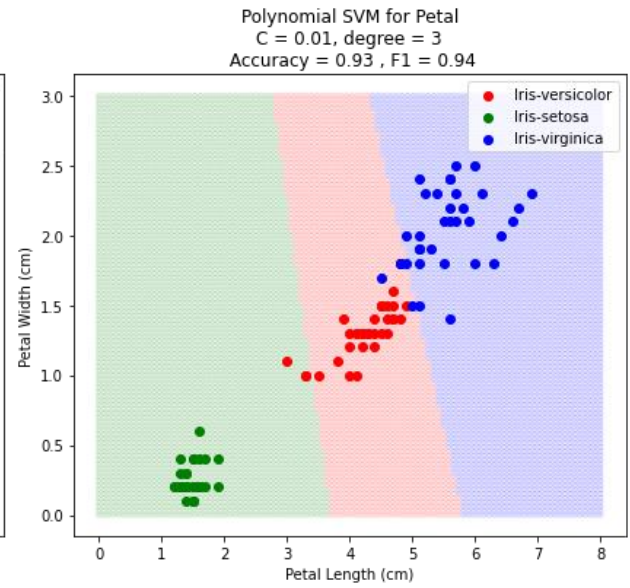
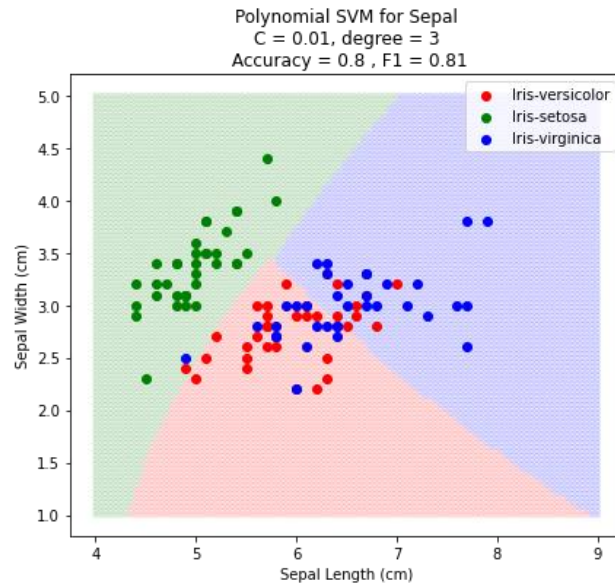
(الف)

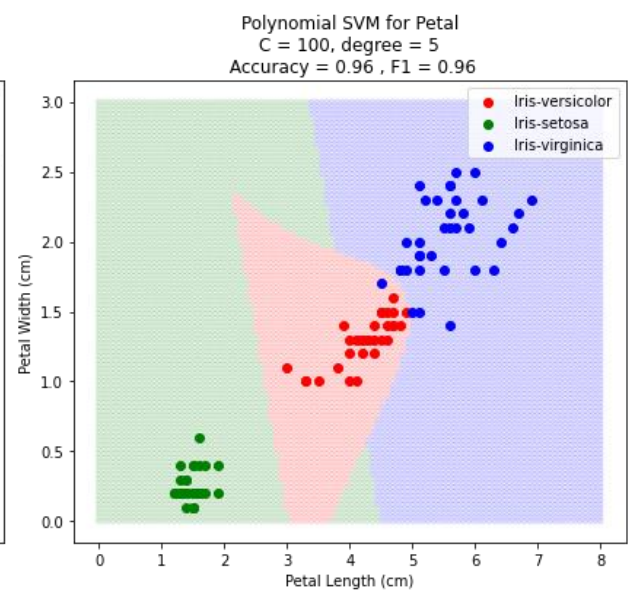
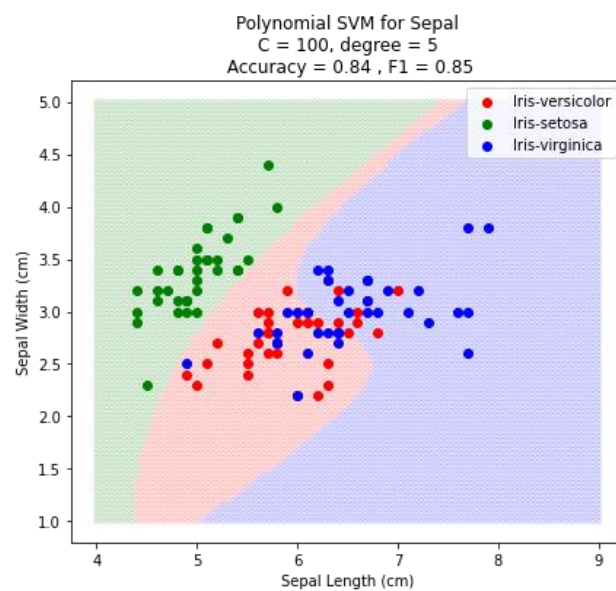
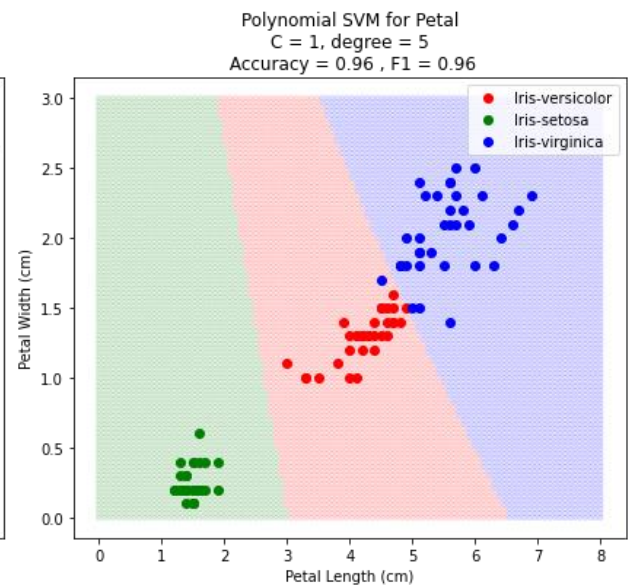
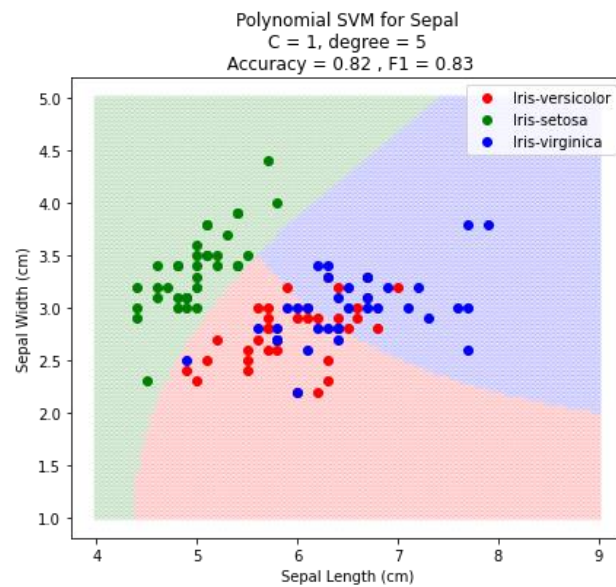
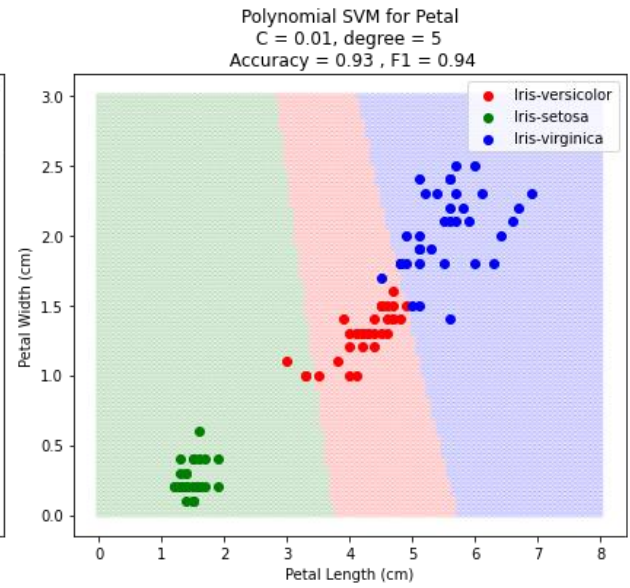
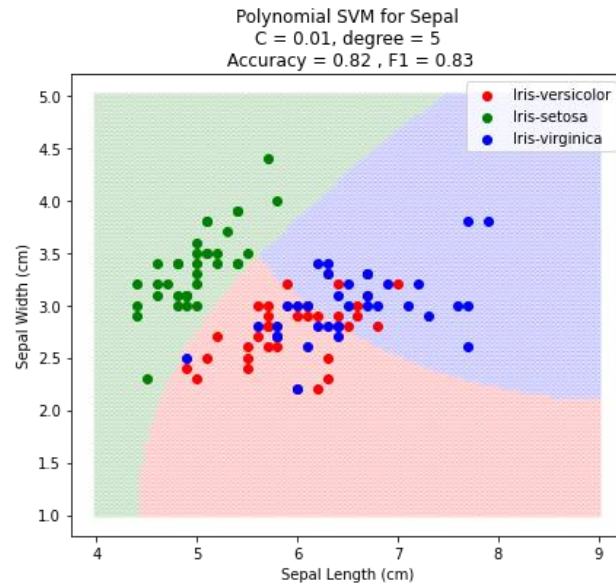


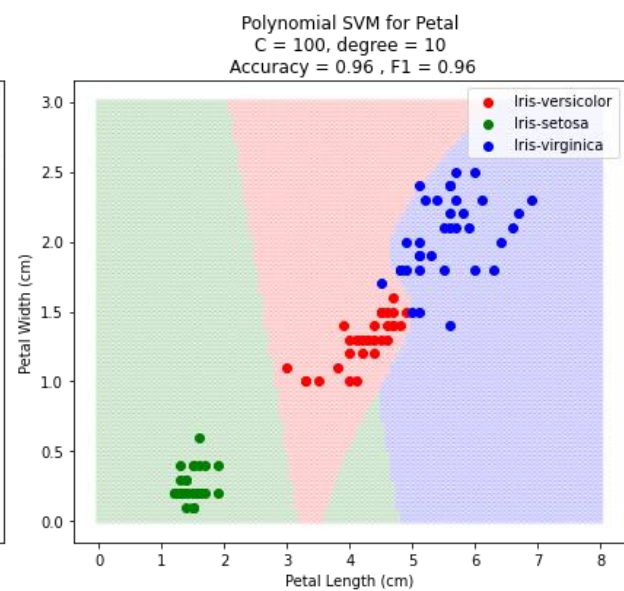
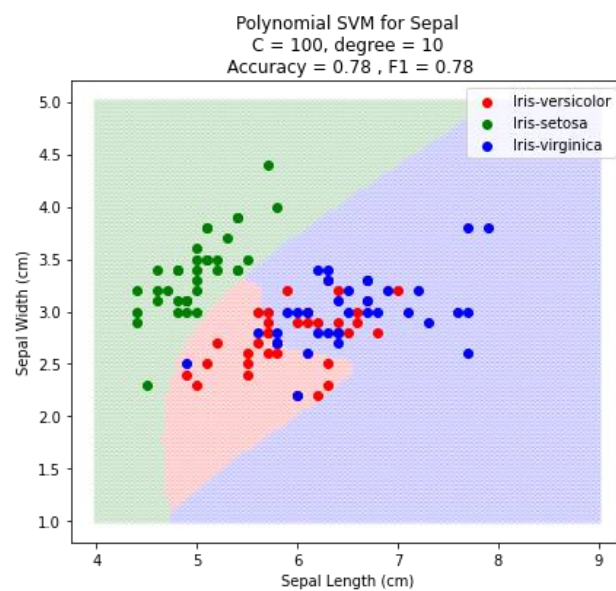
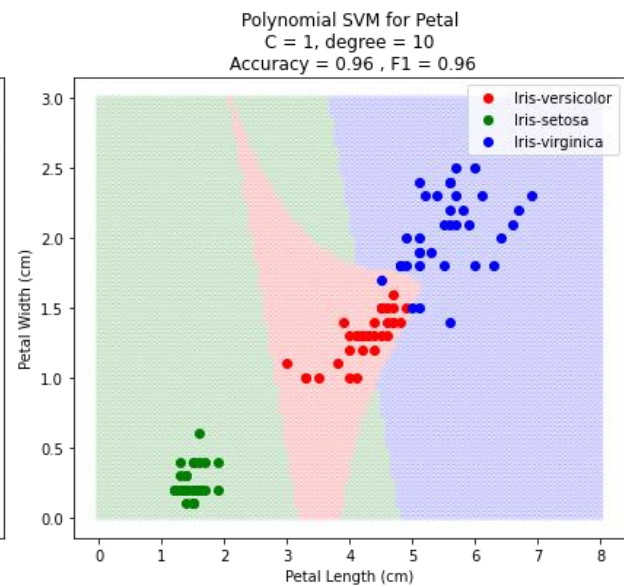
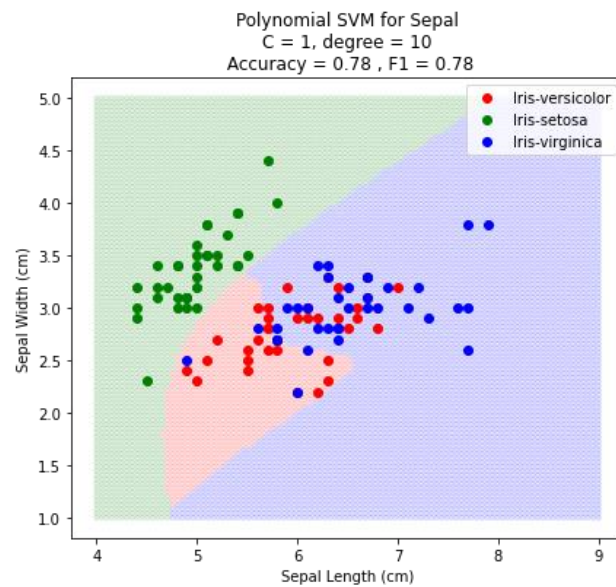
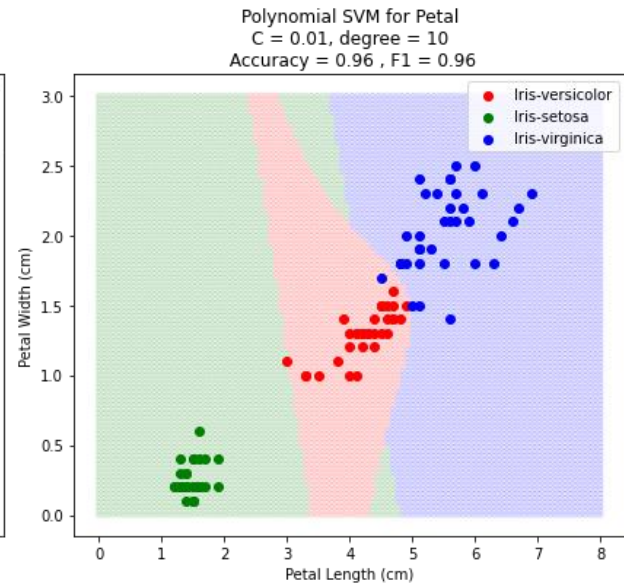
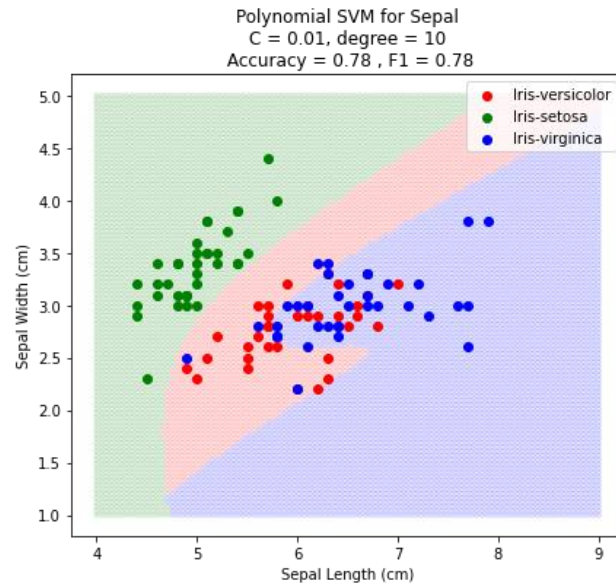
ب) نمودارهای این قسمت در صفحات آتی و در ادامه آورده شده است. برای کرنل‌های چندجمله‌ای، RBF و سیگموید در هر صفحه پارامتر C تغییر می‌کند و پارامتر دیگر در آن صفحه ثابت است. برای پارامتر C در این سه کرنل مقادیر 0.1 ، 1 و 100 تست شده است. برای کرنل چندجمله‌ای درجه ۳، ۵ و ۱۰ تست شده است. برای کرنل RBF مقدار گامای 0.25 ، 0.5 و 1 تست شده است. نهایتاً برای کرنل سیگموید مقادیر گامای 0.0025 ، 0.005 و 0.01 تست شده است.

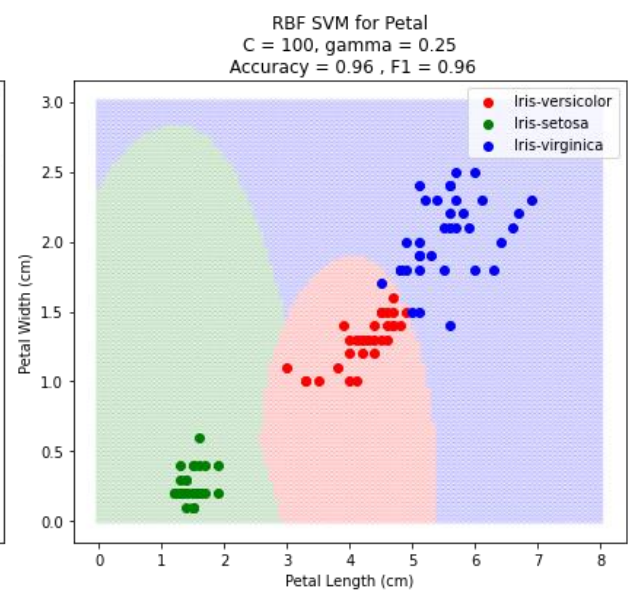
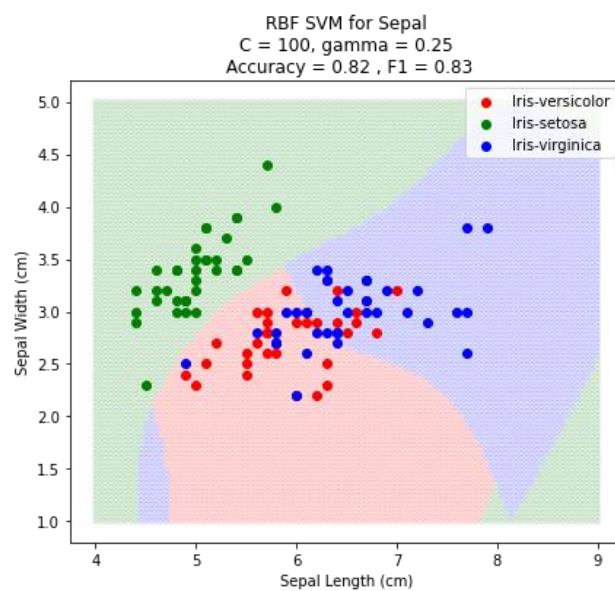
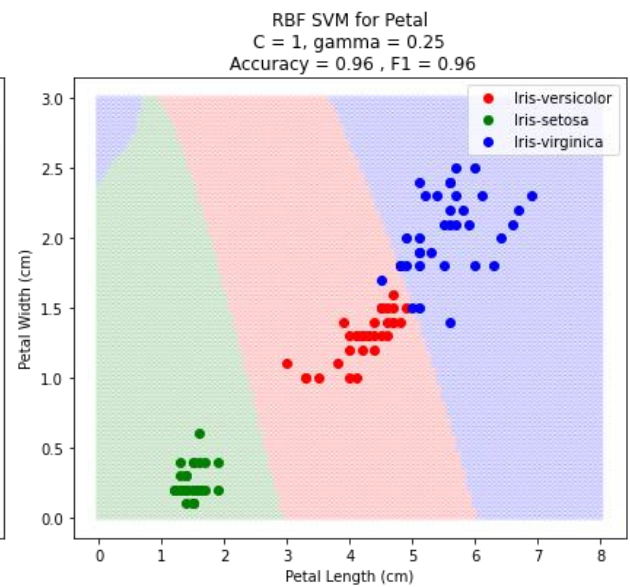
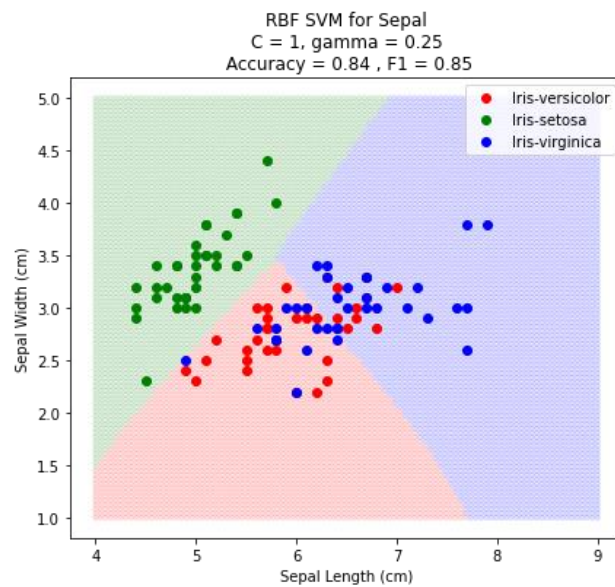
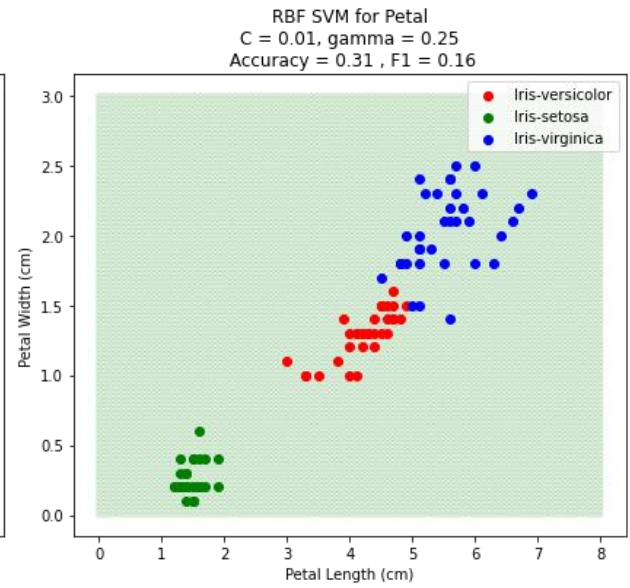
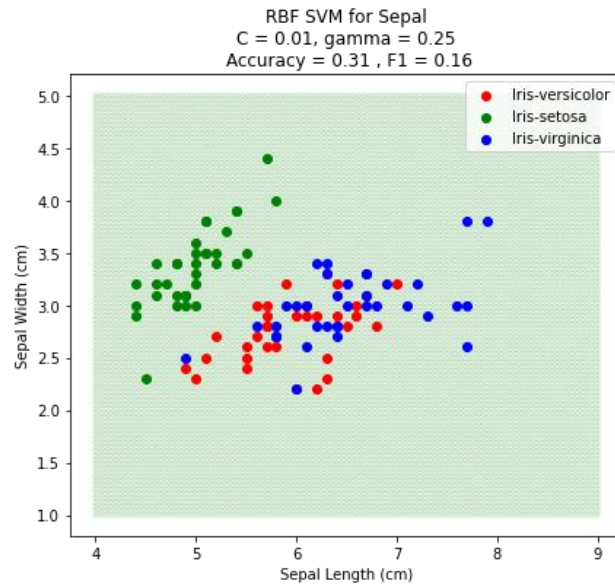


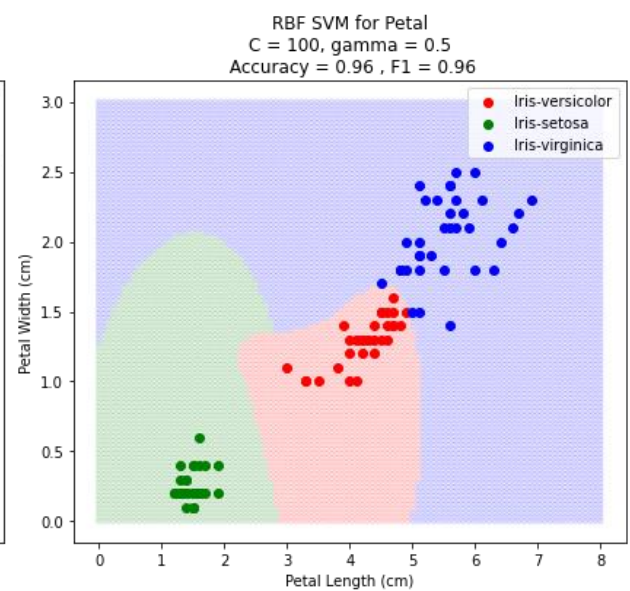
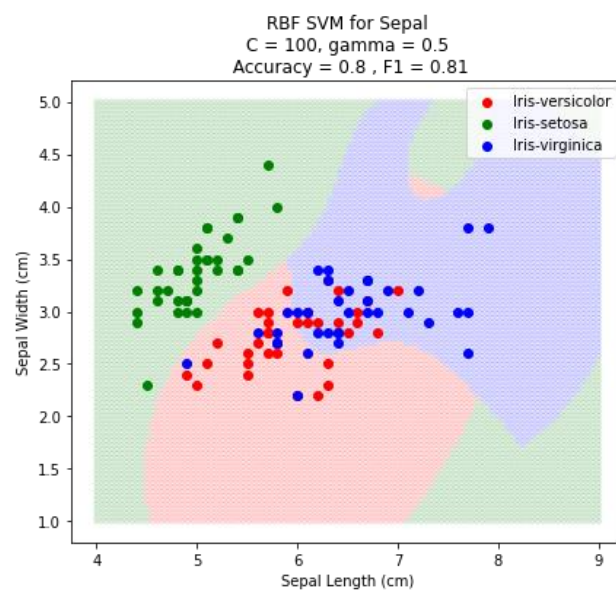
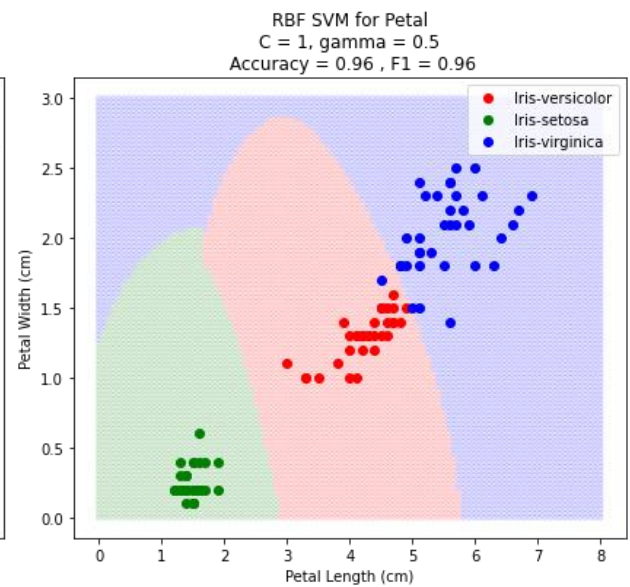
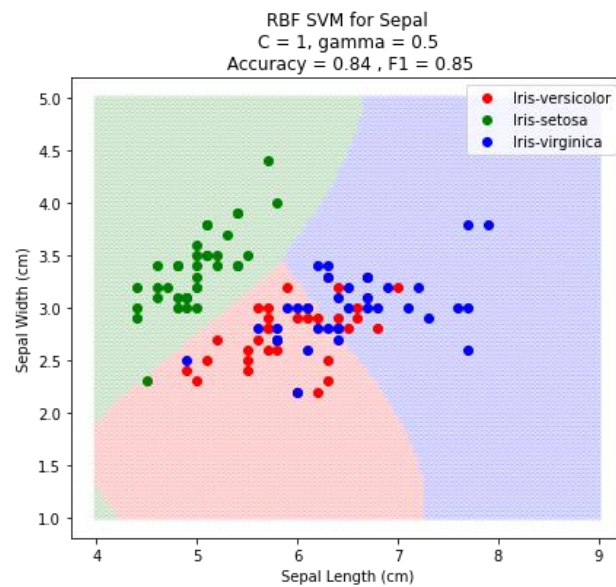
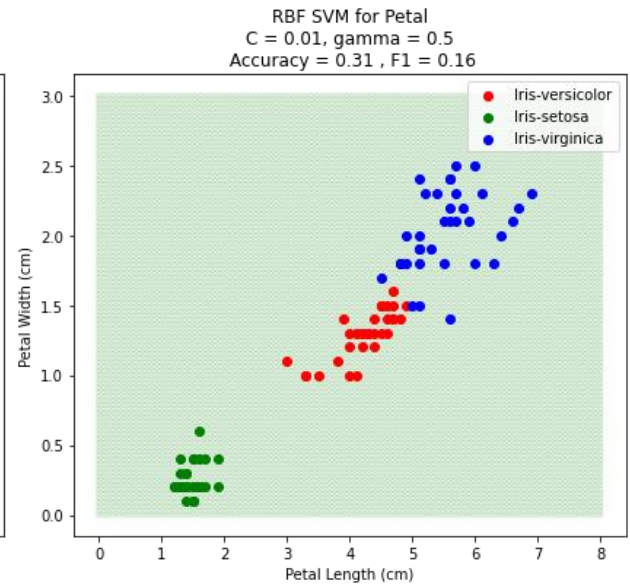
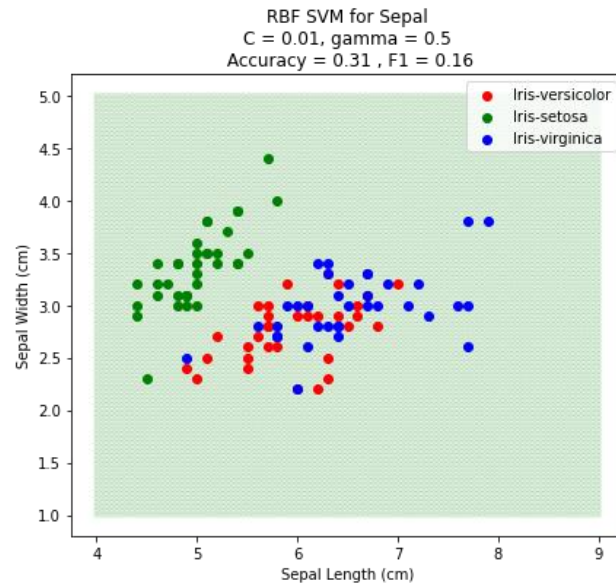


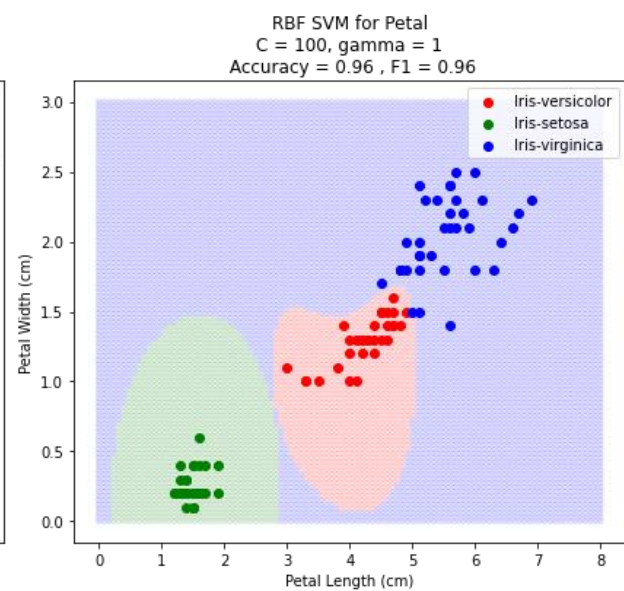
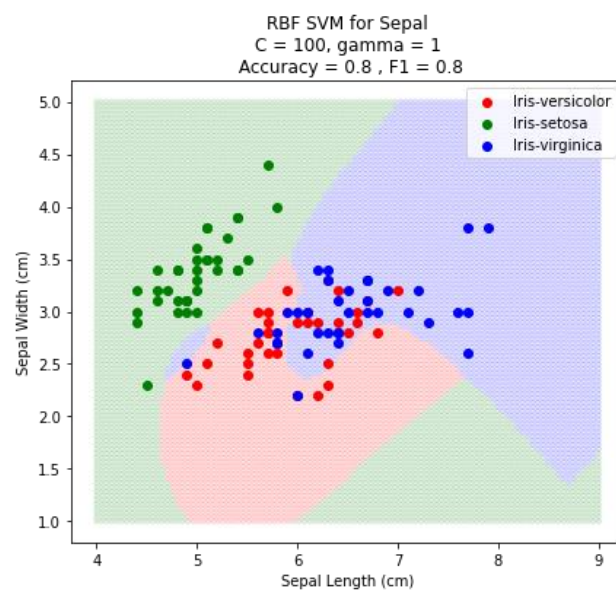
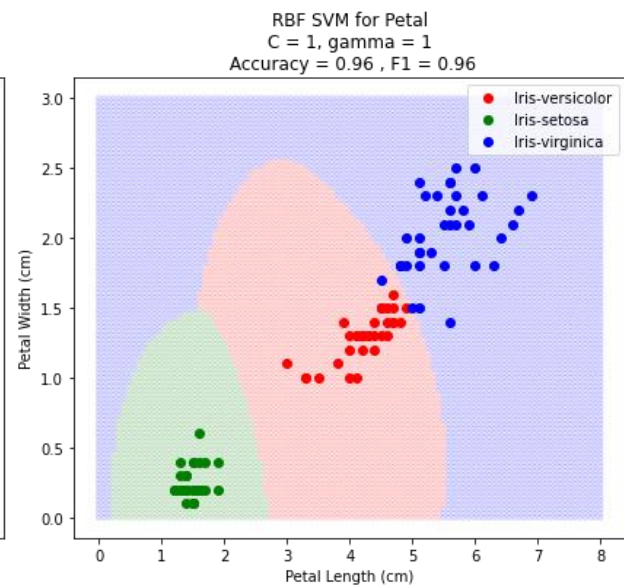
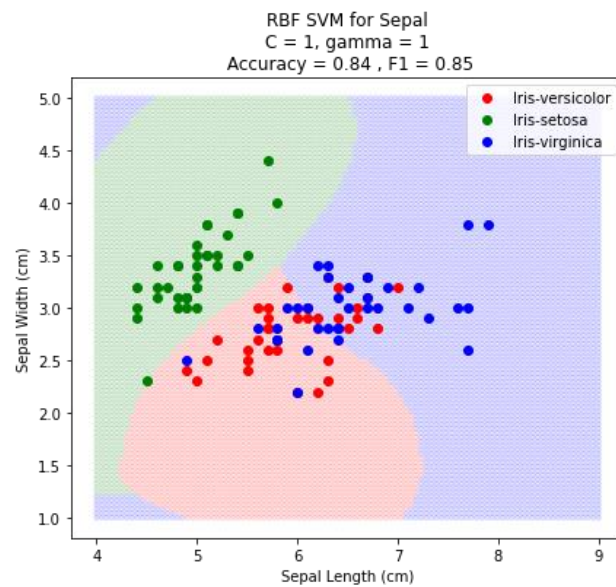
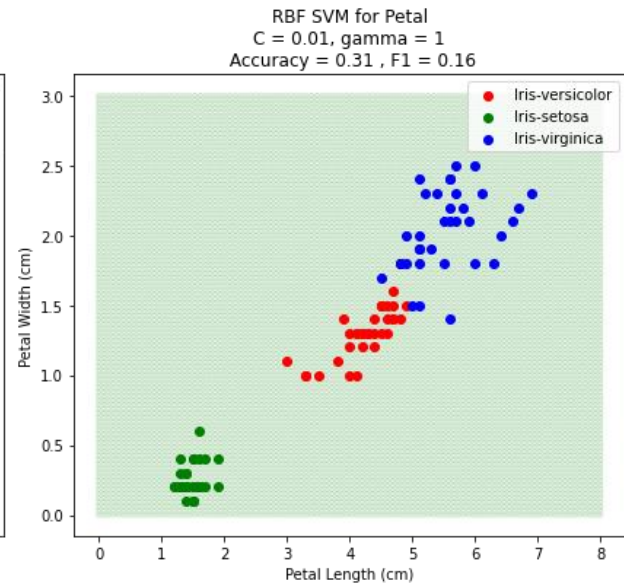
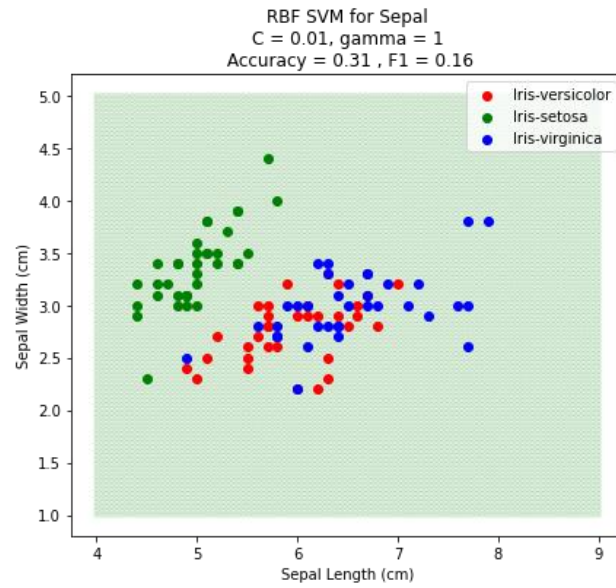


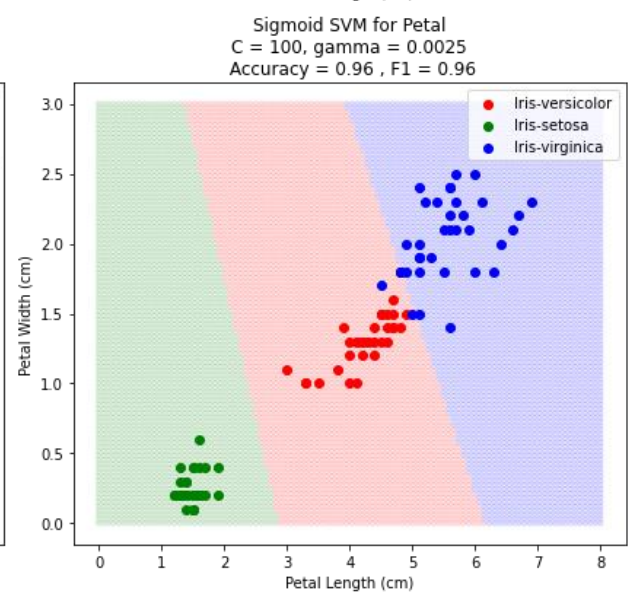
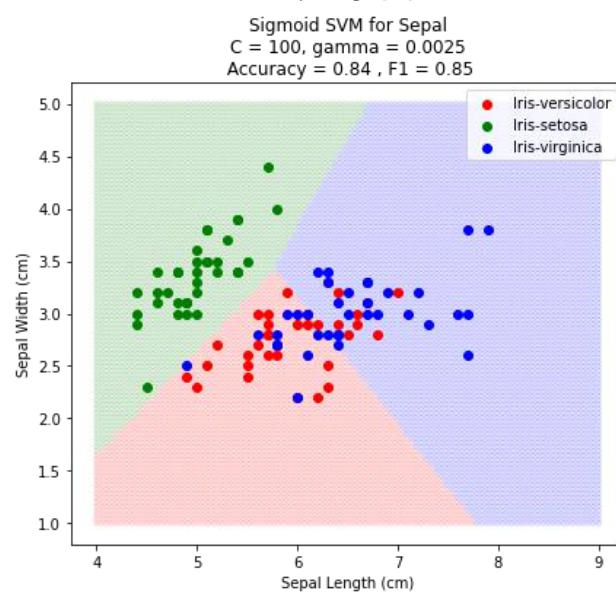
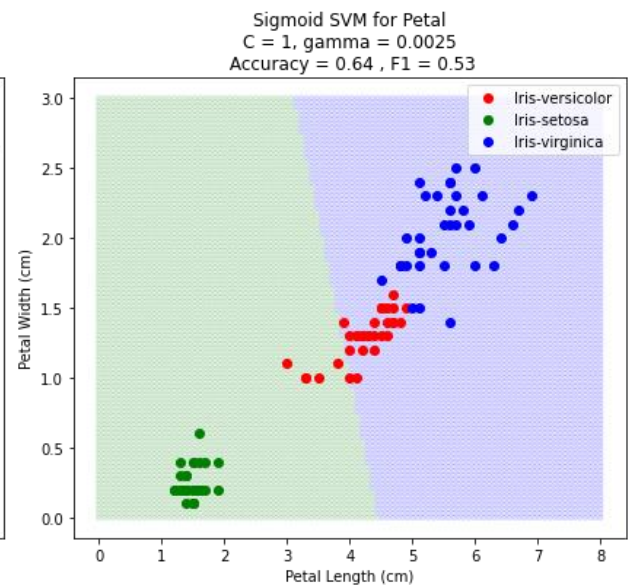
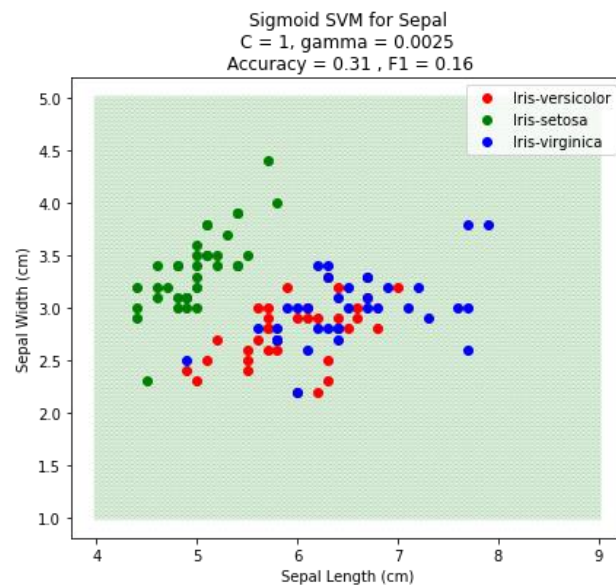
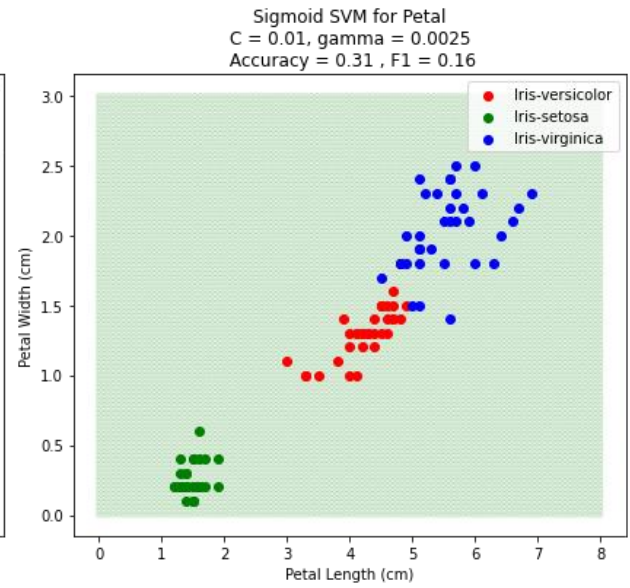
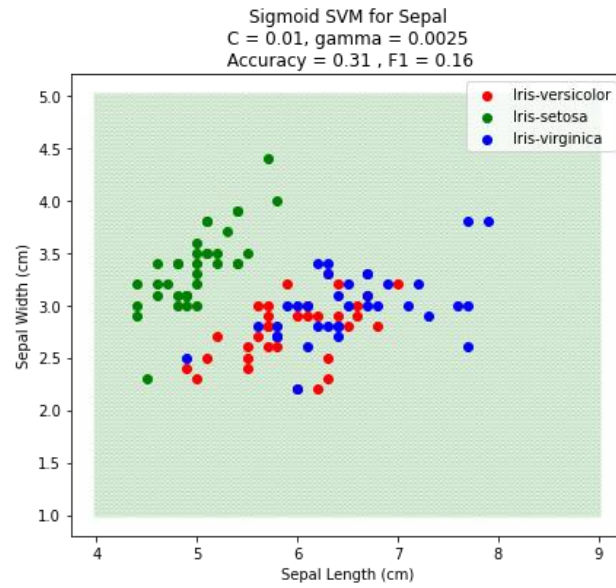


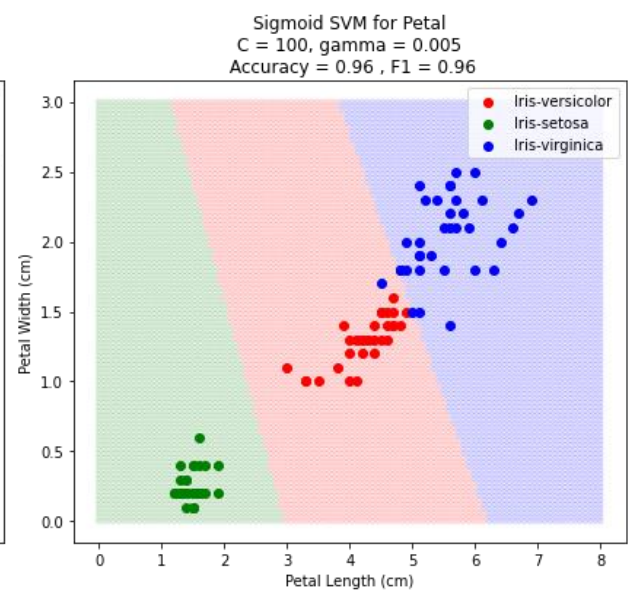
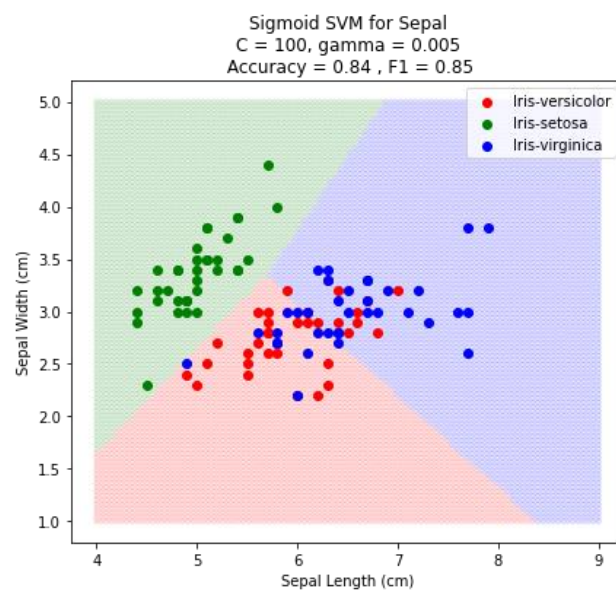
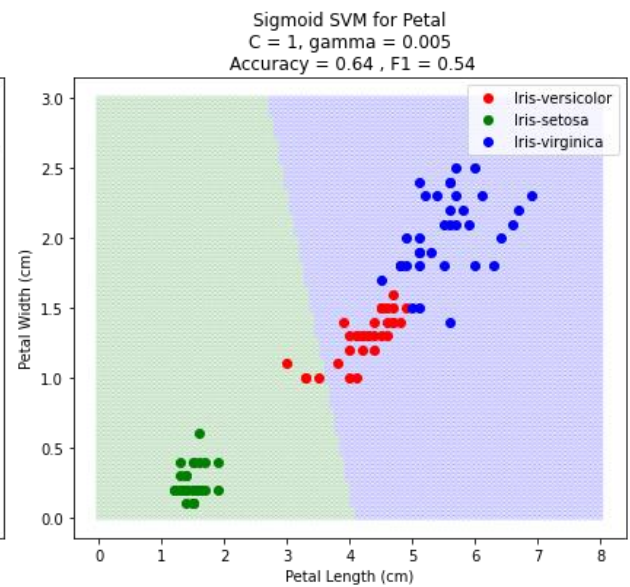
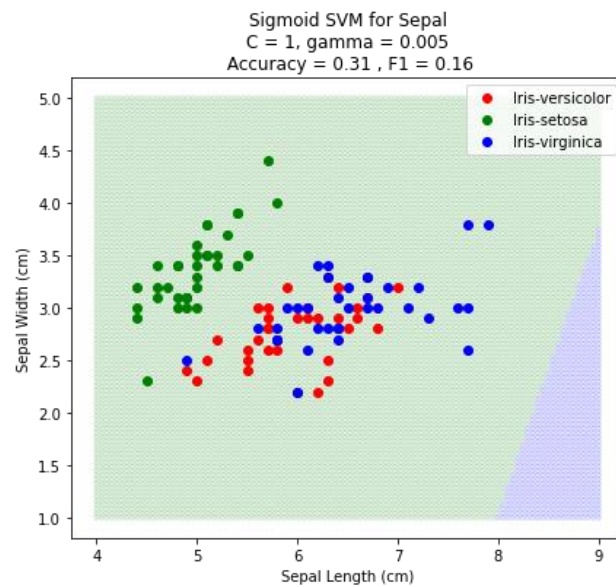
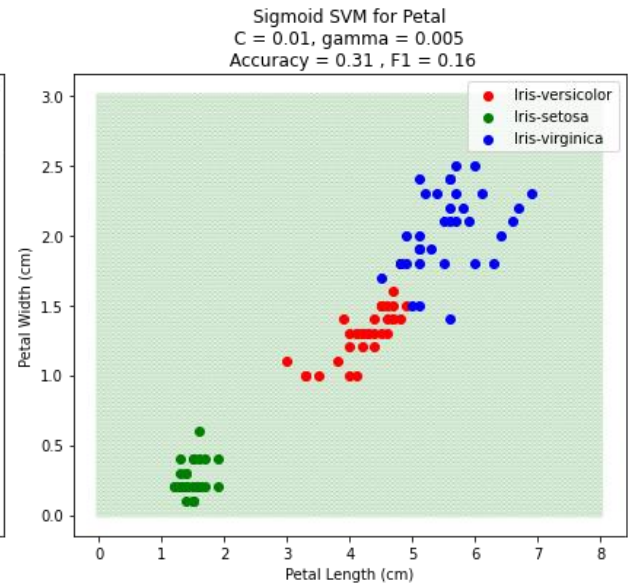
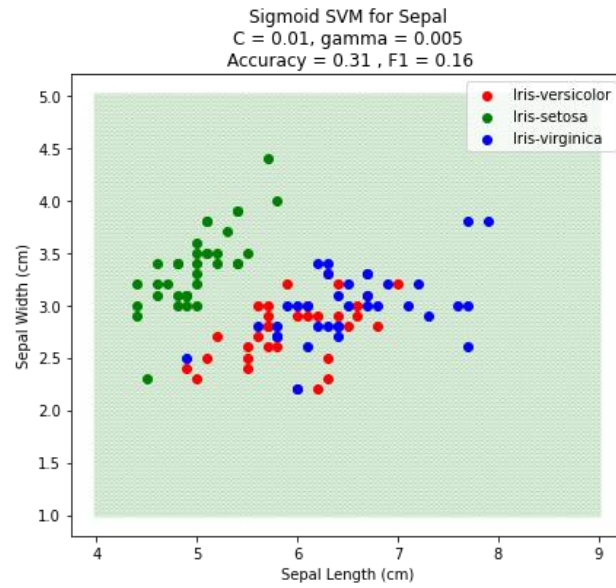


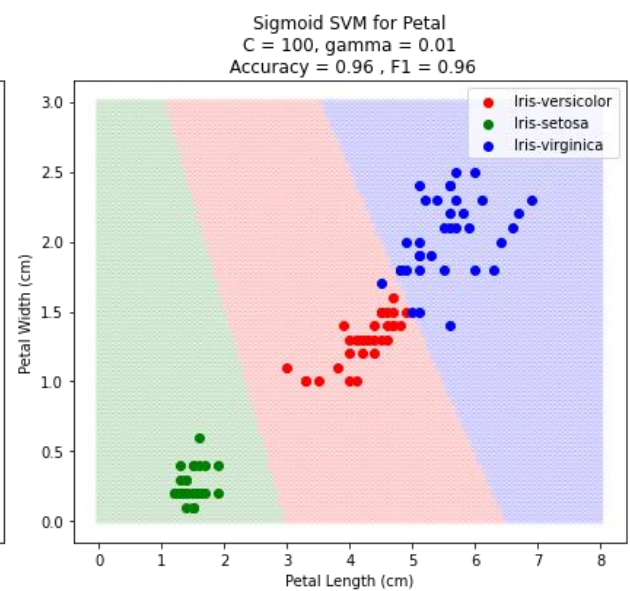
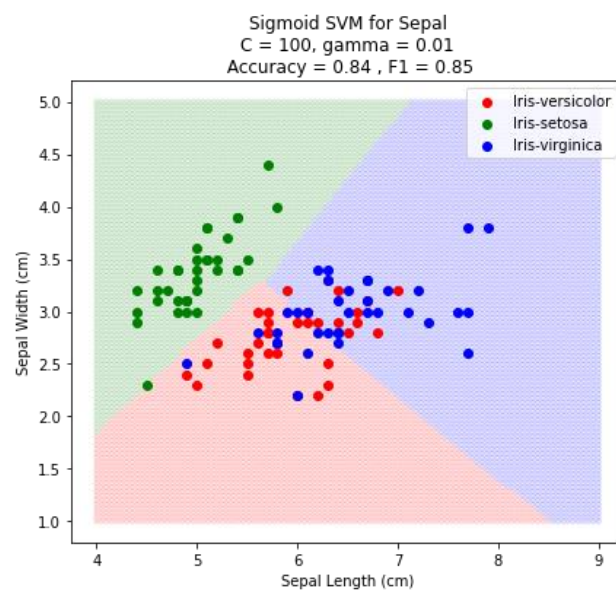
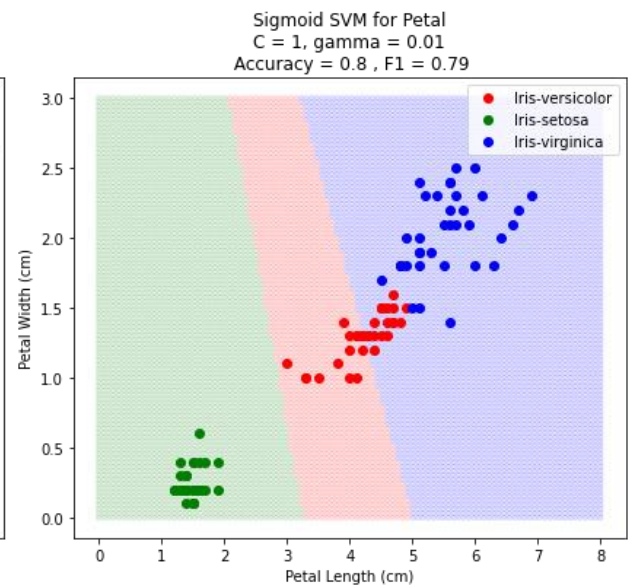
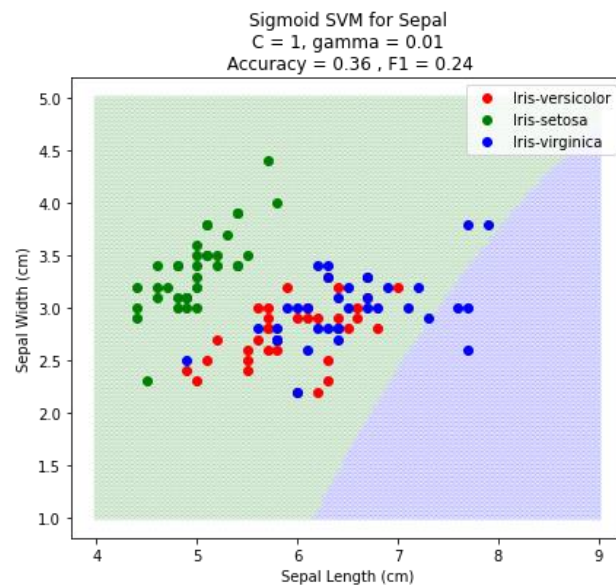
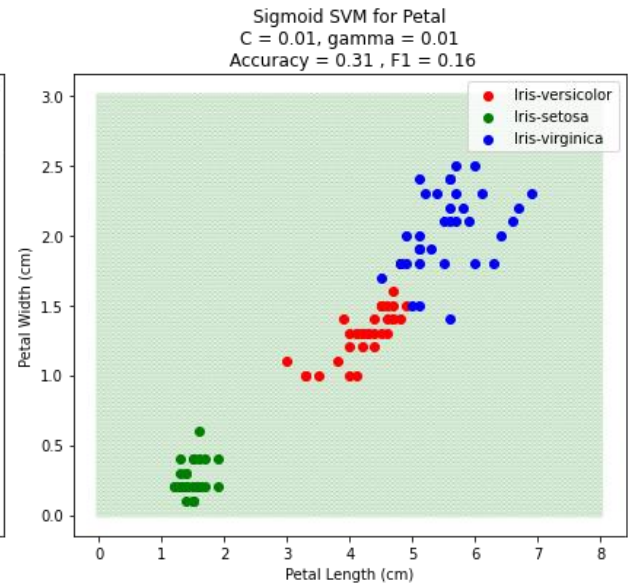
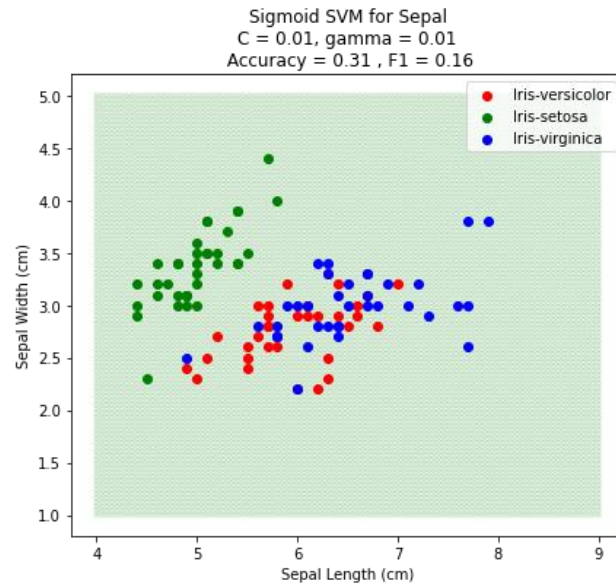












ج) در نمودارهای قسمت ب این مقادیر برای هر قسمت ذکر شده است. خلاصه این‌ها را در جداول زیر می‌توانید ببینید:

جدول دقت کرنل خطی

Sepal		Petal		C
F1	Accuracy	F1	Accuracy	
۰/۵۴	۰/۶۴	۰/۵۵	۰/۶۴	۰/۰۱
۰/۶۶	۰/۷۱	۰/۷۲	۰/۷۳	۰/۱
۰/۷۷	۰/۷۸	۰/۹۱	۰/۹۱	۱
۰/۷۸	۰/۷۸	۰/۹۶	۰/۹۶	۱۰
۰/۵۹	۰/۶۷	۰/۹۸	۰/۹۸	۱۰۰

جدول دقت کرنل چندجمله‌ای برای ویژگی‌های Petal

Degree = ۱۰		Degree = ۵		Degree = ۳		
F1	Accuracy	F1	Accuracy	F1	Accuracy	
۰/۹۶	۰/۹۶	۰/۹۴	۰/۹۳	۰/۹۴	۰/۹۳	C = ۰/۰۱
۰/۹۶	۰/۹۶	۰/۹۶	۰/۹۶	۰/۹۶	۰/۹۶	C = ۱
۰/۹۶	۰/۹۶	۰/۹۶	۰/۹۶	۰/۹۶	۰/۹۶	C = ۱۰۰

جدول دقت کرنل چندجمله‌ای برای ویژگی‌های Sepal

Degree = ۱۰		Degree = ۵		Degree = ۳		
F1	Accuracy	F1	Accuracy	F1	Accuracy	
۰/۷۸	۰/۷۸	۰/۸۳	۰/۸۲	۰/۸۱	۰/۸۰	C = ۰/۰۱
۰/۷۸	۰/۷۸	۰/۸۳	۰/۸۲	۰/۸۳	۰/۸۲	C = ۱
۰/۷۸	۰/۷۸	۰/۸۵	۰/۸۴	۰/۸۳	۰/۸۲	C = ۱۰۰

جدول دقت کرنل RBF برای ویژگی‌های Petal

Gamma = ۱		Gamma = ۰/۵		Gamma = ۰/۲۵		
F1	Accuracy	F1	Accuracy	F1	Accuracy	
۰/۱۶	۰/۳۱	۰/۱۶	۰/۳۱	۰/۱۶	۰/۳۱	C = ۰/۰۱
۰/۹۶	۰/۹۶	۰/۹۶	۰/۹۶	۰/۹۶	۰/۹۶	C = ۱
۰/۹۶	۰/۹۶	۰/۹۶	۰/۹۶	۰/۹۶	۰/۹۶	C = ۱۰۰

جدول دقت کرنل RBF برای ویژگی‌های Sepal

Gamma = ۱		Gamma = ۰/۵		Gamma = ۰/۲۵		
F1	Accuracy	F1	Accuracy	F1	Accuracy	
۰/۱۶	۰/۳۱	۰/۱۶	۰/۳۱	۰/۱۶	۰/۳۱	C = ۰/۰۱
۰/۸۵	۰/۸۴	۰/۸۵	۰/۸۴	۰/۸۵	۰/۸۴	C = ۱
۰/۸	۰/۸	۰/۸۱	۰/۸۰	۰/۸۳	۰/۸۲	C = ۱۰۰

جدول دقت کرنل سیگموئید برای ویژگی‌های Petal

Gamma = ۰/۰۱		Gamma = ۰/۰۰۵		Gamma = ۰/۰۰۲۵		
F1	Accuracy	F1	Accuracy	F1	Accuracy	
۰/۱۶	۰/۳۱	۰/۱۶	۰/۳۱	۰/۱۶	۰/۳۱	C = ۰/۰۱
۰/۷۹	۰/۸۰	۰/۵۴	۰/۶۴	۰/۵۳	۰/۶۴	C = ۱
۰/۹۶	۰/۹۶	۰/۹۶	۰/۹۶	۰/۹۶	۰/۹۶	C = ۱۰۰

جدول دقت کرنل سیگموئید برای ویژگی‌های Sepal

Gamma = ۰/۰۱		Gamma = ۰/۰۰۵		Gamma = ۰/۰۰۲۵		
F1	Accuracy	F1	Accuracy	F1	Accuracy	
۰/۱۶	۰/۳۱	۰/۱۶	۰/۳۱	۰/۱۶	۰/۳۱	C = ۰/۰۱
۰/۲۴	۰/۳۶	۰/۱۶	۰/۳۱	۰/۱۶	۰/۳۱	C = ۱
۰/۸۵	۰/۸۴	۰/۸۵	۰/۸۴	۰/۸۵	۰/۸۴	C = ۱۰۰

د) برای کرنل خطی و برای ویژگی‌های مربوط به Petal که داده‌های هر کلاس تا حد مناسبی از یکدیگر جدا شده‌اند، افزایش هرچه بیشتر C ما را به نتایج بهتری می‌رساند ولی برای ویژگی‌های مربوط به Sepal که داده‌ها تا حدی باهم مخلوط هستند می‌بینیم که دقت برای $C = 100$ دقت کاهش پیدا می‌کند و باید C یک مقدار معمولی‌تری اتخاذ کند. از نمودارها هم متوجه می‌شویم که برای مقادیر C پایین این امکان وجود دارد که یک کلاس بسیار کوچک شود. برای مقدار $C=100$ ویژگی‌های Sepal هم همین اتفاق رخ داده است.

برای کرنل‌های چندجمله‌ای، مقادیر بالای C منجر به نتایج بهتری شده است. در مورد ویژگی‌های Petal چون داده‌های آموزشی دارای نویز نیست درجات بالا هم نتایج مناسبی دارد اما برای ویژگی‌های Sepal درجه ۱۰ باعث بیش‌برازش می‌شود و اندازه درجه نه چندان زیاد و نه چندان کم مانند درجه ۵ مناسب است. از روی نمودارها هم می‌توان دید که برای درجه ۱۰، خروجی بیش‌برازش شده و یک شکل تیز حاصل شده است که نمی‌تواند مناسب باشد.

برای کرنل RBF اولین چیزی که مشهود است نتایج مناسب‌تر این مدل در مقایسه با سایر مدل‌هاست. با این نتایج خیلی نمی‌توان در مورد پارامترها صحبت کرد ولی می‌بینیم مانند سایر کرنل‌ها مقدار C ی پایین اصلاً خوب نیست و کل داده‌ها در یک کلاس قرار می‌گیرند. برای گاما به نظر مقادیر پایین‌تر کمی مناسب‌تر بوده‌اند ولی چندان قطعی نیست. از روی نمودارها، در مورد $C=0.01$ می‌بینیم مدل تمام داده‌ها را در فراوان‌ترین کلاس قرار داده است که اصلاً مطلوب نیست. در مورد گاما هم می‌بینیم با مقادیر پایین بیشتر داده‌ها در یک ناحیه بزرگ و نرم قرار گرفته‌اند اما برای گاما‌های بالا این ناحیه بریدگی‌های زیادی پیدا کرده که مطلوب نیست و باعث بیش‌برازش می‌شود.

برای کرنل سیگموید همانند قبل مقدار C بالا نتایج به مراتب بهتری را رقم زده است. مقادیر گامایی که برای این قسمت استفاده شده است نسبت به مقادیر گامای کرنل RBF خیلی کمتر است چراکه برای مقادیر گامای هم اندازه با RBF نتایج بسیار بدی دریافت می‌شد. پس مقدار گامای کرنل سیگموید نباید چندان زیاد باشد ولی در عین حال اگر خیلی هم کم باشد باز نتیجه مناسبی نخواهد داشت. در نمودارهای متناسب با این کرنل هرچه C بیشتر شده است مدل به کلاس‌های کوچک اهمیت بیشتری داده

است و فضا به نحو معقول‌تری جدا شده است. در مورد گاما هم وضعیت تقریباً مشابه است یعنی وقتی گاما مقدار کمی دارد کلاس‌های کوچک سهم کمتری از فضا را دارند و بعضاً نادیده گرفته می‌شوند.

سوال ۲

الف) ماتریس درهم‌ریختگی و معیارهای دقت به شرح زیر است:

	۰	۱	۲	۳	۴	۵	۶	۷	۸	۹
۰	۳۱۵	۴	۰	۰	۰	۰	۱	۳	۴۰	۰
۱	۰	۱۸۸	۱۳۵	۱	۲	۰	۰	۰	۰	۳۸
۲	۰	۱۸	۳۴۲	۱	۱	۰	۰	۲	۰	۰
۳	۰	۷	۰	۳۲۸	۰	۰	۰	۰	۰	۱
۴	۰	۲	۰	۱	۳۵۴	۰	۰	۰	۰	۷
۵	۰	۰	۰	۱۱۶	۱۳	۱۶۲	۰	۰	۶	۳۸
۶	۳	۰	۲	۳	۰	۰	۲۹۷	۲۹	۲	۰
۷	۰	۲۲	۱۰	۳	۰	۳	۳	۲۹۴	۰	۲۹
۸	۱۶	۰	۰	۰	۰	۱۲	۲	۲۲	۲۸۳	۱
۹	۰	۵۲	۰	۶۳	۱۳	۰	۲	۰	۲	۲۰۴

F1	Recall	Precision	Accuracy
۷۸/۴۹	۷۸/۹۷	۸۰/۶۴	۷۹/۱

ب و ج) معیارهای دقت برای هر یک از حالات ۵، ۱۰، ۲۰ و ۵۰ درخته در جدول زیر ذکر شده است:

F1	Recall	Precision	Accuracy	Number Trees
۴۶/۴۴	۵۱/۳۴	۵۴/۸۷	۵۲/۳۴	۵
۴۷/۸	۵۳/۷۸	۶۳/۶۶	۵۴/۸۶	۱۰
۵۹/۳۶	۶۲/۹۷	۶۳/۸۸	۶۳/۶۹	۲۰
۵۵/۱۷	۵۹/۰۶	۶۰/۱۹	۶۱/۰۳	۵۰

د) با استفاده از پارامترهای پیش‌فرض مدل می‌توان به دقت مناسبی رسید:

F1	Recall	Precision	Accuracy
۹۶/۷۱	۹۶/۷۹	۹۶/۷۳	۹۶/۶۸

اما برای آنکه از مناسب بودن این پارامترها اطمینان حاصل شود، مجموعه آموزشی را به دو قسمت ۷۰٪ و ۳۰٪ تقسیم کردم؛ قسمت ۳۰٪ برای اعتبارسنجی و قسمت ۷۰٪ برای آموزش مدل. نهایتاً برای پارامترهای `gamma`، `max_depth`، `subsample` و `learning_rate` تعداد حالت انتخاب کردم و مدلی که بهترین `Accuracy` را روی مجموعه اعتبارسنجی کسب کرده بود، به عنوان مدل نهایی انتخاب کردم. معیارهای دقت برای این مدل در جدول زیر گزارش شده‌اند که نزدیک به دقت‌های پارامترهای پیش‌فرض ولی کمی کمتر است:

F1	Recall	Precision	Accuracy
۹۶/۳۳	۹۶/۴۴	۹۶/۳۷	۹۶/۳۱