عنوان پایان‌نامه‌ی کارشناسی ارشد

ارایه‌ی یک روش بهبود یافته برای پیش‌بینی ضرورت بستری‌شدن بیماران کووید ۱۹ در بخش مراقبت‌های ویژه با استفاده از تکنیک‌های ترکیبی داده‌کاوی

دانشجو: مهنام پدرام

دانشکده: مکانیک، برق و کامپیوتر

گروه تخصصی: مهندسی نرم‌افزار

استاد راهنما: خانم دکتر مریم رستگارپور

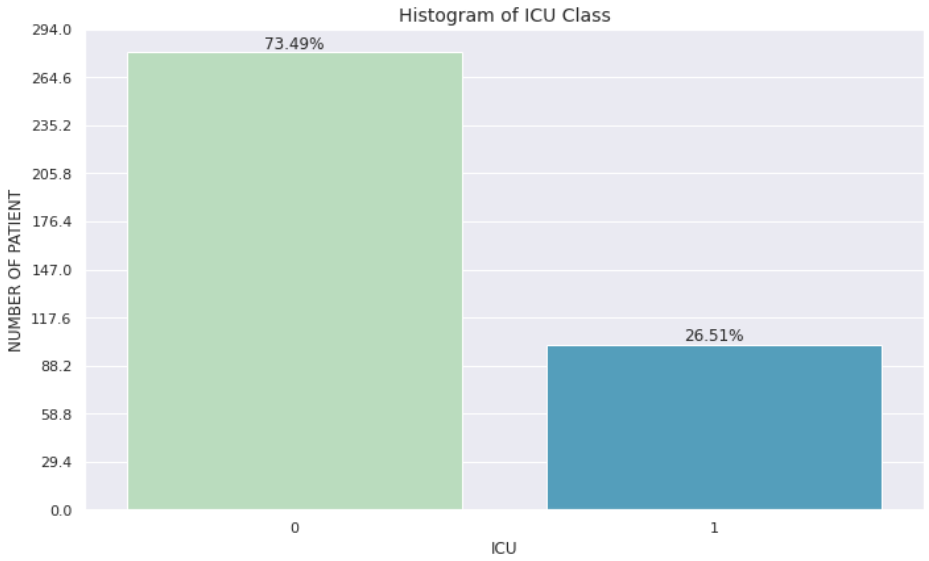
گزارش پیشرفت: شماره ۴ –۱۴۰۱/۶/۲۸

۱– مقدمه:

با توجه به اینکه نتایج گزارش شده در گزارشات قبلی قابل مقایسه با مقادیر State of the art گزارش شده در سایر مقالات نبود و همچنین روش های مختلف جایگذاری نمونه های خالی و نمونه افزایی چندان تغییری در نتایج ایجاد نکرد، در این گزارش، بررسی دقیق تری روی مجموعه دادگان و معانی فیزیولوژیک ویژگی های ثبت شده انجام شده است. همانگونه که در ادامه توضیح داده شده است با اعمال چند تغییر کوچک در کدهای قبلی بهبود قابل ملاحظه ای در نتایج دیده می شود. به علاوه در این گزارش، عملکرد مدل XGBoost نیز بررسی شده و همچنین معیار مساحت زیر منحنی[1] نیز محاسبه و نشان داده شده است.

۲– حذف نمونه های تکراری:

به دنبال بازنگری ماتریس ویژگی ها، ویژگی غیر عددی 'PATIENT_VISIT_IDENTIFIER' که به نوعی نشان دهنده ی تعداد دفعات مراجعه ی یک بیمار به بیمارستان است از ادامه ی محاسبات حذف شد. برای جایگذاری داده‌های ثبت نشده نیز از ترکیب روش‌های جلوسو و بازگشتی استفاده شد. یکی از نکاتی که در گزارشات قبلی در نظر گرفته نشده بود، تعداد ردیف های تکراری و حذف آنها بود. در ارزیابی‌های اخیر، متوجه شدیم که ۲۴۳ ردیف تکراری در دادگان وجود داشت که حذف آنها توزیع دو کلاس را به صورت زیر به دنبال دارد که در آن ۵۱/۲۶٪ نمونه‌ها به هرحال به ICU انتقال داده شده‌اند.

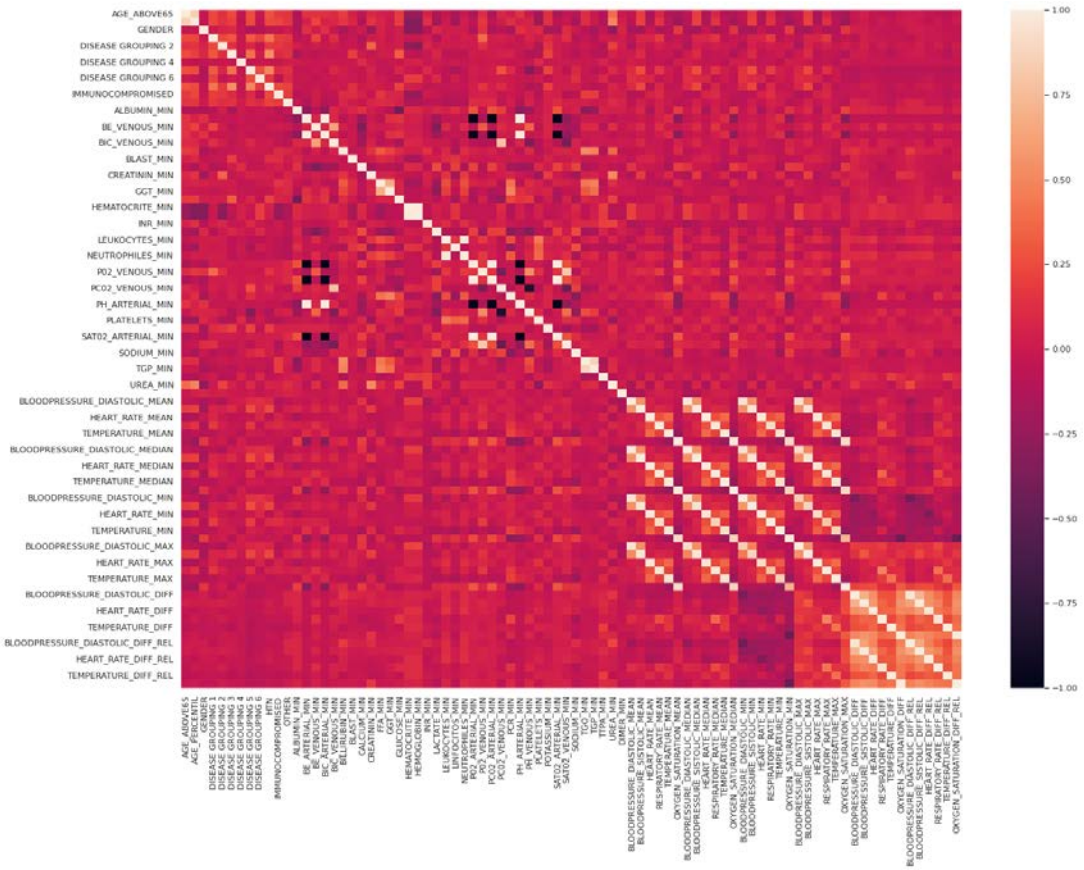


شکل ۱ توزیع افراد منتقل شده به **ICU** را براساس جنسیت بیماران نشان می‌دهد.

۳–بررسی همبستگی ویژگی های مختلف:

برای ایجاد درک بهتری نسبت به همبستگی ویژگی‌های ثبت شده، ماتریس همبستگی این ویژگی‌ها به روش Pearson Correlation محاسبه و در شکل ۲ نشان داده شده است. مرتب‌سازی اعداد همبستگی به ترتیب نزولی نشان می‌دهد که مقادیر کمینه‌ی ثبت شده برای ویژگی فیزیولوژیک مرتبط با فشار خون و ... بیشترین همبستگی را با یکدیگر دارند (جدول ۱).

[1] Area under the curve (AUC)



شکل ۲: ماتریس همبستگی متغیرهای فیزیولوژیک ثبت شده پس از نرمالیزه شدن در بازه [۱ ۱-].

جدول ۱: فهرست ویژگی‌های به ترتیب همبستگی بر اساس شاخص پیرسون

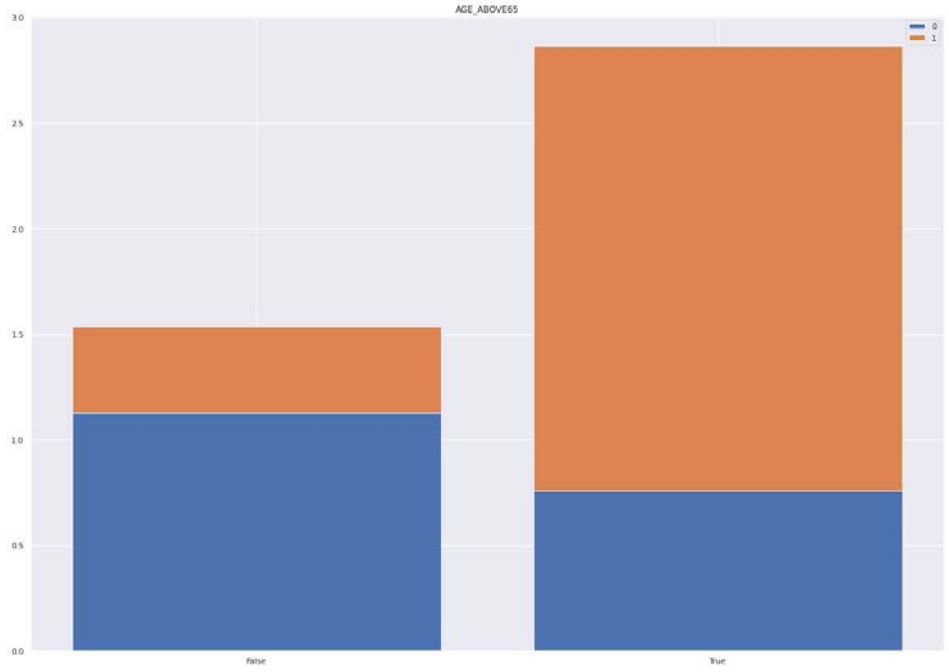
	Feature_1	Feature_2	Pearson Correlation
1128	BE_ARTERIAL_MIN	PH_ARTERIAL_MIN	1
1300	BIC_ARTERIAL_MIN	SAT02_ARTERIAL_MIN	1
1125	BE_ARTERIAL_MIN	PC02_ARTERIAL_MIN	1
1291	BIC_ARTERIAL_MIN	P02_ARTERIAL_MIN	1
1123	BE_ARTERIAL_MIN	P02_ARTERIAL_MIN	1
...	...	...	...
546	DISEASE GROUPING 4	SODIUM_MIN	0.00028
1398	BIC_VENOUS_MIN	BLOODPRESSURE_DIASTOLIC_MEDIAN	0.000155
1362	BIC_VENOUS_MIN	BLAST_MIN	0.000153
520	DISEASE GROUPING 4	BIC_VENOUS_MIN	0.000113
47	AGE_ABOVE65	DIMER_MIN	0.000008
3486 rows × 3 columns			

با توجه به معنی آماری Pearson correlation، می‌توان آن دسته از ویژگی‌هایی که ضریب همبستگی بالاتر از ۹۹٪ دارند را حذف نمود که این اطلاعات بعداً (در صورت رضایتبخش نبودن نتایج) در مرحله‌ی کاهش تعداد ویژگی‌ها به کار گرفته خواهد شد. فهرست کامل این ویژگی‌ها در جدول ۲ نشان داده شده است.

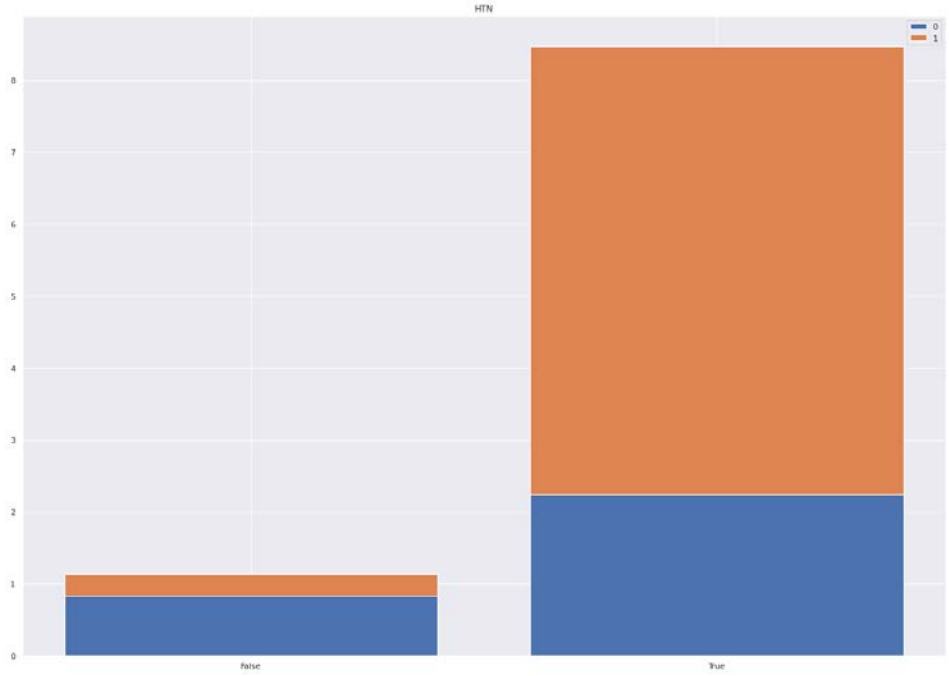
جدول ۲: فهرست کامل ویژگی‌های با ضریب همبستگی بالاتر از ٪۹۹

	Feature_1	Feature_2	Pearson Correlation
1128	BE_ARTERIAL_MIN	PH_ARTERIAL_MIN	1
1300	BIC_ARTERIAL_MIN	SAT02_ARTERIAL_MIN	1
1125	BE_ARTERIAL_MIN	PC02_ARTERIAL_MIN	1
1291	BIC_ARTERIAL_MIN	P02_ARTERIAL_MIN	1
1123	BE_ARTERIAL_MIN	P02_ARTERIAL_MIN	1
2812	PC02_ARTERIAL_MIN	SAT02_ARTERIAL_MIN	1
1132	BE_ARTERIAL_MIN	SAT02_ARTERIAL_MIN	1
2640	P02_ARTERIAL_MIN	PH_ARTERIAL_MIN	1
1296	BIC_ARTERIAL_MIN	PH_ARTERIAL_MIN	1
1107	BE_ARTERIAL_MIN	BIC_ARTERIAL_MIN	1
1293	BIC_ARTERIAL_MIN	PC02_ARTERIAL_MIN	1
2644	P02_ARTERIAL_MIN	SAT02_ARTERIAL_MIN	1
2808	PC02_ARTERIAL_MIN	PH_ARTERIAL_MIN	1
2637	P02_ARTERIAL_MIN	PC02_ARTERIAL_MIN	1
3064	PH_ARTERIAL_MIN	SAT02_ARTERIAL_MIN	1
6466	TEMPERATURE_DIFF	TEMPERATURE_DIFF_REL	0.999672
6551	OXYGEN_SATURATION_DIFF	OXYGEN_SATURATION_DIFF_REL	0.998998
4256	HEART_RATE_MEAN	HEART_RATE_MEDIAN	0.996628
4171	BLOODPRESSURE_SISTOLIC_MEAN	BLOODPRESSURE_SISTOLIC_MEDIAN	0.996262
4086	BLOODPRESSURE_DIASTOLIC_MEAN	BLOODPRESSURE_DIASTOLIC_MEDIAN	0.991894
4426	TEMPERATURE_MEAN	TEMPERATURE_MEDIAN	0.990118

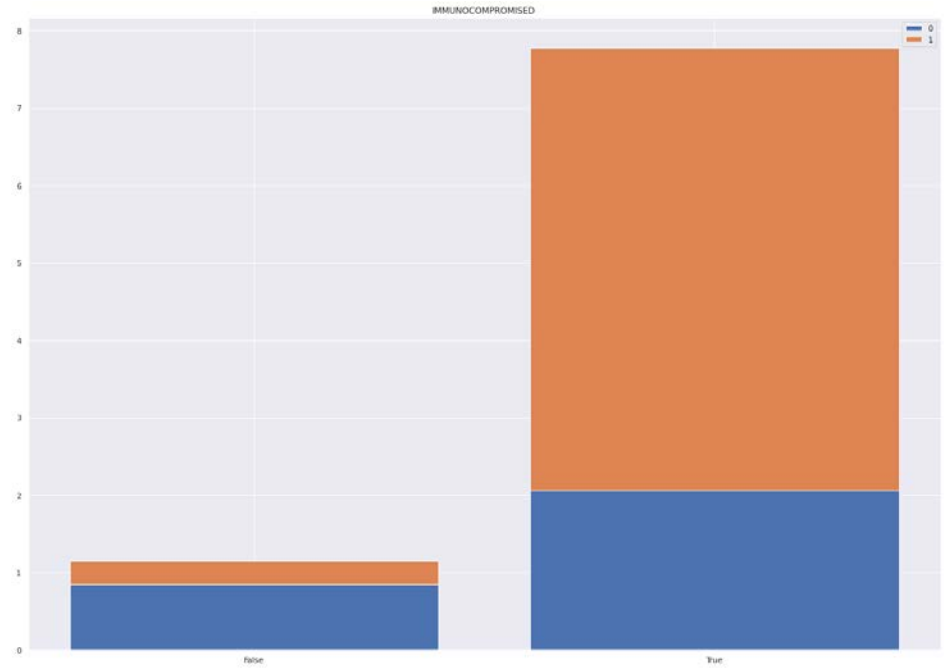
همچنین در پی محاسبه‌ی همبستگی این متغیرها با خروجی‌های مطلوب مدل، یعنی بردار برچسب‌ها (اعداد باینری نشان دهنده‌ی بستری با عدم بستری در ICU) در نظر گرفته نشده، به نظر می‌رسد که سن بالای ۶۵ سال، متغیر HTN و وضعیت سیستم ایمنی با احتمال بستری در ICU رابطه‌ی مستقیمی داشته باشند (شکل ۳).


AGE_ABOVE65
0
1
3.0
2.5
2.0
1.5
1.0
0.5
0.0
False
True

الف


HTN
0
1
8
7
6
5
4
3
2
1
0
False
True

ب

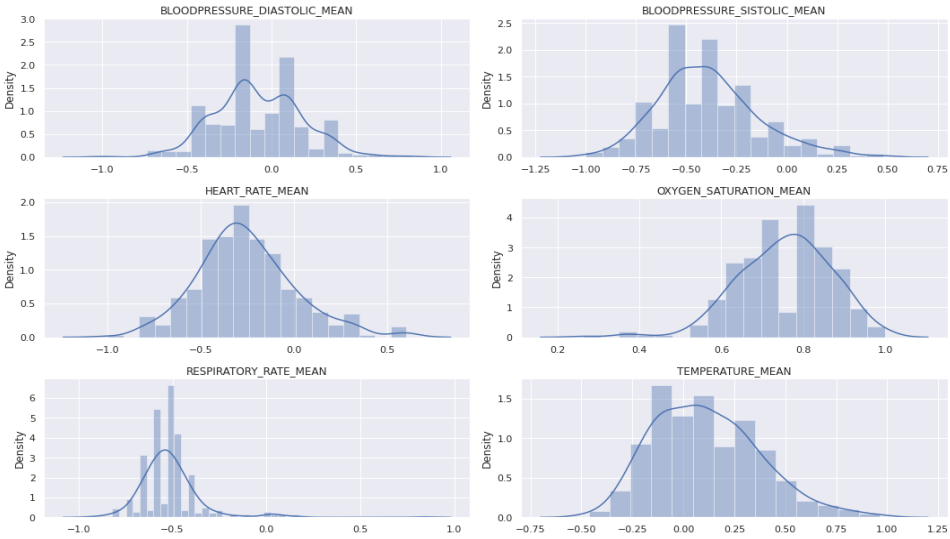


ج

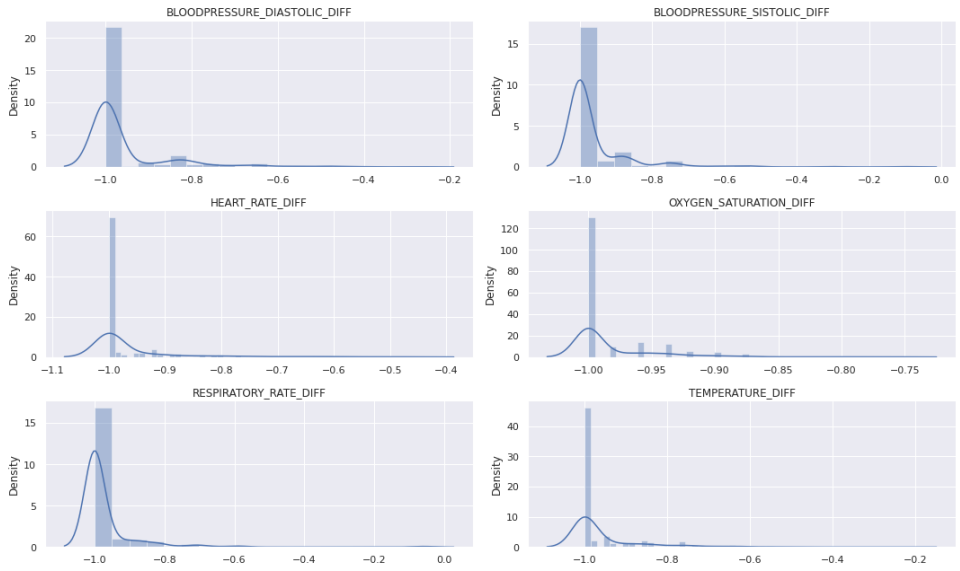
شکل ۳: نمایش ارتباط متغیرهای **الف-** سن بالای ۶۵ سال، **ب-** HTN و **ج-** وضعیت سیستم ایمنی با احتمال بستری در ICU.

۴– دسته بندی ویژگی ها

در ادامه، ویژگی ها به صورت دستی و براساس اینکه کمینه، بیشینه، میانگین، تفاوت و میانه ی یک متغیر فیزیولوژیک را اندازه گیری میکنند دسته بندی شدند تا توزیع آنها در مجموعه دادگان بررسی شود. به عنوان مثال در شکل های ۴ و ۵، این توزیع ها برای میانگین متغیرها و تفاوت عددی آنها از یک اندازه گیری تا اندازه گیری بعدی نشان داده شده است. شایان ذکر است که اعداد ثبت شده به گونه ای نرمالیزه شده اند که در بازه ی [۱و ۱-] قرار گیرند.



شکل۴: توزیع متغیرهای فیزیولوژیکی که به صورت میانگین گزارش شده‌اند.



شکل ۵: توزیع متغیرهای فیزیولوژیکی که به صورت تفاوت عددی آنها از یک اندازه‌گیری تا اندازه‌گیری بعدی نشان داده شده است.

۵– پیش بینی احتمال بستری در ICU

در این گزارش از مدل های K-نزدیک‌ترین همسایگی، Random Forest و XGBoost برای پیش‌بینی احتمال بستری شدن در ICU استفاده شده و علاوه بر معیارهای ارزیابی مانند دقت و معیار F-۱، مساحت زیر منحنی نیز گزارش شده‌اند. برخلاف گزارش‌های پیشین، هایپرپارامترهای این مدل‌ها با استفاده از cross validation و با در نظر گرفتن ۱۰ Folds تنظیم شده‌اند.

با توجه به این‌که در این پروژه با یک مساله‌ی طبقه‌بندی دو کلاسه و با مجموعه دادگان نامتعادل رو به رو هستیم، منحنی‌های ROC[۲] نیز رسم شده‌اند. در حالت ایده‌آل، تلاش بر این است که نرخ نمونه‌های به درستی مثبت تشخیص داده شده از نرخ نمونه‌هایی که به اشتباه مثبت تشخیص داده شده‌اند بیشتر شوند و منحنی به گوشه بالا و سمت چپ متمایل گردد (شکل ۶). بنابراین، افزایش مساحت زیر این منحنی به عنوان معیار برای تنظیم هایپرپارامترها در نظر گرفته شده است. منحنی ROC برای مدلی که توانایی تمایز میان دو کلاس مختلف را نداشته باشه (no skill model) یک خط با شیب ۱ خواهد بود.

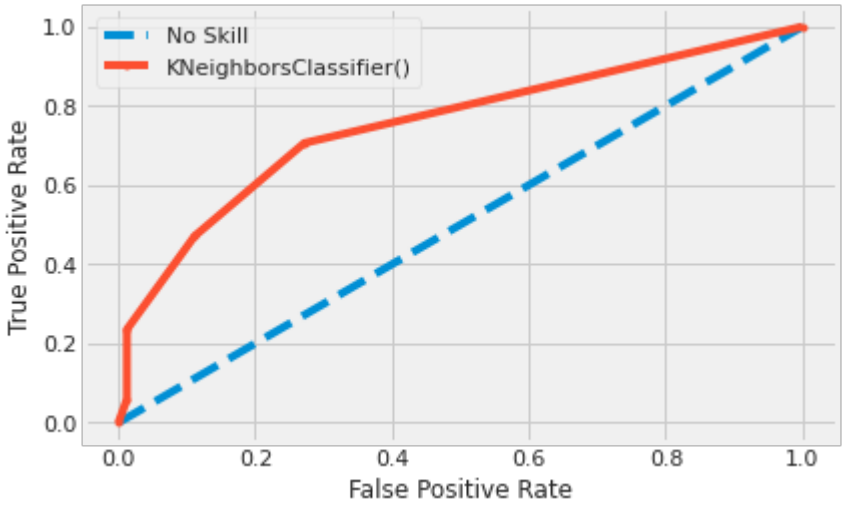
در ابتدا، یک مدل K-نزدیک‌ترین همسایگی با هایپرپارامترهای تنظیم شده تعلیم داده شد که نتایج آن در شکل ؟ نشان داده شده است.

KNN

```
CV model accuracy:  %77.48 +/- %5.67
CV model f_1 score:  %38.76 +/- %19.72
CV model roc_auc:  %83.06 +/- %6.98
```

[2] Receiver Operating Characteristic

```
Validation accuracy score: %76.52
Validation f_1 score: %37.21
Validation ROC_AUC score: %61.15
--------
No Skill: ROC AUC=%50.000
KNeighborsClassifier(): ROC AUC=%75.091
```

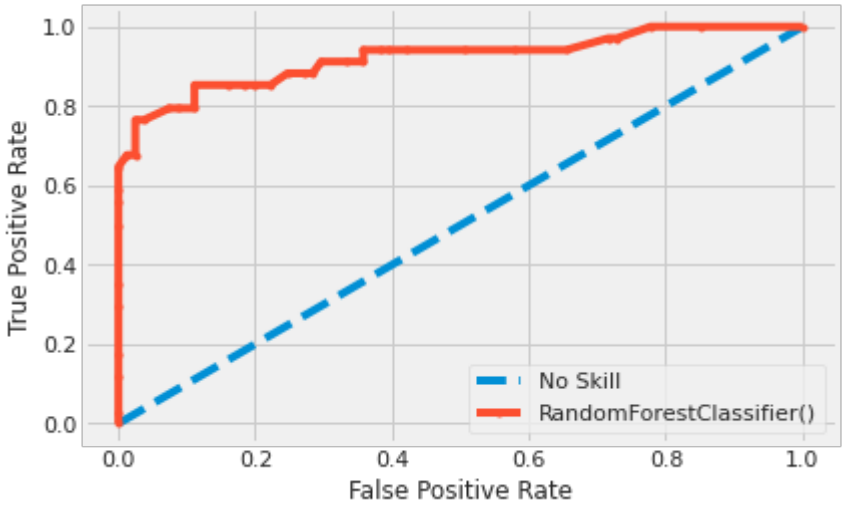


شکل ۶: منحنی ROC مربوط به عملکرد مدل K-نزدیک‌ترین همسایگی

سپس یک مدل Random Forest تعلیم داده شده و ارزیابی شده است که با استفاده از آن، مساحت زیر منحنی به طرز چشمگیری افزایش یافت.

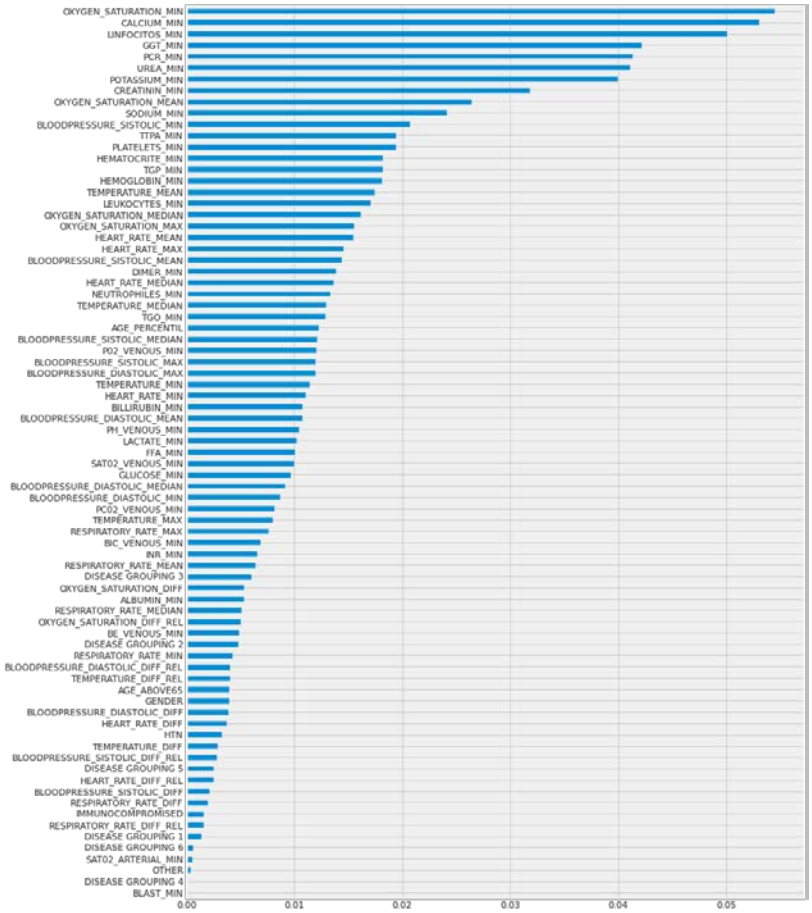
Random Forest

```
CV model accuracy:  %92.85 +/- %4.56
CV model f_1 score:  %82.87 +/- %11.82
CV model roc_auc:  %97.72 +/- %2.65
Validation accuracy score: %89.57
Validation f_1 score: %78.57
Validation ROC_AUC score: %82.35
--------
No Skill: ROC AUC=%50.000
RandomForestClassifier(): ROC AUC=%92.121
```



شکل ۷: منحنی ROC مربوط به عملکرد مدل Random Forest

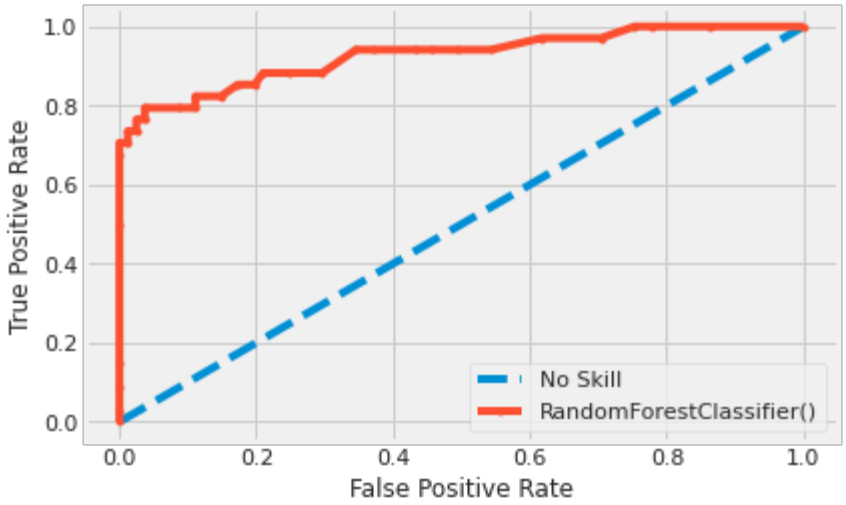
شکل ۸ ویژگی‌هایی را نشان می‌دهد که در تصمیم گیری مدل Random Forest بیشترین تأثیر را داشته‌اند. قابل توجه است که اغلب این ویژگی ها مقادیر کمینه‌ی متغیرهای فیزیولوژیکی مانند میزان اکسیژن اشباع خون، کلسیم و ... هستند.



شکل ۸: متغیرهای فیزیولوژیکی که بیشترین تاثیر را در عملکرد مدل Random Forest دارند.

تعلیم و ارزیابی مدل Random Forest روی بردار کاهش یافته‌ی ویژگی‌ها که متشکل از ۸۰٪ ویژگی‌ها به ترتیب تاثیرگذاری آن‌هاست نتایج زیر را به دنبال دارد که بهبود اندکی را در مساحت زیر منحنی نشان می‌دهد.

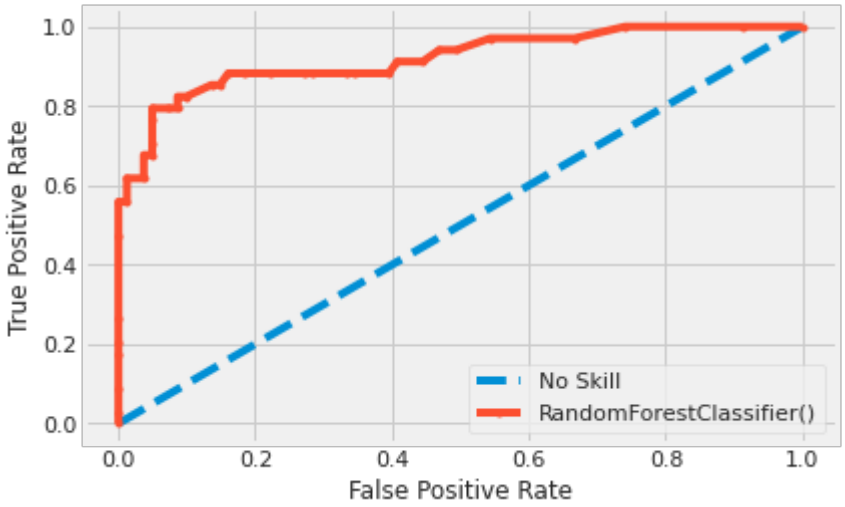
```
CV model accuracy:  %92.85 +/- %3.91
CV model f_1 score:  %83.07 +/- %10.51
CV model roc_auc:  %97.27 +/- %3.71
Validation accuracy score: %88.70
Validation f_1 score: %76.36
Validation ROC_AUC score: %80.88
--------
No Skill: ROC AUC=%50.000
RandomForestClassifier(): ROC AUC=%92.647
```



شکل ۹: منحنی ROC مربوط به عملکرد مدل Random Forest پس از کاهش ویژگی‌ها

شبیه‌سازی‌های بیشتر نشان می‌دهد که اگر از روش recursive feature elimination (RFE) برای استخراج ویژگی‌ها استفاده کنیم، بهبودی در نتایج دیده نمی‌شود .

```
CV model accuracy:  %92.09 +/- %4.60
CV model f_1 score:  %81.92 +/- %11.30
CV model roc_auc:  %96.74 +/- %4.15
Validation accuracy score: %86.96
Validation f_1 score: %73.68
Validation ROC_AUC score: %79.65
--------
No Skill: ROC AUC=%50.000
RandomForestClassifier(): ROC AUC=%91.957
```



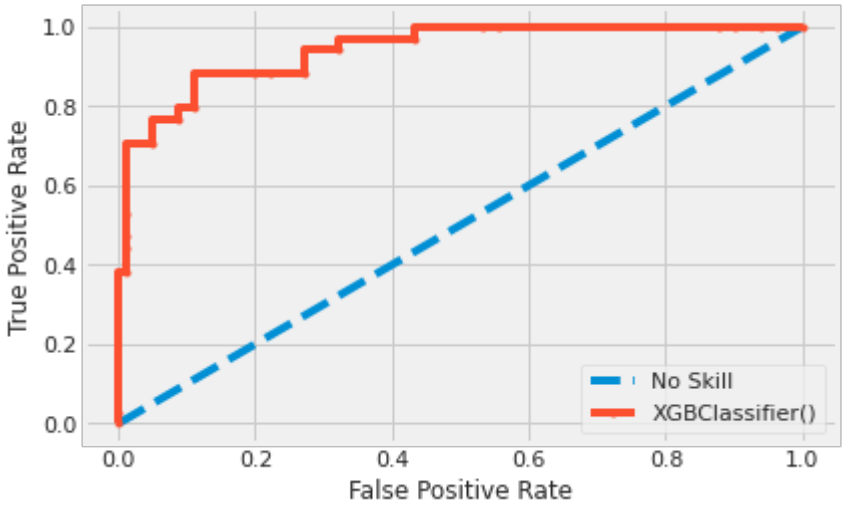
شکل ۱۰: منحنی ROC مربوط به عملکرد مدل Random Forest پس از کاهش ویژگی‌ها به روش RFE

پس از ترکیب روش سیستماتیک Cross validation با روش Grid Search برای تنظیم ابرپارامترهای مدل، مشخص شد که یک مدل Random Forest با ۱۰۰ تخمین زننده و معیار آنتروپی می‌تواند منجر به بهترین نتایج در این شبیه‌سازی‌ها شود. در تنظیم این پارامترها اولویت با بیشینه‌ی AUC در نظر گرفته شده.

```
{'criterion': 'entropy', 'max_depth': 10, 'max_features': 'log2', 'n_estimators': 100}
Validation accuracy: %85.22
Validation f1_score: %66.67
Validation ROC_AUC: %75.00
-----
No Skill: ROC AUC=%50.000
RandomForestClassifier: ROC AUC=%94.154
```

با توجه به این‌که در سایر مقالات گزارش‌هایی نیز از اثر بخشی روش XGBoost منتشر شده است، در ادامه، یکی از این مدل‌ها نیز ابتدا بدون تنظیم هایپرپارامترها و سپس با تنظیم آن‌ها به روش cross-validation تعلیم داده شده و ارزیابی شده‌اند. همان‌گونه که دیده می‌شود، پس از تنظیم پارامترها مساحت زیر منحنی بالاتر از ۹۶٪ می‌شود که بالاترین عدد در میان تمام مدل‌های امتحان شده است.

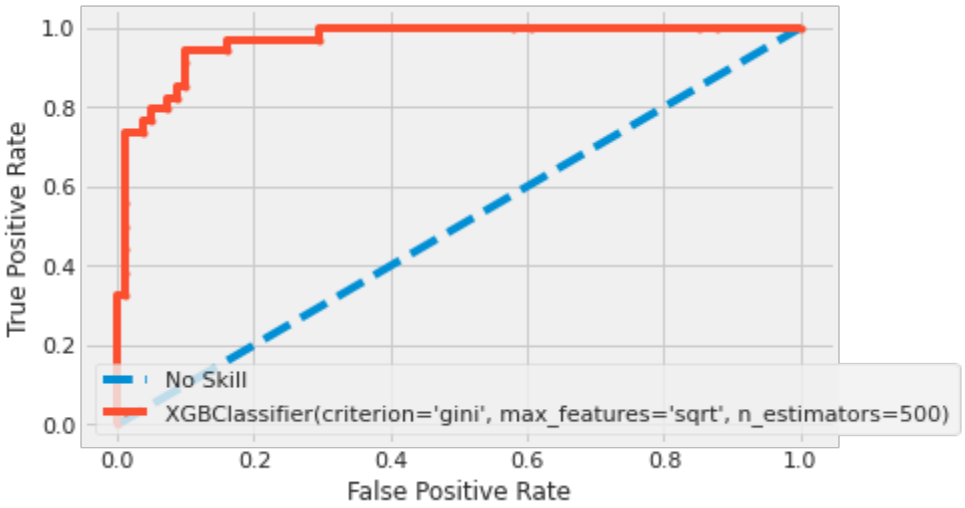
```
CV model accuracy:  %92.05 +/- %4.37
CV model f_1 score:  %81.76 +/- %11.43
CV model roc_auc:  %97.70 +/- %2.29
Validation accuracy score: %90.43
Validation f_1 score: %81.36
Validation ROC_AUC score: %84.68
--------
No Skill: ROC AUC=%50.000
XGBClassifier(): ROC AUC=%94.263
```



شکل ۱۱: منحنی ROC مربوط به عملکرد مدل XGBoost

پس از تنظیم ابرپارامترها:

```
{'criterion': 'gini', 'max_depth': 3, 'max_features': 'sqrt', 'n_estimators': 500}
Validation accuracy: %90.43
Validation f1_score: %81.36
Validation ROC_AUC: %84.68
-----
No Skill: ROC AUC=%50.000
XGBClassifier: ROC AUC=%96.550
```



شکل ۱۲: منحنی ROC مربوط به عملکرد مدل XGBoost پس از تنظیم ابرپارامترها

شایان توجه است که پس از تنظیم ابرپارامترها، متوجه شدیم که شاخص انتروپی برای مدل Random Forest و شاخص ضریب جینی برای مدل XGBoost منجر به نتایج بهتری می‌شود.

۶– نتیجه‌گیری و گام‌های پیش رو:

پس از تغییراتی در مرتب‌سازی دادگان و اجرای 5-Fold Cross Validation، نتایج حاصل از تعلیم مدل‌های Random Forest و XGBoost به مقادیری قابل مقایسه با آن‌چه در مقالات مرجع گزارش شده است نزدیک شد. در ادامه مجدداً شبیه‌سازی‌هایی روی روش‌های جبران عدم تعادل تعداد نمونه‌ها انجام می‌شود به این امید که دقت پیش‌بینی مدل از سایر روش‌ها بهتر شود.