



پروژه
"یادگیری ماشین"
گروه دوم

نام استاد:

دکتر اعرابی-ابوالقاسمی

تهیه کننده :

۸۱۰۱۹۹۲۸۹

عرفان میرزایی

۸۱۰۱۹۹۲۳۹

حافظ قائمی



پاسخ به پرسش های مربوط به مقاله

۱- خلاصه ای از مقدمه

در مقدمه مقاله توضیحاتی درباره بیماری پارکینسون، علایم آن و روش های پزشکی تشخیص آن داده شده است که در زیر خلاصه ای از آن آورده شده است.

بر طبق گفته سازمان بهداشت جهانی (WHO)، بیماری های عصبی (مانند پارکینسون) از مهم ترین تهدید ها برای سلامت جوامع انسانی می باشد. بیماری پارکینسون یک بیماری عصبی می باشد که شاخصه اصلی آن کمبود تولید دوپامین در بخش میانی مغز (که ماده سیاه نامیده می شود) می باشد. دلیل این امر نیز تحلیل رفتن نورون های بخش ماده سیاه مغز می باشد که علت آن می تواند به سن، تاریخچه خانوادگی و حتی در معرض مواد آفت زدا قرار گرفتن مربوط باشد.

علایم این بیماری را می توان به دو دسته کلی علایم حرکتی و علایم غیر حرکتی تقسیم نمود. از علایم حرکتی می توان به لرزش های ناخواسته، تغییر در صحبت و تغییر در دست خط فرد اشاره کرد. علایم غیر حرکتی نیز شامل به هم خوردگی ساعت خواب، افسردگی و نگرانی می شود.

تشخیص بیماری پارکینسون در سه مرحله صورت می پذیرد؛ در مرحله اول علایم بیماری توسط معاینات بالینی تایید می شوند. در گام دوم به فرد مشکوک دارو داده می شود. اگر دارو در بهبود شخص تاثیر داشته باشد آنگاه به احتمال زیاد فرد به پارکینسون مبتلا شده است. در هر حال مرحله سومی برای موارد مشکوک وجود دارد؛ در این مرحله با انجام تست های آزمایشگاهی (که ممکن است برای همه مردم در دسترس نباشد) شک را بر طرف می کنند.

انجام مراحل درمانی بالا ممکن است برای یک فرد زمان بسیار زیادی به طول بینجامد. بنابراین لزوم وجود روش های سریع تر کاملاً احساس می شود. تست های بررسی دست خط و صدا از در دسترس ترین تست های موجود هستند.

در این مقاله از دیتای Vaiciukynas و همکاران همراه با ویژگی های صوتی یکسان استفاده شده است. نوآوری اصلی این مقاله ارزیابی عملکرد دیگر طبقه بند ها بدون ترکیب مجموعه های ویژگی ها می باشد.



۲- در تحقیق از چه مجموعه داده ای استفاده شده است؟

مجموعه داده این تحقیق از دو تسک صوتی تهیه شده است و به دو گونه متفاوت تقسیم شده است. ۱- فونتیک ها ۲- گفتار. این تسک صوتی توسط دو میکروفون متفاوت انجام شده است. ۱- Acoustic cardioid ۲- میکروفون گوشی هوشمند Samsung Galaxy Note 3. میکروفون در ۱۰ سانتی متری منبع صدا قرار گرفته است و همچنین فرمت صوتی که استفاده شده است mono PCM wav می باشد. گونه فونتیک شامل صدای حرف صدا دار a می باشد که در یک سطح بلندی صدا و فرکانس مناسب صدا ضبط شده است. مدت زمان این تسک پنج ثانیه می باشد و سه بار تکرار شده است. گونه گفتار شامل تلفظ یک عبارت به زبان لیتوانی می باشد که توسط یک فرد بومی این زبان انجام شده است. مجموعه داده شامل اطلاعات ۹۹ فرد (مرد و زن) می باشد. میانگین سن افراد سالم ۴۱,۸ و میانگین سن افراد مبتلا ۶۱,۵ می باشد.

۳- آیا داده های توسط خود نویسندگان جمع آوری شده است یا خیر؟

خیر از داده هایی که در مقاله Vaiciukynas و همکاران وجود داشته، استفاده شده است.

۴- از چه روش هایی برای پیش پردازش داده ها و انتخاب ویژگی ها استفاده شده است؟

سیگنال های صوتی توسط نرم افزار Praat Software Toolbox پیش پردازش شده اند. در چکیده مقاله گفته شده است که از ۱۸ الگوریتم استخراج ویژگی استفاده شده است اما نامی از روش هایی که از آن ها استفاده کرده اند برده نشده است.

۵- از چه مدل هایی برای طبقه بندی/خوشه بندی/درون یابی استفاده شده است؟

از روش های طبقه بندی kNN، MLP، OPF و SVM استفاده شده است.



۶- عملکرد مدل ها با چه اطلاعاتی گزارش شده است؟ آیا این گزارش دقیق است یا خیر؟ در صورتی که پاسخ منفی است، شیوه بهتری برای گزارش عملکرد مدل پیشنهاد دهید.

در این مقاله از تکنیک hold out برای محاسبه متریک های مختلف استفاده شده است و برای طبقه بند های مختلف، متریک های accuracy، sensitivity، specificity گزارش شده است. همچنین EER، AUC و زمان اجرای هر الگوریتم بر روی مجموعه های آموزش و تست آورده شده است. تست های آماری متفاوتی نیز برای ارزیابی نتایج نهایی صورت گرفته است. در نهایت نیز ماتریس confusion بهترین طبقه بند ها آورده شده است.

در این مقاله معیار های متنوعی در جداول مختلف گزارش شده اند و می توان گفت که گزارش بسیار دقیق و خوبی انجام شده است.

۷- نتیجه گیری و دستاورد های پژوهش

در این مطالعه روشی برای تشخیص بیماری پارکینسون با کمک ترکیب روش های استخراج مجموعه ویژگی و تکنیک های یادگیری ماشین ارائه شده است. در این مقاله از ۱۸ روش استخراج ویژگی و ۱۴ الگوریتم طبقه بند استفاده شده و نتایج آن ها بر اساس معیار های EER و accuracy ارزیابی شده است. نتایج آزمون های آماری نشان می دهد که تفاوت چشم گیری در استفاده از نوع میکروفون وجود ندارد؛ بنابراین این باور تقویت می شود که می توان از میکروفون تلفن های همراه برای تشخیص بیماری پارکینسون استفاده نمود. صدا های فونتیکی نسبت به صداهای گفتاری، بخش صدا دار و بخش بی صدا نتایج بسیار بهتری به همراه داشتند. در گونه دیتا های فونتیک که با میکروفون audio cardioid ضبط شده اند، مجموعه ویژگی Yaffe و طبقه بند 1-nearest neighbor و در گونه دیتاهای فونتیک که با میکروفون گوشی هوشمند ضبط شده اند، مجموعه ویژگی KTU و طبقه بند 1-nearest neighbor بهترین نتیجه را داشته است. نتایج بدست آمده نشان می دهد که روش به کار گرفته شده در این مطالعه نتایج نسبتا بهتری نسبت به روش Vaiciukynas داشته است. اهمیت بالینی این روش در آن است که این امکان را برای پزشک فراهم می کند تا با استفاده از برنامه نصب شده بر روی گوشی همراه، بیماری پارکینسون را تشخیص دهد.



پیش از آنکه به توضیح روش های پیش پردازش مورد استفاده در پروژه پردازیم، توضیحاتی در مورد معیار های اندازه گیری عملکرد مدل ها بحث می کنیم.

در این مسئله با توجه به اینکه تمامی ویژگی های موجود در مجموعه داده به صورت عددی است و هم چنین تعداد داده های مربوط به هر کلاس برابر نیست و در واقع مجموعه داده نامتعادل (Unbalanced) است نمی توانیم تنها از معیاری مانند "دقت" استفاده کنیم و باید در کنار آن از معیار دیگری مانند ماتریس در هم ریختگی و "F1_Score" نیز استفاده کنیم که این کار باعث می شود هم "Recall" و هم "Precision" را در نظر بگیریم. البته در یک نگاه دیگر با توجه به اینکه این طبقه بندی مربوط به تشخیص بیماری پارکینسون است می توان اهمیت "Recall" را بالاتر در نظر گرفت و به آن توجه بیشتری کرد. از طرف دیگر، به عنوان معیار دیگر برای سنجش عملکرد طبقه بند که میزان تمایزپذیری ویژگی ها رو نیز مشخص کند می توانیم از "AUC" استفاده کنیم که در واقع همان سطح زیر منحنی ROC می باشد.

برای بدست آوردن شهودی از مجموعه داده ها ابتدا نگاهی به مجموعه داده ها می اندازیم. مجموعه داده ها شامل ۷۵۶ نمونه است که هر کدام از نمونه ها دارای ۷۵۵ ویژگی هستند که از بین این ۷۵۶ نمونه، ۵۶۴ نمونه مربوط به کلاس "۱" یعنی بیماران مبتلا به پارکینسون و باقی مربوط به کلاس "۰" یعنی افراد عادی اختصاص دارد. در ادامه ستون "id" را از مجموعه داده ها حذف می کنیم و مجموعه داده ها را در ابتدا به دو قسمت آموزش و آزمون به نسبت ۴ به ۱ تقسیم می کنیم. در تمامی مراحل این پروژه صرفاً از مجموعه آموزش برای بهینه سازی پارامترهای مدل و هم چنین پارامترهای ساختاری استفاده کردیم و از مجموعه آزمون تنها یک بار و در انتها صرفاً برای مقایسه بین مدل های آموزش دیده استفاده شده است. برای Cross-Validation در تمامی مدل ها از روش k-fold با $k = 5$ استفاده شده است و در تمامی مدل ها برای سه معیار "دقت"، "F1_Score" و "AUC" جداگانه این کار صورت گرفته و نتایج حاصل از آن بررسی شده است.

روش های پیش پردازش و انتخاب ویژگی

برای یک سازی داده ها با استفاده از کتابخانه Scikit-Learn و با استفاده از دستور StandardScaler() داده ها را به فرم نرمال استاندارد تبدیل می کنیم.

برای روش های انتخاب ویژگی از روش ICA استفاده نمی کنیم؛ زیرا این روش بیشتر برای مواردی استفاده می شود که دنبال پیدا کردن منبع های مستقل تولید داده ها باشیم استفاده می شود.



هم چنین از روش LDA نیز استفاده نکردیم؛ زیرا با استفاده از آن باید داده ها را به یک ویژگی تبدیل کنیم که نمی تواند جدایی پذیری مورد نظر را حاص کند تنها برای نشان دادن درستی این فرض در قسمت SVM استفاده کردیم که در ادامه ی گزارش خواهد آمد.

دو روش اصلی مورد استفاده در این پروژه استفاده از کاهش بعد با استفاده از PCA با و بدون استفاده از Whitening است که در هر قسمت توضیحات مربوط به آن آمده است.

برای حل مشکل نامتعادل بودن مجموعه داده از روش SMOTE داده ی مصنوعی برای مجموعه آموزش درست کردیم و بنابراین در مجموعه آموزش تعداد داده های هر دسته برابر است ولی در مجموعه آزمون همون مجموعه اصلی و بدون تغییر است.

روش های طبقه بندی

۱.۱. طبقه بندهای Generative :

• طبقه بندی بهینه بیز با استفاده از تخمین Parzen

در طراحی این مدل با استفاده از دستور Pipeline و هم چنین GridSearchCV کتابخانه ی Scikit-Learn استفاده کردیم که در سه مرحله استاندارد سازی، کاهش بعد و طراحی طبقه بند را به صورت موازی انجام دادیم و در مورد بهترین پارامتر های آن سرچ انجام دادیم این کار به این علت انجام شده است که دقیقا داده هایی که در مرحله برای folding استفاده میشود در تمامی این مراحل یکسان باشد که این کار از نشئت اطلاعات جلوگیری می کند.

طبقه بندی های Generative ، معیارهای "دقت" ، "F1_Score" و "AUC" هر سه نتایج یکسانی را رقم می زدند و حاصل نتیجه به صورت زیر است:

الف) بدون استفاده از Whitening

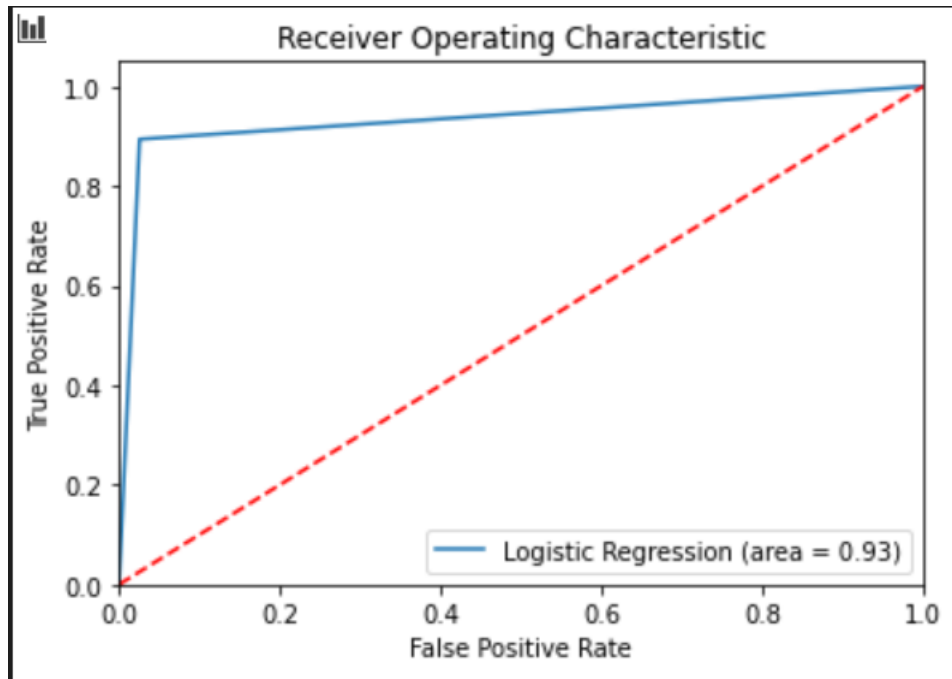
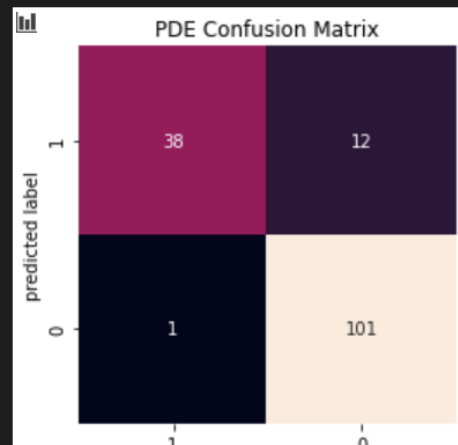
Model			
exponential	0.336	ACC	0.939
exponential	0.336	F1	0.934
exponential	0.336	AUC	0.939



پروژه یادگیری ماشین
عرفان میرزایی-حافظ قائمی
۸۱۰۹۹۲۳۹_۸۱۰۱۹۹۲۸۹



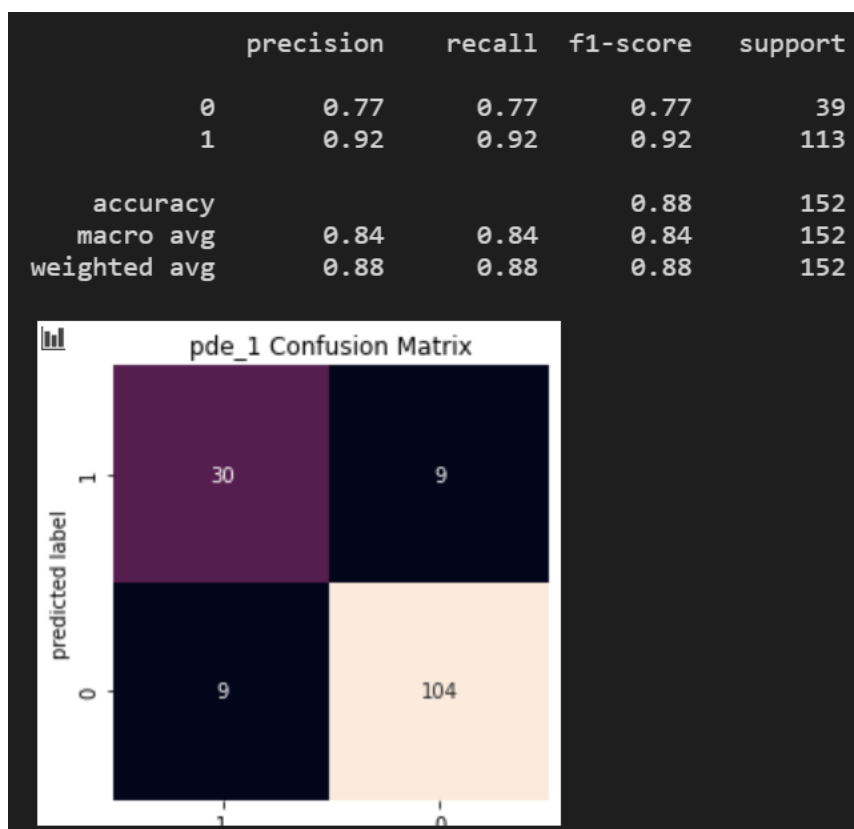
	precision	recall	f1-score	support
0	0.76	0.97	0.85	39
1	0.99	0.89	0.94	113
accuracy			0.91	152
macro avg	0.88	0.93	0.90	152
weighted avg	0.93	0.91	0.92	152

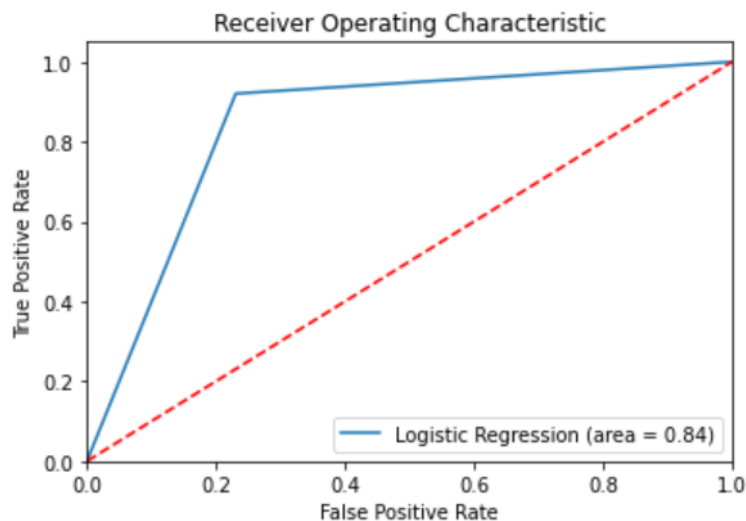




(ب) با استفاده از Whitening

Model			
exponential	0.336	ACC	0.935
exponential	0.336	F1	0.931
exponential	0.336	AUC	0.935





که همان طور که مشاهده می شود دقت این طبقه بند در حالت بدون استفاده از Whitening عملکرد بهتری دارد.

- طبقه بندی بهینه بیز با استفاده از تخمین KNN

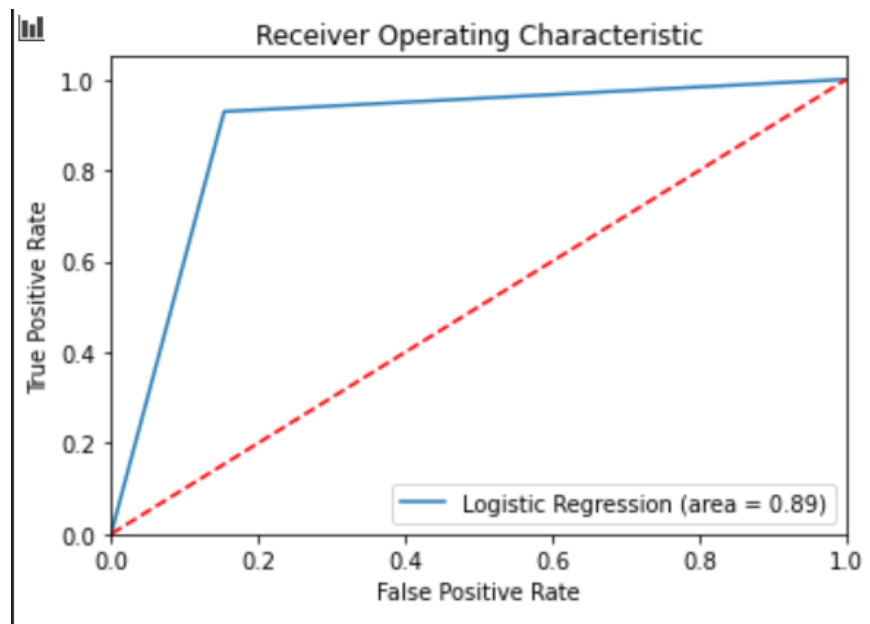
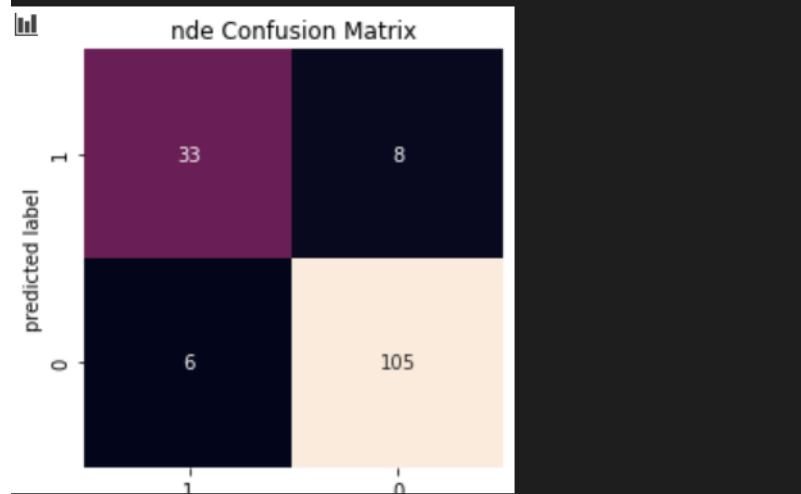
طبقه بندی های **Generative**، معیارهای "دقت"، "F1_Score" و "AUC" هر سه نتایج یکسانی را رقم می زدند و حاصل نتیجه به صورت زیر است:

الف) بدون استفاده از Whitening

Model(Components)			
1	1	ACC	0.942
1	1	F1	0.939
1	1	AUC	0.942



	precision	recall	f1-score	support
0	0.80	0.85	0.83	39
1	0.95	0.93	0.94	113
accuracy			0.91	152
macro avg	0.88	0.89	0.88	152
weighted avg	0.91	0.91	0.91	152

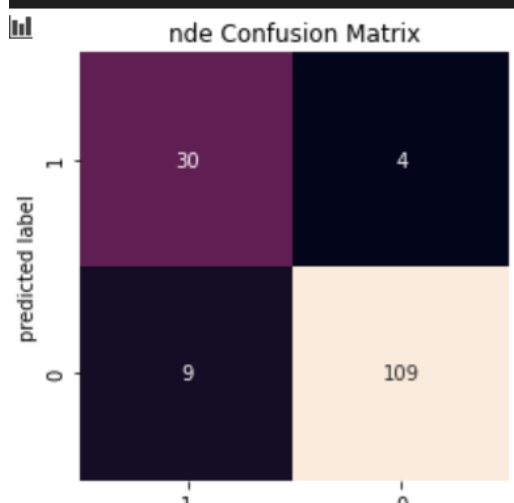


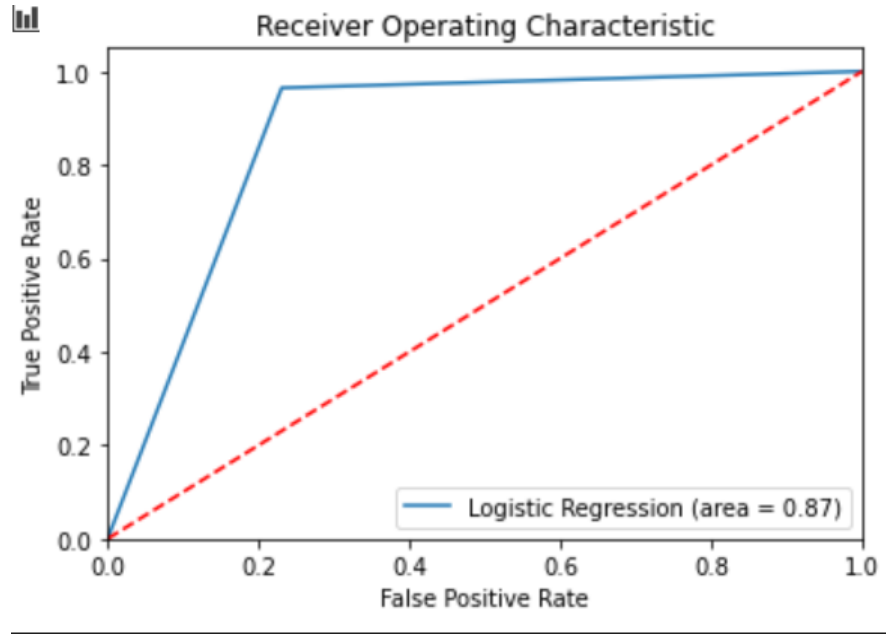


ب) با استفاده از Whitening

Model(Components)			
2	1	ACC	0.938
2	1	F1	0.936
2	1	AUC	0.938

	precision	recall	f1-score	support
0	0.88	0.77	0.82	39
1	0.92	0.96	0.94	113
accuracy			0.91	152
macro avg	0.90	0.87	0.88	152
weighted avg	0.91	0.91	0.91	152





که همان طور که مشاهده می شود دقت این طبقه بند در حالت بدون استفاده از Whitening عملکرد بهتری دارد.

- طبقه بندی بهینه بیز با استفاده از تخمین GMM

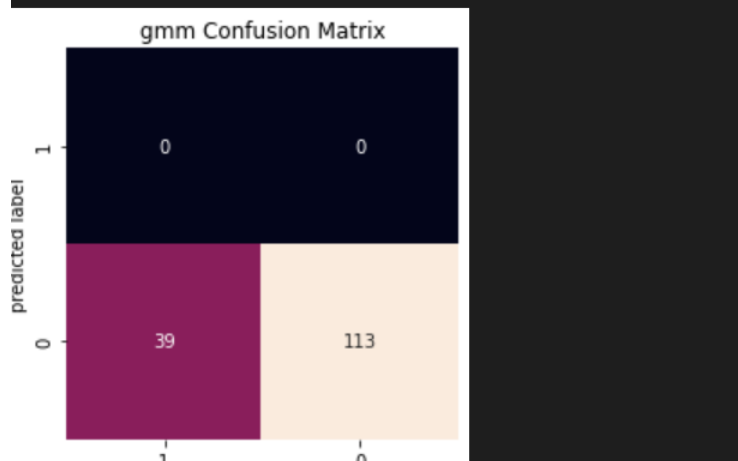
طبقه بندی های **Generative**، معیارهای "دقت"، "F1_Score" و "AUC" هر سه نتایج یکسانی را رقم می زدند و حاصل نتیجه به صورت زیر است:

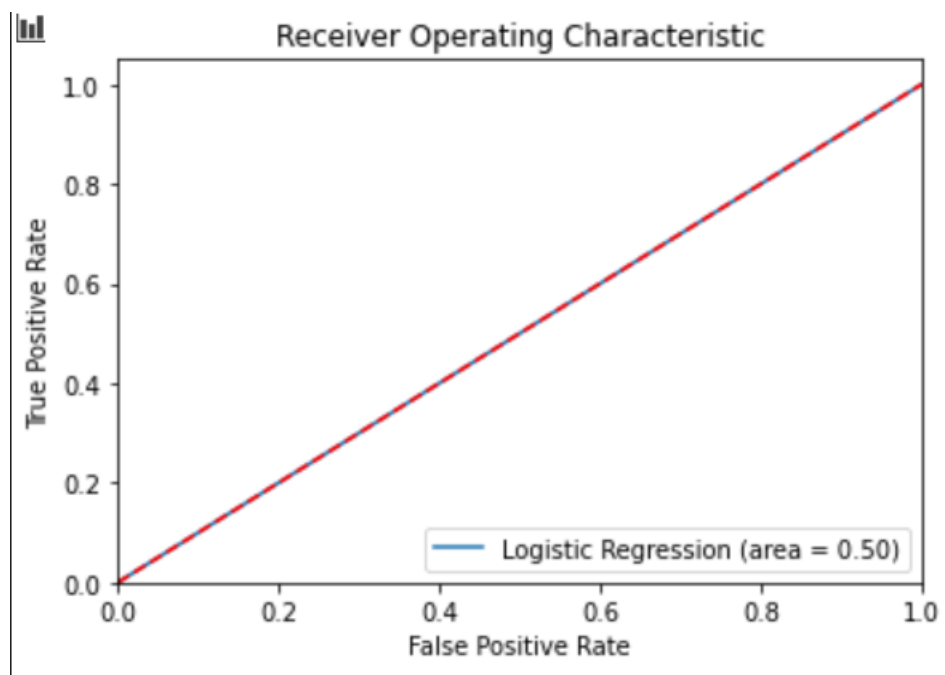
الف) بدون استفاده از Whitening

Model(Components)			
2	3	ACC	0.96
2	3	F1	0.961
2	3	AUC	0.96



	precision	recall	f1-score	support
0	0.00	0.00	0.00	39
1	0.74	1.00	0.85	113
accuracy			0.74	152
macro avg	0.37	0.50	0.43	152
weighted avg	0.55	0.74	0.63	152





Whitening (ب) با استفاده از

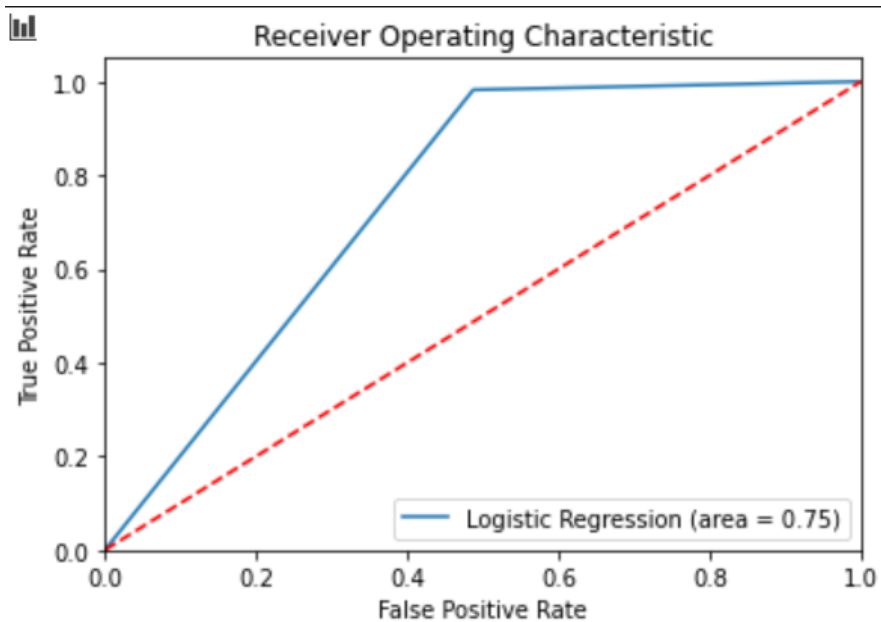
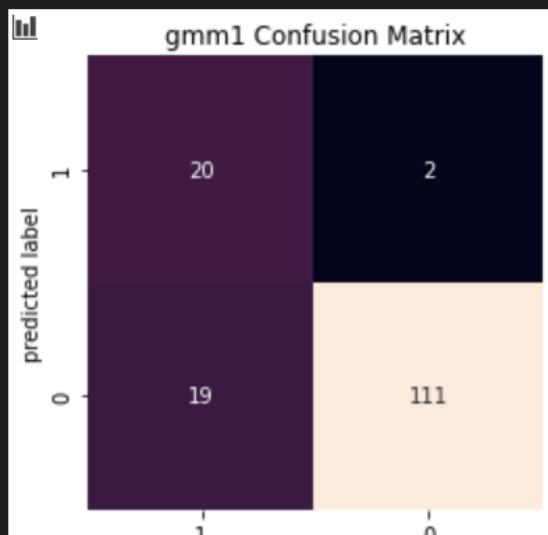
Model(Components)			
2	1	ACC	0.95
2	1	F1	0.952
2	1	AUC	0.95



پروژه یادگیری ماشین
عرفان میرزایی-حافظ قائمی
۸۱۰۹۹۲۳۹_۸۱۰۱۹۹۲۸۹



	precision	recall	f1-score	support
0	0.91	0.51	0.66	39
1	0.85	0.98	0.91	113
accuracy			0.86	152
macro avg	0.88	0.75	0.78	152
weighted avg	0.87	0.86	0.85	152





که همان طور که مشاهده می شود دقت این طبقه بند در حالت با استفاده از Whitening عملکرد بهتری دارد.

۱.۲. طبقه بندهای Discriminative :

• Logistic regression

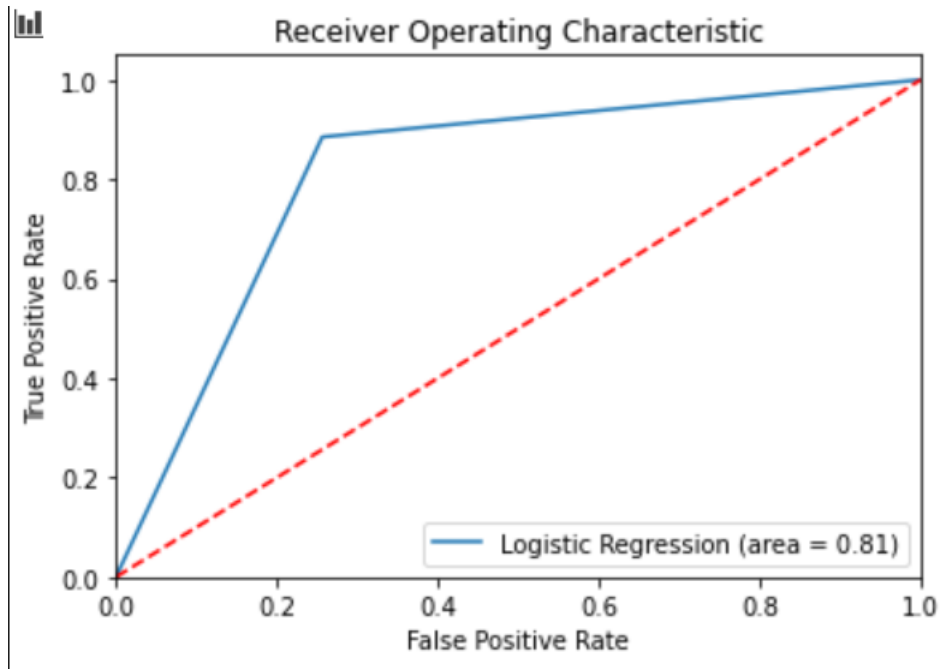
طبقه بندی های Discriminative ، معیارهای "دقت" ، "F1_Score" و "AUC" هر سه نتایج تقریباً یکسانی را رقم می زدند که از آوردن تک تک نتایج پرهیز شده (در کد موجود است) و اما حاصل نتیجه به صورت زیر است:

best parameters of the model are:

```
{'Log_Reg__C': 1, 'Log_Reg__penalty': 'l2', 'Log_Reg__solver': 'lbfgs', 'pca__n_components': 0.995, 'pca__whiten': False}
```

Mean accuracy of the model with 5-fold is: 0.9168999386126458





SVM •

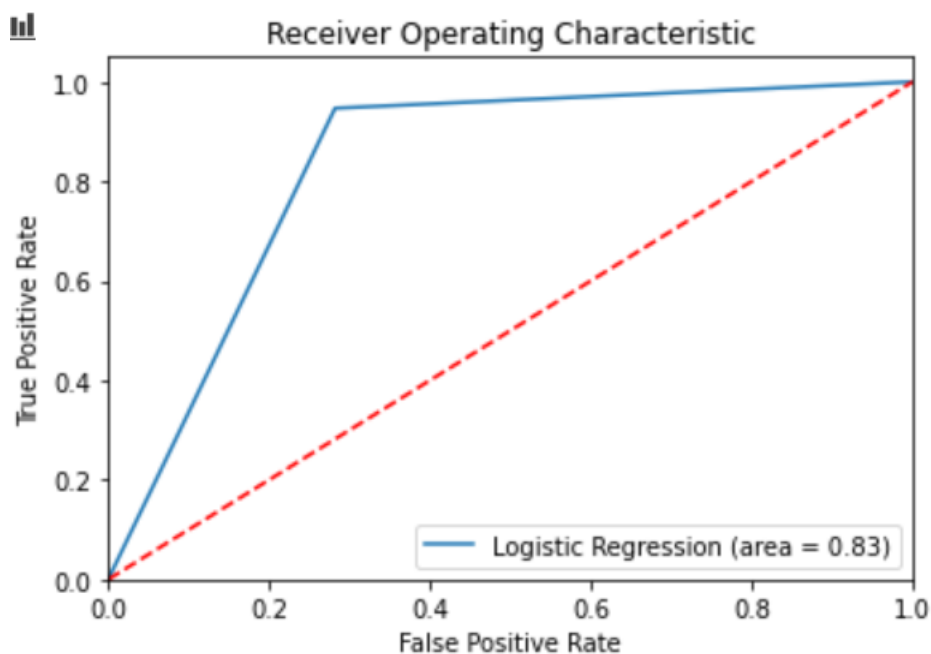
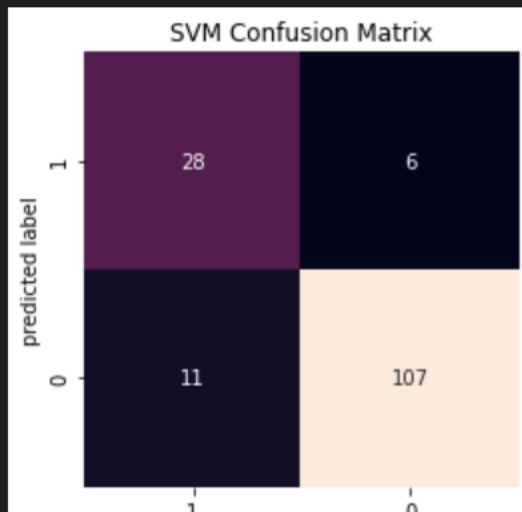
طبقه بندی های **Discriminative**، معیارهای "دقت"، "F1_Score" و "AUC" هر سه نتایج تقریباً یکسانی را رقم می زدند که از آوردن تک تک نتایج پرهیز شده (در کد موجود است) و اما حاصل نتیجه به صورت زیر است:

best parameters of the model are: {'pca__n_components': 0.999, 'pca__whiten': False, 'svm__C': 10, 'svm__kernel': 'rbf'}

Mean accuracy of the model with 5-fold is: 0.9634561080417434



	precision	recall	f1-score	support
0	0.82	0.72	0.77	39
1	0.91	0.95	0.93	113
accuracy			0.89	152
macro avg	0.87	0.83	0.85	152
weighted avg	0.89	0.89	0.89	152



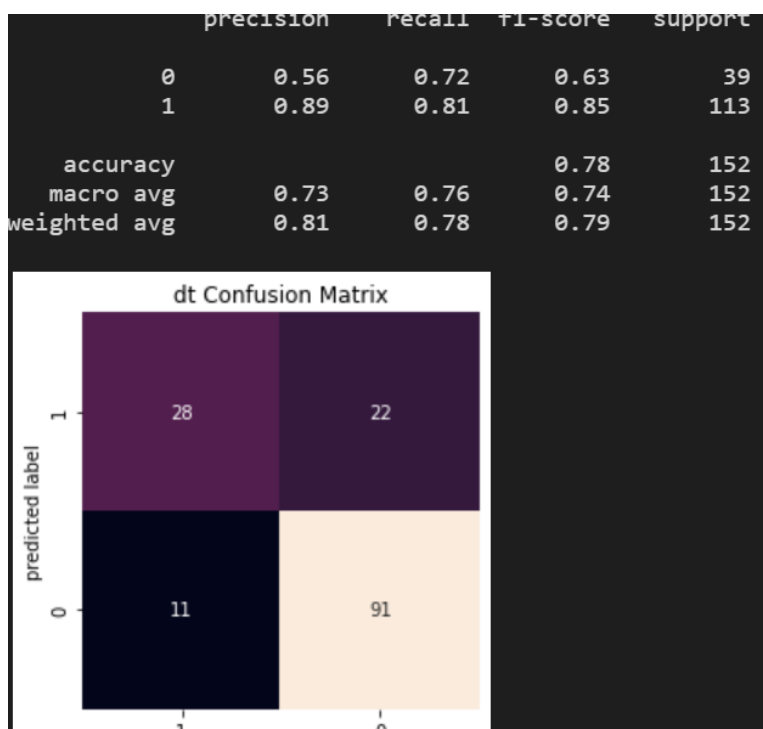


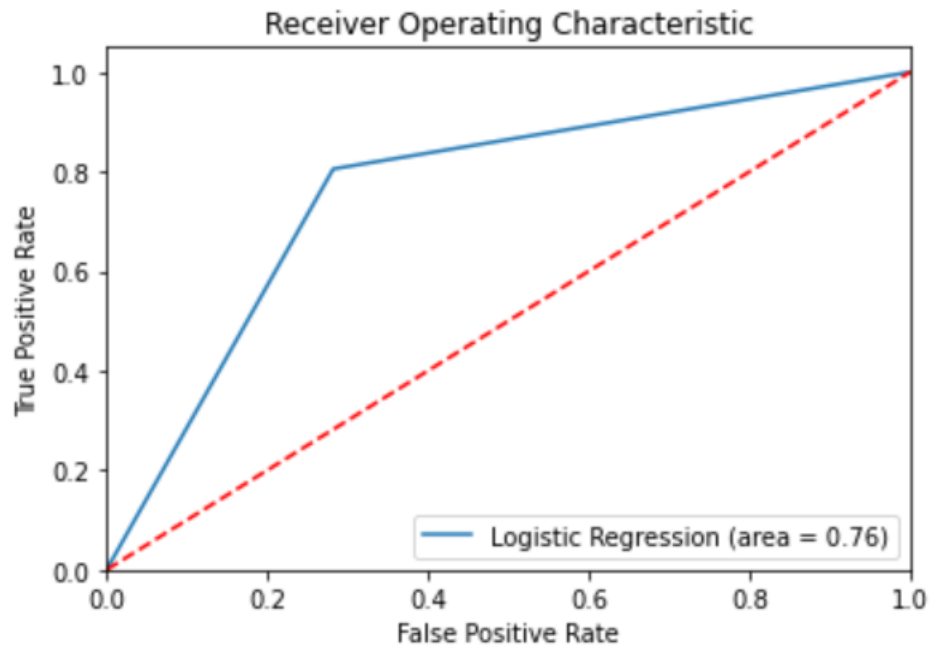
• Decision Tree

طبقه بندی های **Discriminative**، معیارهای "دقت"، "F1_Score" و "AUC" هر سه نتایج تقریباً یکسانی را رقم می زدند که از آوردن تک تک نتایج پرهیز شده (در کد موجود است) و اما حاصل نتیجه به صورت زیر است:

best parameters of the model are: {'dt__max_depth': 12, 'dt__min_samples_split': 2, 'pca__n_components': 0.8, 'pca__whiten': False}

Mean accuracy of the model with 5-fold is: 0.8448127685696747





KNN •

طبقه بندی های **Discriminative**، معیارهای "دقت"، "F1_Score" و "AUC" هر سه نتایج تقریباً یکسانی را رقم می زدند که از آوردن تک تک نتایج پرهیز شده (در کد موجود است) و اما حاصل نتیجه به صورت زیر است:

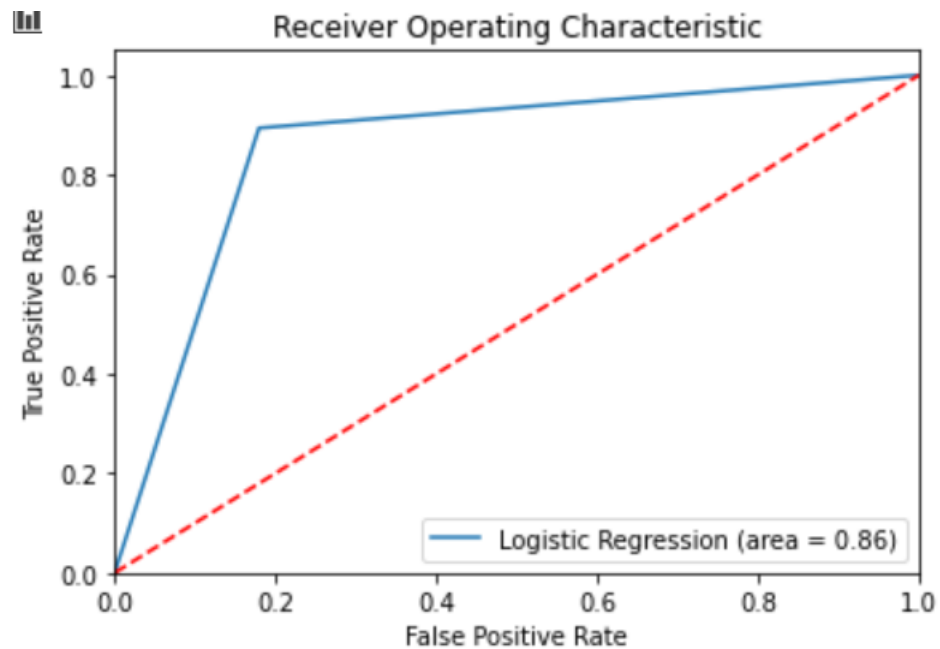
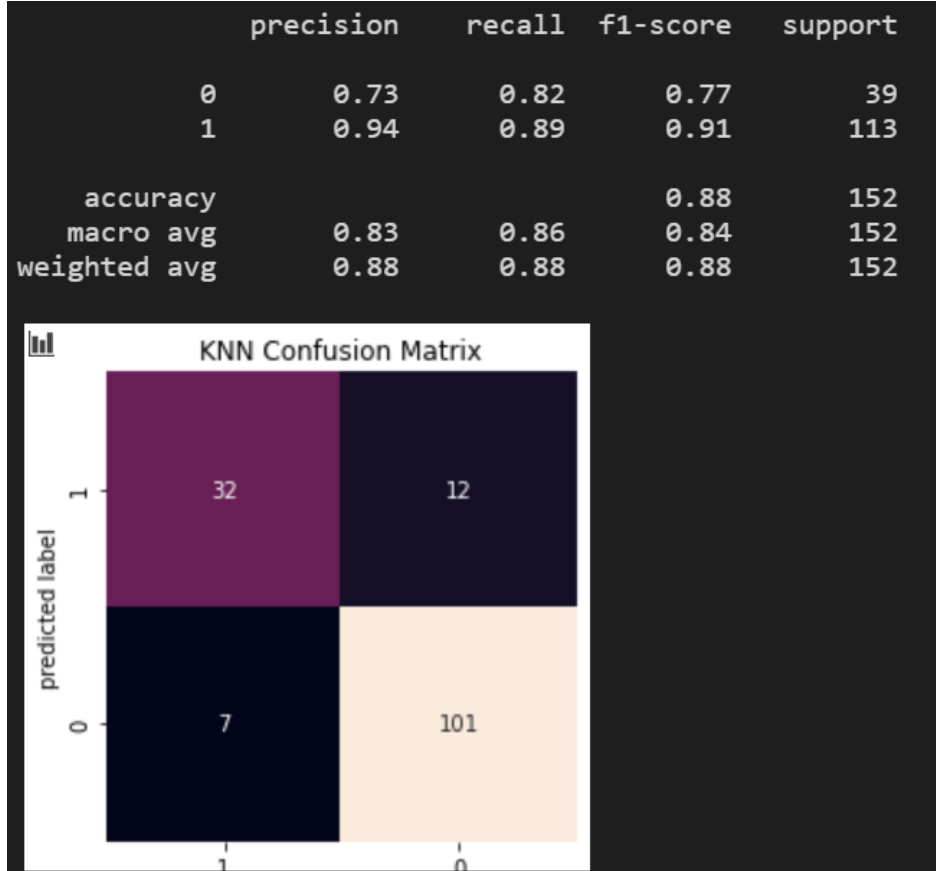
best parameters of the model are:

```
{'knn__n_neighbors': 1, 'pca__n_components': 0.85, 'pca__whiten': True}
```

Mean accuracy of the model with 5-fold is: 0.9390546347452424



پروژه یادگیری ماشین
عرفان میرزایی-حافظ قائمی
۸۱۰۹۹۲۳۹_۸۱۰۱۹۹۲۸۹





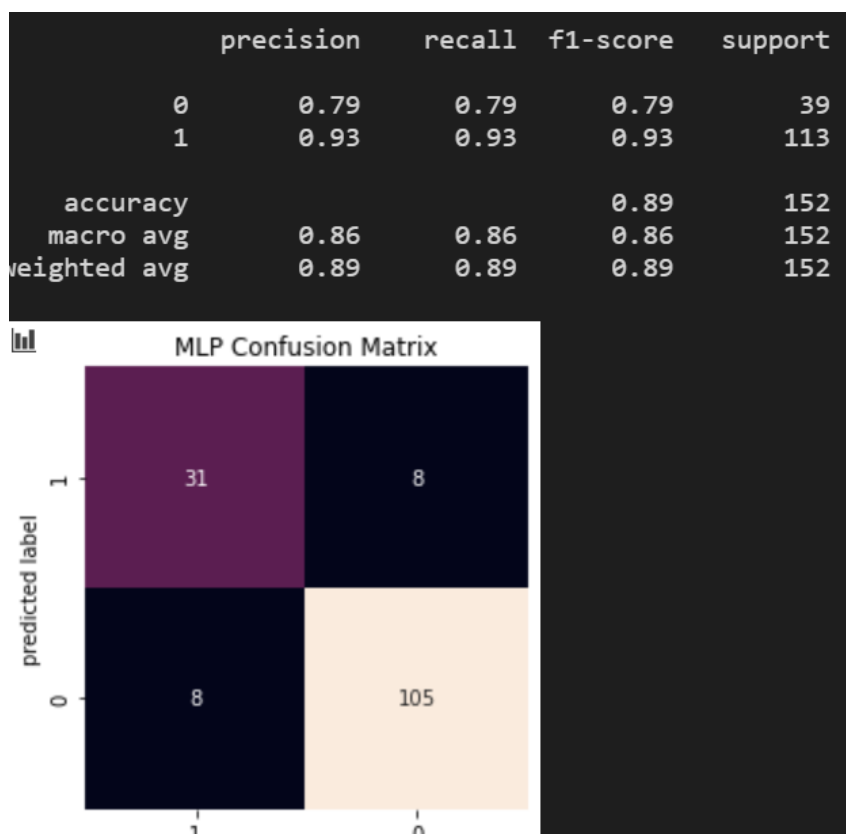
MLP •

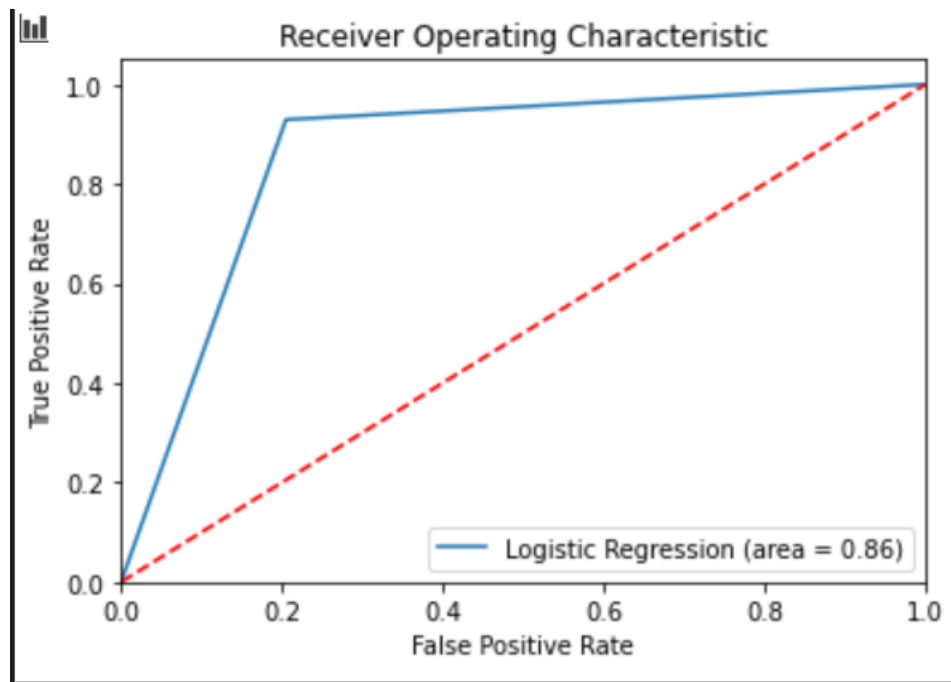
طبقه بندی های **Discriminative**، معیارهای "دقت"، "F1_Score" و "AUC" هر سه نتایج تقریباً یکسانی را رقم می زدند که از آوردن تک تک نتایج پرهیز شده (در کد موجود است) و اما حاصل نتیجه به صورت زیر است:

best parameters of the model are:

```
{'mlp__alpha': 0.0001, 'mlp__batch_size': 256, 'mlp__hidden_layer_sizes': 200, 'mlp__learning_rate': 'invscaling', 'mlp__learning_rate_init': 0.001, 'mlp__max_iter': 100, 'mlp__momentum': 0.9, 'mlp__solver': 'adam', 'pca__n_components': 0.9, 'pca__whiten': False}
```

Mean accuracy of the model with 5-fold is: 0.9556783302639656





۱.۳. مدل نهایی

در میان طبقه بند های **Generative**، طبقه بندی بهینه بیز با استفاده از تخمین Parzen بهترین طبقه بندی را کسب کرده است. دقت این طبقه بندی بر روی داده های تست ۹۱ درصد است و هم چنین F_Score آن برابر ۰,۹۲ است و AUC آن برابر ۰,۹۳ است.

هم چنین طبقه بندی بهینه بیز با استفاده از تخمین KNN دارای دقتی بر روی داده های تست برابر با ۹۱ درصد است و هم چنین F_Score آن برابر ۰,۹۲ است و AUC آن برابر ۰,۸۹ است.

در میان طبقه بند های **Discriminative** طبقه بند MLP بهترین عملکرد را در طبقه بندی کسب کرده است. دقت این طبقه بندی بر روی داده های تست ۸۹ درصد است و هم چنین F_Score آن برابر ۰,۸۹ است و AUC آن برابر ۰,۸۶ است.

و هم چنین طبقه بند SVM با دقتی بر روی داده های تست برابر ۸۹ درصد است و هم چنین F_Score آن برابر ۰,۸۹ است و AUC آن برابر ۰,۸۳ است.

که بهترین عملکرد مربوط به طبقه بندی بهینه بیز با استفاده از تخمین Parzen است.



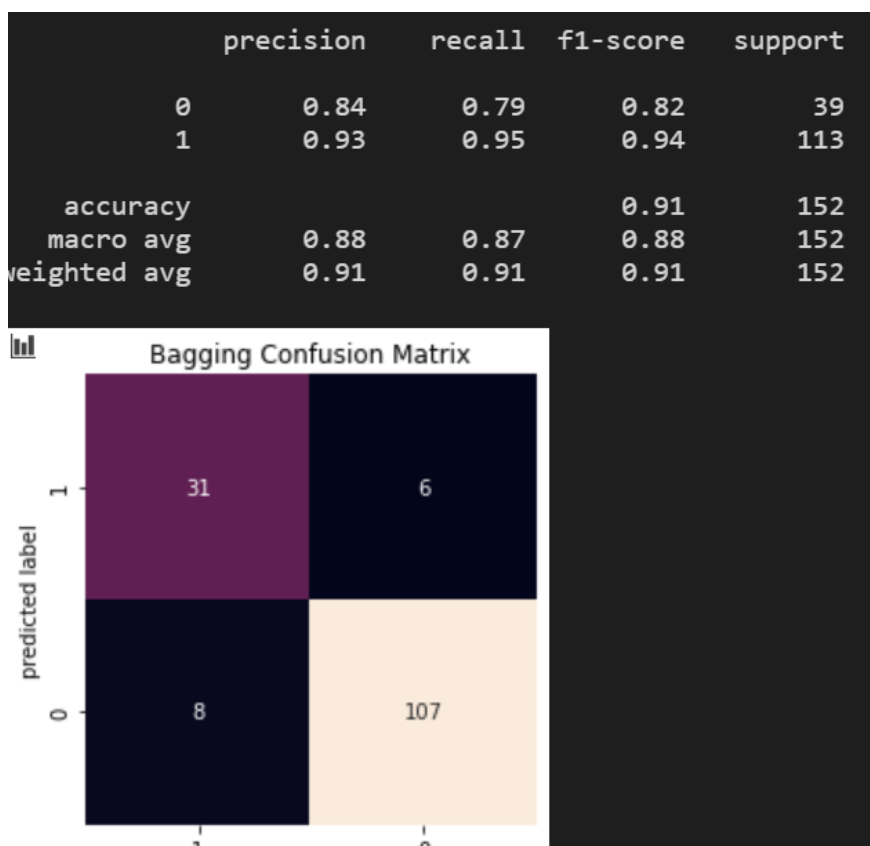
به صورت کلی عملکرد طبقه بند های **Generative**، بهتر بوده است که با توجه به اینکه داده های مصنوعی تولید شده و در نظر گرفتن دانش پیشین این امر توجیه پذیر است.

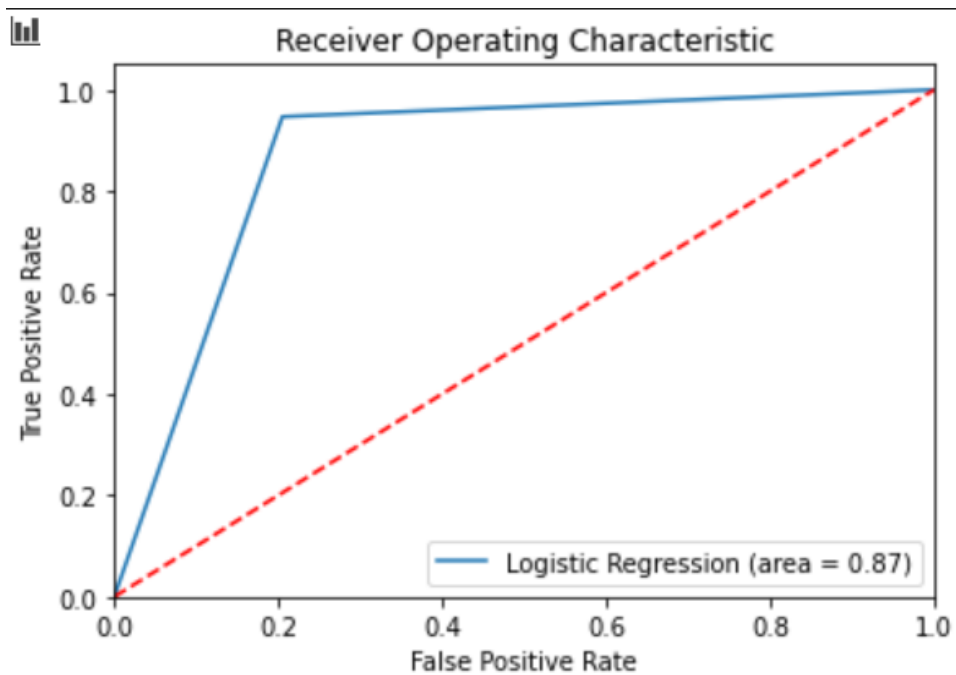
۱.۴. روش های تجمیعی

best parameters of the model are:

```
{'b__max_features': 0.5, 'b__n_estimators': 100,  
'pca__n_components': 0.8, 'pca__whiten': False}
```

Mean accuracy of the model with 5-fold is: 0.9301780233271947





که با مقایسه عملکرد این روش ها با روش های قبلی مشاهده می شود که عملکرد آن بهبود پیدا کرده است.