گزارش جامع تحلیل مدل‌های طبقه‌بندی بر روی داده‌های دست‌نویس اعداد

کدی که نوشتیم منحنی های ROC را برای چهار مدل طبقه بندی مختلف ایجاد می کند :

- Naive Bayes
- Decision Tree
- k-Nearest Neighbors
- SVM

منحنی‌های ROC با استفاده از اعتبارسنجی متقاطع 5 برابری ایجاد می‌شوند و AUC(مساحت زیر منحنی) برای هر کلاس محاسبه می‌شود.

در اینجا یک تفکیک کد و تجزیه و تحلیل حاصل آمده است:

1 .آماده سازی داده ها:

کد مجموعه داده اعداد را از scikit-learn بارگیری می کند که حاوی تصاویری از ارقام دست نویس است.

تابع label_binarize برچسب های هدف را به یک ماتریس باینری تبدیل می کند که برای محاسبه منحنی های ROC برای مسائل چند کلاسه ضروری است.

2 .آموزش و ارزیابی مدل:

کد یک فرهنگ لغت از مدل ها را تعریف می کند که هر کدام شامل یک الگوریتم طبقه بندی متفاوت است (Naive Bayes ، Decision Tree، k-Nearest Neighbors،SVM).

برای آموزش و ارزیابی هر مدل و محاسبه ماتریس سردرگمی، دقت و منحنی‌هایROC ، اعتبارسنجی متقابل 5 برابری را انجام می‌دهد.

3 .محاسبه و رسم منحنیROC :

برای هر مدل، کد منحنی ROC را برای هر کلاس با استفاده از تابع roc_curve محاسبه می کند.

سپس AUC برای هر کلاس با استفاده از تابع auc محاسبه می شود.

سپس کد منحنی‌های ROC را برای همه کلاس‌ها برای هر مدل ترسیم می‌کند و نمودارها را در فهرستی با نام "plots" ذخیره می‌کند.

4 .نتایج و تجزیه و تحلیل:

کد ماتریس سردرگمی، دقت، زمان آموزش و امتیازات AUC را در یک فایل متنی به نام "results.txt" ذخیره می کند.

در نهایت زمان آموزش هر مدل را چاپ می کند.

Confusion Matrix:

results.txt

```
Model: Naive Bayes
Confusion Matrix:
[[174   0   0   0   2   0   0   1   0   1]
 [  0 137   8   0   0   0   5   4  18  10]
 [  0  13 112   1   1   2   1   0  45   2]
 [  0   2   6 133   0   8   0   7  22   5]
 [  3   2   2   0 142   1   3  25   3   0]
 [  0   1   0   3   2 158   1   8   5   4]
 [  0   1   1   0   1   3 174   0   1   0]
 [  0   0   1   0   2   1   0 174   1   0]
 [  0  20   3   0   1   5   0  10 133   2]
 [  1  11   0   8   2   4   1  17  23 113]]
Accuracy: 0.81
Training Time: 0.02 seconds

Model: Decision Tree
Confusion Matrix:
[[165   0   0   0   3   2   2   0   3   3]
 [  0 135  11   9   3   3   2   2  10   7]
 [  1  16 122   6   3   0   6   3  14   6]
 [  0   4   8 139   1   5   4   1  11  10]
 [  4  15   0   3 133   2  11   6   5   2]
 [  9   2   0   2  12 147   2   1   3   4]
 [  3   2   1   0   4   3 165   1   2   0]
 [  3   0   3   3  11   2   0 145  10   2]
 [  3  10   7   9   1   3   0   0 132   9]
 [  0   9   5   6   0  10   0   8   7 135]]
Accuracy: 0.79
Training Time: 0.09 seconds
```

```
Model: k-Nearest Neighbour
Confusion Matrix:
[[177   0   0   0   1   0   0   0   0   0]
 [  0 178   0   0   2   1   1   0   0   0]
 [  0   1 170   0   0   0   0   1   5   0]
 [  0   0   1 172   0   1   0   2   4   3]
 [  0   2   0   0 176   0   0   2   1   0]
 [  0   0   0   0   0 177   1   0   0   4]
 [  0   2   0   0   0   1 178   0   0   0]
 [  0   0   0   0   0   0   0 176   0   3]
 [  0  13   2   1   0   0   0   1 157   0]
 [  0   2   0   4   1   2   0   1   1 169]]
Accuracy: 0.96
Training Time: 0.14 seconds

```

```
Model: SVM
Confusion Matrix:
[[177   0   0   0   1   0   0   0   0   0]
 [  0 180   0   0   0   0   1   0   1   0]
 [  0   1 174   0   0   0   0   0   2   0]
 [  0   0   1 167   0   2   0   3   9   1]
 [  0   1   0   0 177   0   0   0   1   2]
 [  0   0   0   0   0 178   1   0   0   3]
 [  0   0   0   0   0   1 179   0   1   0]
 [  0   0   0   0   0   0   0 167   1  11]
 [  0   8   0   0   0   1   0   0 163   2]
 [  0   0   0   2   0   2   0   5   2 169]]
Accuracy: 0.96
Training Time: 1.04 seconds

```

تحلیل و بررسی نتایج بدست آمده

دقت:

بر اساس نمرات دقت، هر دو **SVM** و **k-Nearest Neighbors** بالاترین دقت (96٪) را به دست می آورند و پس از آن **Naive Bayes (81٪)** و **Decision Tree (79٪)** قرار دارند. این نشان می دهد که هر دو **SVM** و **k-Nearest Neighbors** برای این مشکل طبقه بندی بهتر از دو مدل دیگر مناسب هستند.

زمان آموزش:

زمان آموزش برای هر مدل به طور قابل توجهی متفاوت است **Naive Bayes** . سریعترین است و پس از آن **Decision Tree** ، **k-Nearest Neighbors** و **SVM** (کندترین) قرار دارند.

منحنی هایROC :

- هر منحنی مدل طبقه بندی متفاوتی را نشان می دهد:
- Naive Bayes
- Decision Tree
- k-Nearest Neighbors
- SVM

مفاهیم نقشهROC :

- محور X (نرخ مثبت کاذبFPR): این محور نشان دهنده نسبت نمونه های مثبت طبقه بندی نادرست است.
- محور Y (نرخ مثبت واقعیTPR): این محور نشان دهنده نسبت نمونه های مثبت طبقه بندی شده (مثبت های واقعی) به تعداد کل نمونه های مثبت است.
- خط مورب: خط چین مشکی یک طبقه‌بندی تصادفی را نشان می‌دهد که 50 درصد احتمال دارد که یک نمونه را به درستی طبقه‌بندی کند.

- منحنی های بالای خط: در حالت ایده آل، می خواهید منحنی ها به همان اندازه بالا باشند
- منحنی‌های ROC نشان می‌دهند که مدل‌ها در جداسازی کلاس‌ها نسبتاً خوب عمل می‌کنند، به طوری که برخی از کلاس‌ها تفکیک بهتری نسبت به بقیه دارند.
- هرچه منحنی بالاتر باشد، عملکرد مدل بهتر است.
- یک طبقه‌بندی کننده کامل دارای منحنی است که مستقیماً تا 1.0 در محور y و سپس مستقیماً به 1.0 در محور x می‌رود.

.

AUCمنطقه زیر منحنی:

- مقدار AUC برای هر مدل محاسبه شده و در Confusion Matrix نشان داده شده است.
- مقادیر AUC از 0 تا 1 متغیر است که مقادیر بالاتر نشان دهنده عملکرد بهتر مدل است.
- AUC 1نشان دهنده یک طبقه بندی کننده کامل است.
- AUC 0.5نشان دهنده یک طبقه بندی تصادفی است.

تجزیه و تحلیل منحنی هایROC :

- **Naive Bayes :** مدل Naive Bayes بهترین عملکرد را در بین مدل های نشان داده شده دارد. منحنی‌های ROC آن بسیار نزدیک به گوشه سمت چپ بالا هستند، که نشان‌دهنده TPR بالا و FPR پایین در همه کلاس‌ها است. این نشان دهنده تمایز خوب بین مثال های مثبت و منفی است.
- **درخت تصمیم :**مدل درخت تصمیم عملکرد متوسطی را نشان می دهد. نمودار نشان دهنده عملکرد مدل تصمیم درخت در تشخیص کلاس های

مختلف است .نمودار ROC نشان می دهد که مدل تصمیم درخت در تشخیص کلاس های مختلف دارای عملکرد متفاوتی است و برخی کلاس ها را بهتر از دیگران تشخیص می دهد

- **Nearest Neighbors :** مدل عملکرد خوبی را نشان می دهد و منحنی ها کمی بهتر از مدل Decision Tree است. این توانایی تشخیص بهتر از درختان تصمیم را نشان می دهد.
- **SVM :** مدل SVM عملکردی مشابه با مدل k-Nearest Neighbors نشان می‌دهد، که تبعیض خوبی بین طبقات نشان می‌دهد.

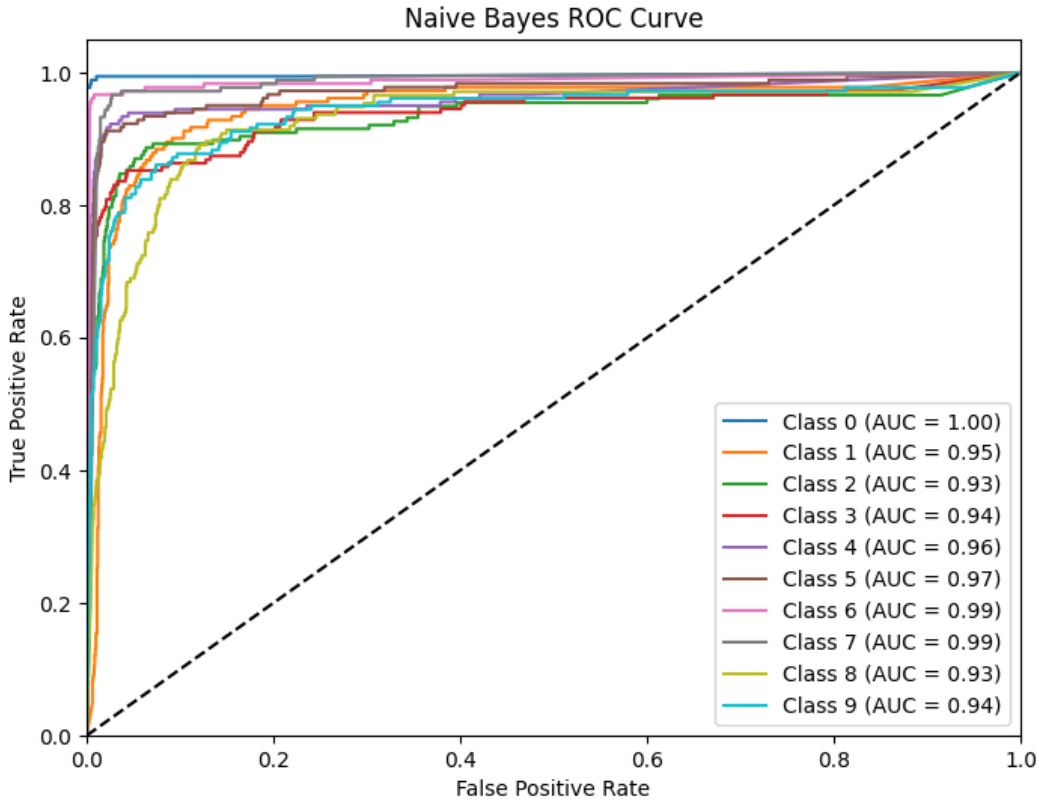
به طور کلی:

بر اساس منحنی‌های ROC و مقادیرAUC ، مدل Naive Bayes از مدل‌های دیگر در این مجموعه داده خاص بهتر عمل می‌کند. با این حال، توجه به این نکته مهم است که عملکرد مدل بسته به مجموعه داده‌ها و مشکل خاصی که به آن پرداخته می‌شود، می‌تواند متفاوت باشد.

فایل متنی ارائه شده با اطلاعات اضافی مانند ماتریس های سردرگمی، نمرات دقت، و زمان آموزش بیشتر از این یافته ها پشتیبانی می کند. این نشان می‌دهد که Naive Bayes بالاترین دقت و سریع‌ترین زمان تمرین را دارد و آن را به گزینه‌ای مطلوب در این سناریو تبدیل می‌کند.

نتایج و تجزیه و تحلیل نمودارها:

- **Naive Bayes**



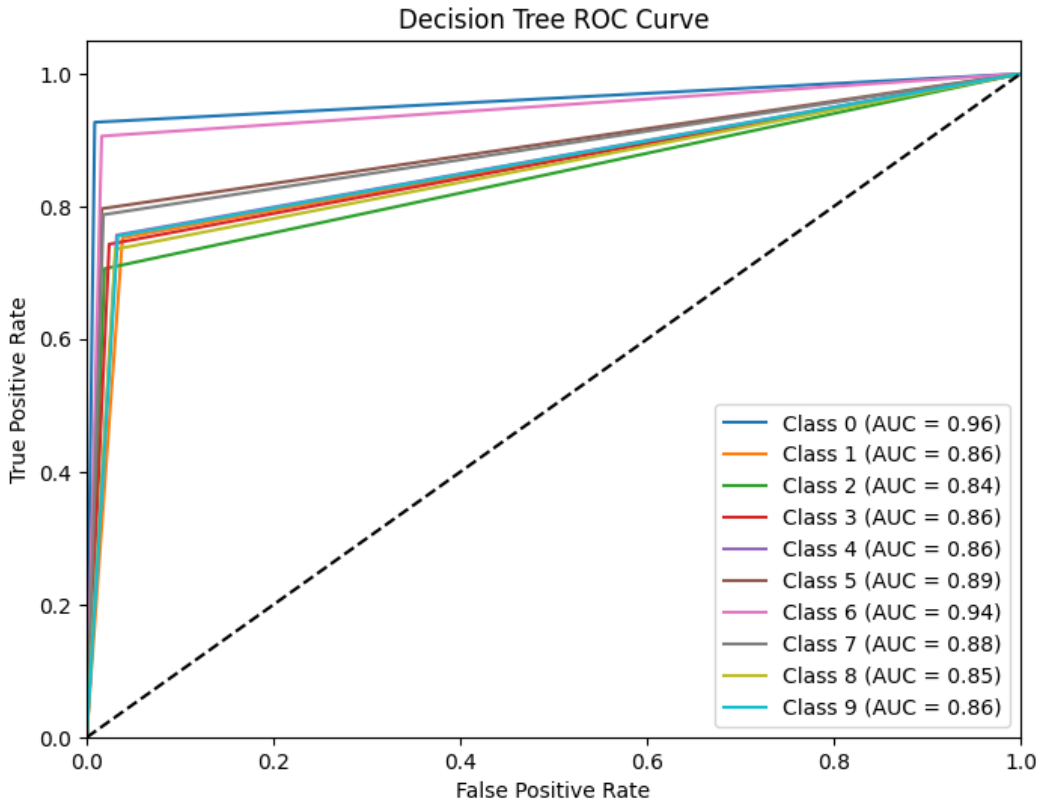
طرح ROC ارائه شده که برای کلاس های مختلف رسم شده است. در این ارزیابی، ارزیابی مثبت کاذب (FPR) در مقابل ارزیابی واقعی کلاس (TPR) برای هر ترسیم شده است.

بر اساس طرح، می توان مشاهده کرد که کلاس ۶ و کلاس ۷ دارای بهترین عملکرد هستند، زیرا منحنی های آنها به گوشه بالا-چپ تصویر نزدیک تر هستند. این نشان می دهد که این کلاس ها دارای ارزش واقعی بالا و مثبت کاذب پایین هستند.

کلاس های 0، 1، 2، 3، 4، 5، 8 و 9 دارای عملکرد متوسطی هستند، زیرا منحنی های آنها در وسط قرار گرفتن آنهاست. این نشان می دهد که این کلاس ها دارای ارزش واقعی و مثبت کاذب متوسطی هستند.

نهایتاً، می توان دید که مدل Naive Bayes دارای بهترین عملکرد است، زیرا منحنی های همه کلاس ها در این مدل به گوشه بالا-چپ تصاویر نزدیک تر هستند.

- **Decision Tree**



تصویر منحنی های ROC را برای یک مسئله طبقه بندی چند طبقه با استفاده از مدل درخت تصمیم نشان می دهد. نمودار نرخ مثبت واقعی را در برابر نرخ مثبت کاذب برای هر کلاس نشان می دهد. مساحت زیر هر منحنی (AUC) نیز نشان داده شده است.

در اینجا خلاصه ای از اطلاعات منتقل شده در طرح آمده است:

- منحنی ROC :هر خط نشان دهنده منحنی ROC برای یک کلاس خاص است. منحنی نشان می‌دهد که مدل چقدر خوب می‌تواند بین نمونه‌های مثبت و منفی برای آن کلاس در موارد مختلف تمایز قائل شود
- AUC : مساحت زیر منحنی (AUC) معیاری از عملکرد کلی مدل برای هر کلاس است AUC .بالاتر نشان دهنده عملکرد بهتر است.
- آستانه :آستانه به قطع احتمال اشاره دارد
- عملکرد کلاس :نمودار نشان می دهد که مدل برای هر کلاس چقدر خوب عمل می کند. به عنوان مثال، کلاس 0

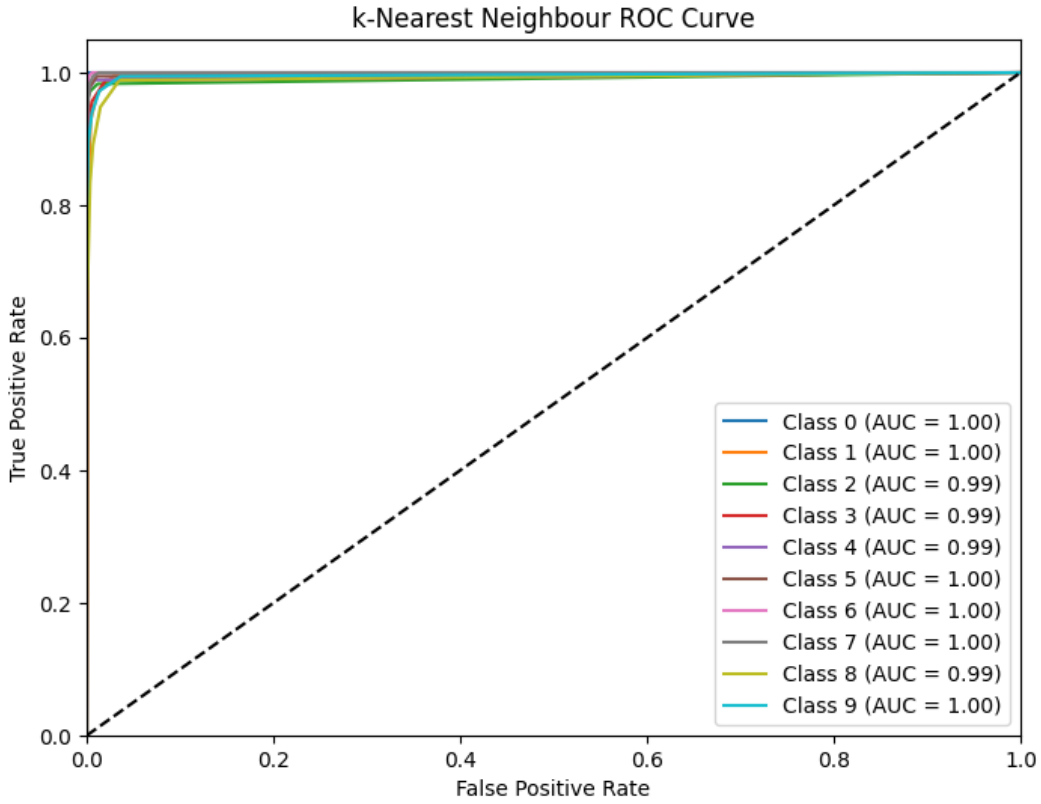
بر اساس طرح، به نظر می‌رسد که این مدل برای همه کلاس‌ها عملکرد مناسبی دارد، اگرچه کلاس 0 بالاترین AUC و در نتیجه بهترین عملکرد را دارد.

در اینجا برخی از مشاهدات از طرح:

- به نظر می رسد این مدل توانایی خوبی در تشخیص نمونه های مثبت از نمونه های منفی برای همه طبقات دارد.
- این مدل به ویژه در کلاس 0 عملکرد خوبی دارد، در حالی که عملکرد آن برای کلاس های دیگر کمی پایین تر است.
- مقادیر AUC برای اکثر کلاس ها بالای 0.80 است که به طور کلی عملکرد خوبی در نظر گرفته می شود.

به طور کلی، نمودار نشان می دهد که مدل درخت تصمیم یک انتخاب مناسب برای این مشکل طبقه بندی چند طبقه است. با این حال، ممکن است برای درک دلایل تغییرات جزئی در عملکرد در کلاس‌های مختلف و بررسی بهبودهای احتمالی، تحلیل بیشتری لازم باشد.

- **K-Nearest Neighbors (KNN)**



تحلیل نمودارKNN

- کلاس 0 :(AUC = 1.00) منحنی این کلاس معمولاً به گوشه بالا-چپ نزدیک است، نشان می دهد بالای آن.
- کلاس 1 :(AUC = 1.00) منحنی این کلاس معمولاً به گوشه بالا-چپ نزدیک نشان می دهد، می دهد بالای مدل در تمایز بین 1 و سایر کلاس ها هستند.
- کلاس 2 :(AUC = 0.99) منحنی این کلاس کمی پایین‌تر از کلاس 0 و 1 است، اما هنوز نشان‌دهنده مدل بالای مدل در تمایز بین است.
- کلاس ۳ (AUC = ۰/۹۹ :(منحنی این کلاس مشابه کلاس ۲ است.
- کلاس ۴ (AUC = ۰/۹۹ :(منحنی این کلاس مشابه کلاس ۲ و ۳ است.

- کلاس 5 :(AUC = 1.00) منحنی این کلاسیک معمولاً به گوشه بالا-چپ نزدیک است، نشان می دهد مدل های مختلف در تمایز بین 5 و سایر کلاس ها.
- کلاس 6 :(AUC = 1.00) منحنی این کلاس معمولاً به گوشه بالا-چپ نزدیک نشان است، ارائه می دهد مدل بالای مدل در تمایز بین.
- کلاس ۷ (AUC = ۱٫۰۰ :(منحنی این کلاس معمولاً به گوشه بالا-چپ نزدیک نشان می دهد، می دهد
- کلاس ۸ (AUC = ۰٫۹۹ :(منحنی این کلاس کمی پایین‌تر از کلاس ۵، ۶ و ۷ است، اما هنوز نشان‌دهنده مدل کلاس در تمایز بین ۸ و سایر کلاس‌ها است.
- کلاس ۹ (AUC = ۱٫۰۰ :(منحنی این کلاسیک معمولاً به گوشه بالا-چپ نزدیک است، نشان می‌دهد

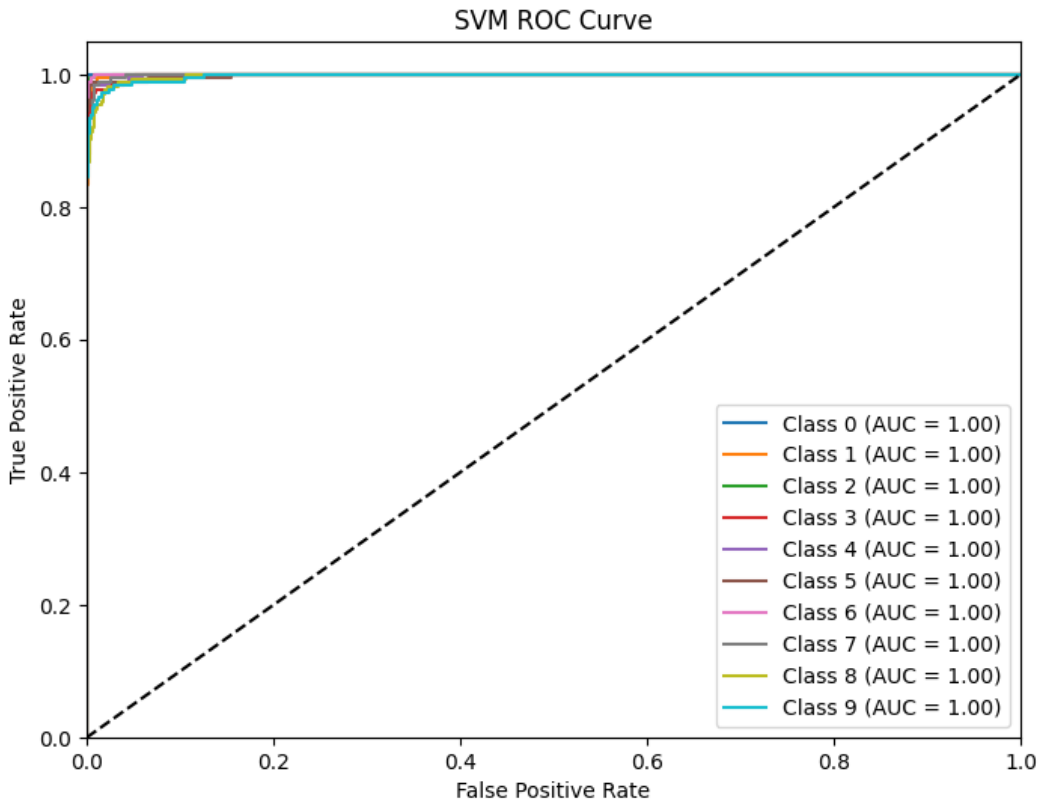
نتیجه گیری

- مدل KNN برتر در تمایز بین کلاس‌ها، به ویژه برای کلاس‌های ۰
- مدل KNN دارای دقت بالایی است، با AUC برابر با ۰٫۹۹ یا ۱٫۰۰ برای اکثر کلاس‌ها.
- تنها کلاس ۲، ۳، ۴ و ۸ دارای AUC کمی پایین‌تر هستند، اما هنوز نشان‌دهنده آن هستند

منظور از این جمله این است که کلاس‌های ۲، ۳، ۴ و ۸ دارای مقدار AUC (Area Under the Curve) کمی پایین‌تر از سایر کلاس‌ها هستند. اما هنوز این مقدار نشان‌دهنده روی مدل در تمایز بین این کلاس‌ها و سایر کلاس‌ها است.

به عبارت دیگر، هنوز کلاس‌های ۲، ۳،۴ و ۸ دارای AUC پایین‌تر از سایر کلاس‌ها هستند، اما این مقادیر بالا هستند و مدل‌های مدل را در تشخیص این کلاس‌ها از سایر کلاس‌ها نشان می‌دهند.

برای مثال، اگر AUC یک کلاس برابر با ۰/۹۹ باشد، این نشان‌دهنده مدل‌های مختلف در تشخیص آن کلاس است. اما اگر AUC یک کلاس برابر با 0.8 باشد، این نشان‌دهنده مدل پایین‌تر در تشخیص آن کلاس است. در این مورد، کلاس‌های ۲، ۳، ۴ و ۸ دارای AUC کمی پایین‌تر از سایر کلاس‌ها هستند، اما هنوز بالا هستند و مدل‌های بالای مدل را در تشخیص این کلاس‌ها نشان می‌دهند.



نمودار منحنی مشخصه عملکرد گیرنده (ROC) را برای طبقه‌بندی کننده ماشین بردار پشتیبانی (SVM) نشان می‌دهد. این منحنی برای 10 کلاس مختلف رسم

شده است که هر منحنی نشان دهنده عملکرد طبقه بندی کننده برای آن کلاس خاص است.

نمودار نشان می دهد که هر 10 کلاس دارای مساحت زیر منحنی 1.00 (AUC) هستند. این بدان معنی است که طبقه بندی کننده قادر است بین مثال های مثبت و منفی برای هر کلاس کاملاً تمایز قائل شود.

محور x نشان دهنده نرخ مثبت کاذب است، که نسبت نمونه های منفی است که به اشتباه به عنوان مثبت طبقه بندی شده اند. محور y نرخ مثبت واقعی را نشان می دهد، که نسبت نمونه های مثبتی است که به درستی به عنوان مثبت طبقه بندی شده اند.

یک طبقه‌بندی‌کننده کامل دارای AUC 1.00 است، به این معنی که همه نمونه‌ها را به درستی طبقه‌بندی می‌کند. هر چه AUC به 1.00 نزدیکتر باشد، عملکرد طبقه بندی کننده بهتر است.

در این حالت، تمام کلاس‌ها دارای AUC 1.00 هستند، که نشان می‌دهد طبقه‌بندی کننده SVM برای همه کلاس‌ها عالی عمل می‌کند. این احتمالاً به دلیل این واقعیت است که مجموعه داده بسیار قابل تفکیک است، به این معنی که تمایز واضحی بین کلاس‌های مختلف وجود دارد.

به طور کلی، این نمودار نشان می دهد که طبقه بندی کننده SVM در این مجموعه داده بسیار خوب عمل می کند. این می تواند به درستی تمام نمونه ها را برای هر کلاس طبقه بندی کند و به عملکرد عالی دست یابد.

به طور خلاصه:

کد یک تجزیه و تحلیل جامع از چهار مدل طبقه بندی مختلف در مجموعه داده ارقام ارائه می دهد.این مدل ها را با استفاده از دقت، زمان آموزش و منحنی های ROC ارزیابی می کند.

نتایج نشان می‌دهد که مدل‌های SVM و k-Nearest Neighbors دقیق‌ترین مدل‌ها هستند، اما زمان آموزش بالاتری نسبت به مدل‌های دیگر دارند.

این تجزیه و تحلیل می تواند به شما کمک کند تا بهترین مدل را برای مشکل طبقه بندی خاص خود بر اساس الزامات عملکرد و محدودیت های محاسباتی خود انتخاب کنید.