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 ۲ –۱۴۰۱/۳/۳۰

۱– مقدمه:

در این گزارش و در ادامه‌ی پایه‌گذاری مدل‌های مرجع برای مقایسه، ابتدا روش Random Forest، که یکی از روش‌های موسوم به یادگیری ترکیبی در داده‌کاوی است، با و بدون اعمال bootstrapping پیاده سازی شده است. سپس و بعد از شناسایی مهمترین ویژگی‌ها، مدل مجدداً تعلیم داده شده و ارزیابی شده است. برخلاف گزارش ۱، در این گزارش از ۷۰٪ نمونه ها برای تعلیم و ۳۰٪ برای ارزیابی استفاده شده است.

همانگونه که در نتایج دیده می‌شود، عدم تعادل تعداد نمونه‌های دو کلاس و تعداد زیاد ویژگی‌ها در مقایسه با تعداد نمونه‌ها (بیماران) همچنان اصلی‌ترین چالش‌هایی هستند که منجر به پایین آمدن دقت مدل و f1-score می‌شوند. بنابراین، برای ایجاد دیدی بهتر نسبت به دادگان و ویژگی‌های ثبت شده، از روش‌های خطی (PCA) و غیر خطی (t-SNE و UMAP) برای نمایش دادگان استفاده شده است.

۲– پیاده سازی Random Forest :

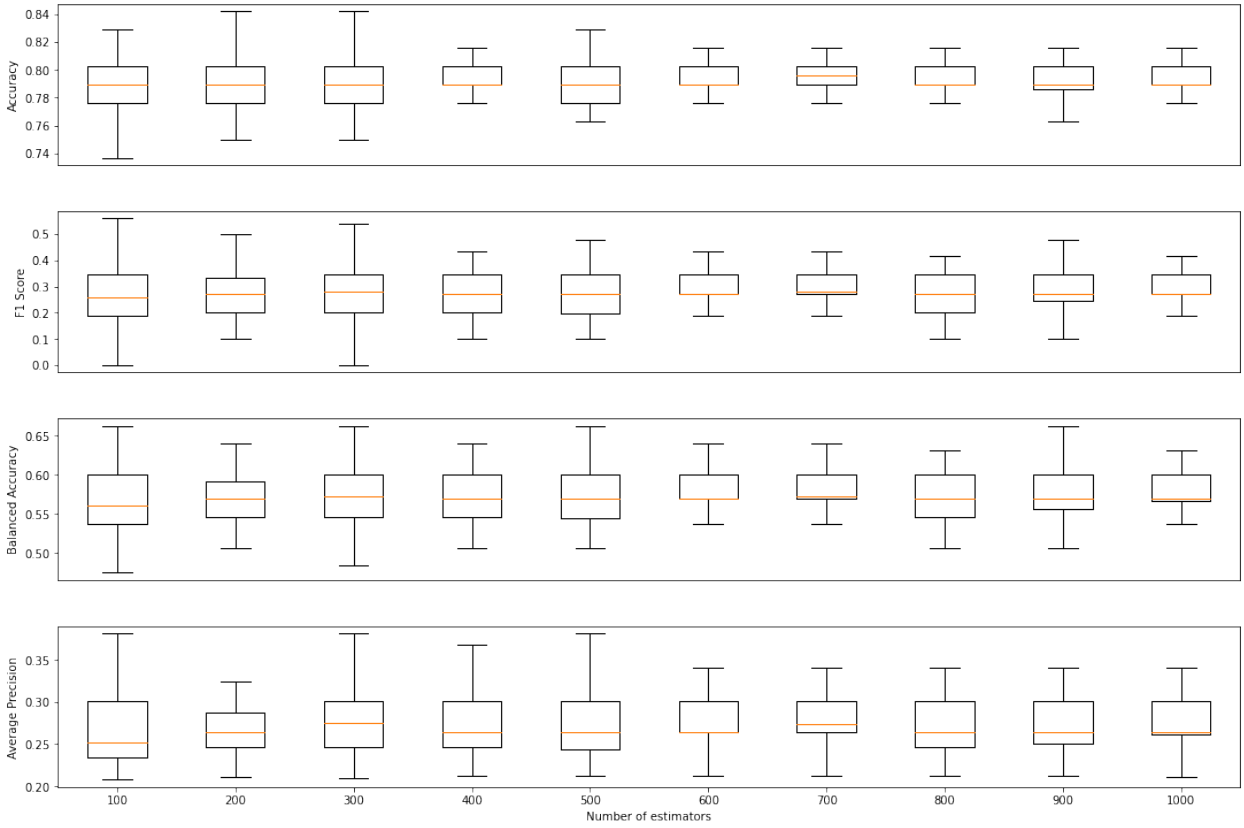
الگوریتم Random Forest یکی از روش‌های یادگیری با سرپرستی و مبتنی بر درخت تصمیم گیری است که در حل مسایل رگرسیون و دسته‌بندی به کار می‌رود. در این روش، بر روی هر یک از نمونه‌هایی از دادگان که به طور تصادفی انتخاب می‌شوند یک درخت تصمیم ایجاد شده، پیش‌بینی هر درخت از خروجی مرتبط با بردار ورودی مورد نظر جمع‌آوری شده و در نهایت، بهترین پیش‌بینی به روش رای‌گیری به عنوان خروجی نهایی معرفی می‌گردد. به منظور پیاده‌سازی این روش، مانند الگوریتم‌های قبلی، ابتدا مجموعه‌ی دادگان به دو زیرمجموعه‌ی تعلیم و ارزیابی دسته‌بندی شده‌اند. سپس بر روی نمونه‌های مستقل و تصادفی از دادگان تعلیم، درختان تصمیم با استفاده از معیارهایی مانند بهره‌ی اطلاعات، بهره‌ی نسبی و ضریب Gini، ایجاد می‌شوند. در حل این مساله‌ی دسته‌بندی، هر درخت به یکی از دو کلاس خروجی رای داده و در نهایت داده‌ی ورودی برچسب کلاسی را که بیشترین رای را دارد می‌پذیرد. یکی از کاربردهای جانبی این روش نیز یافتن مهمترین ویژگی‌ها است که به کاهش بعد مسئله کمک می‌کند. در جدول ۱، نتایج ارزیابی مدل بر روی مجموعه‌ی دادگان ارزیابی گزارش شده است. هر آزمایش ۱۰۰ بار و با Random Seed متفاوت تکرار شده و میانگین و در پرانتز، انحراف معیار پارامترهای ارزیابی مانند دقت، محاسبه و گزارش شده است.

جدول ۱: ارزیابی نتایج الگوریتم Random Forest در دو روش پیاده‌سازی و دو معیار انتخابی (تعداد درختان تصمیم = ۱۰۰)

معیار انتخابی	پیاده سازی bootstrapping	پیاده سازی وزندار الگوریتم	accuracy	f1_score	balanced accuracy	average precision
ضریب Gini	✗	✗	0.7909 (0.0210)	0.2827 (0.0920)	0.5742 (0.0379)	0.2767 (0.0446)
ضریب Gini	✓	✗	0.7772 (0.0183)	0.2068 (0.0878)	0.5443 (0.0335)	0.2455 (0.0326)
ضریب Gini	✓	✓	0.7829 (0.0194)	0.1787 (0.0790)	0.5380 (0.0278)	0.2455 (0.0316)
ضریب Gini	✗	✓	0.7800 (0.0193)	0.2698 (0.0697)	0.5657 (0.0303)	0.2618 (0.0325)
آنتروپی	✗	✗	0.7801 (0.0236)	0.2288 (0.1014)	0.5525 (0.0395)	0.2563 (0.0414)
آنتروپی	✓	✗	0.7826 (0.0193)	0.1658 (0.0810)	0.5342 (0.0286)	0.2425 (0.0310)

آنتروپی	✓	✓	0.7796 (0.0208)	0.1681 (0.0822)	0.5336 (0.0286)	0.2419 (0.0306)
آنتروپی	✗	✓	0.7767 (0.0223)	0.2164 (0.0647)	0.5460 (0.0275)	0.2482 (0.0313)

همانگونه که در جدول ۱ دیده می‌شود، اعمال الگوریتم bootstrapping و وزن‌دار کردن کلاس‌ها در هنگام تعلیم، تنها کاهش انحراف معیار را به همراه دارد و بهبود خاصی در پارامترهای دقت و f1-score ایجاد نمی‌شود. به نظر می‌رسد که بر روی این دادگان عملکرد معیار ضریب Gini نیز بهتر از آنتروپی است. بنابراین، در ادامه، ضریب Gini به عنوان معیار انتخاب شد و بدون اعمال bootstrapping و وزن‌دار کردن کلاس‌ها، اثر تعداد درختان تصمیم (تعداد تخمین) در عملکرد کلی مدل بررسی شد (شکل ۱).



شکل ۱: معیارهای ارزیابی مدل بر حسب تعداد درختان تصمیم (تعداد تخمین)

با توجه به نتایج نشان داده شده در شکل ۱، با افزایش تعداد تخمین‌ها، تغییر پارامتر دقت به صورت نمایی بوده و پس از عدد ۳۰۰ به اشباع می‌رسد. این تغییرات در سایر پارامترها چنین الگوی مشخصی را دنبال نمی‌کند اما در تمامی آن‌ها تعداد ۳۰۰ تخمین، مصالحه‌ی خوبی بین میانگین بالا و انحراف معیار پایین را نشان می‌دهد. بنابراین، پس از این آزمایش، تعداد تخمین بر روی ۳۰۰ ثابت شد و این بار تحلیل با هدف پیدا کردن موثرترین ویژگی‌ها انجام شد.

در جدول ۲، ۱۰ ویژگی که دارای بالاترین اهمیت در تصمیم گیری هستند نشان داده شده است.

جدول ۲: موثرترین ۱۰ ویژگی در تصمیم گیری جنگل تصادفی

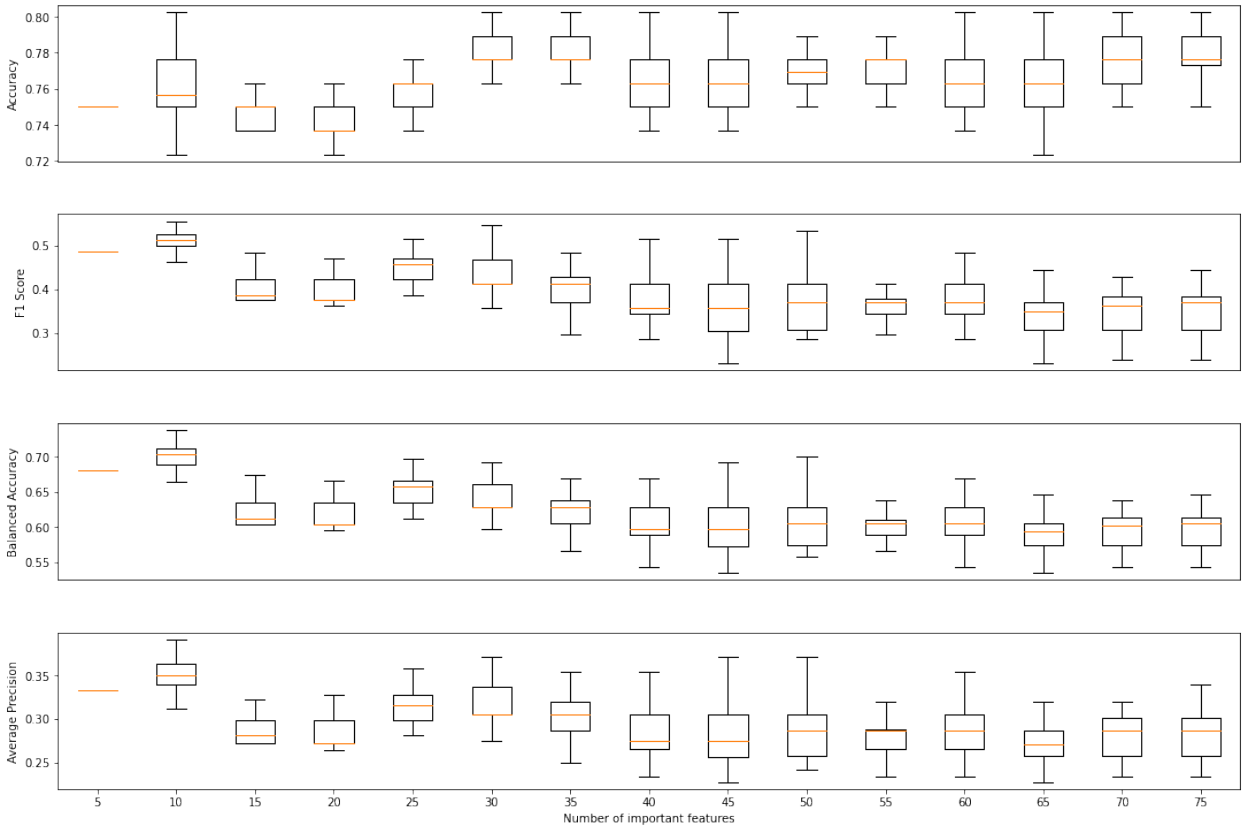
شماره ویژگی	عنوان ویژگی	پنجره زمانی	تاثیر(وزن) ویژگی
317	TEMPERATURE_MIN	4	0.033639
203	PC02_ARTERIAL	3	0.020229
119	PC02_VENOUS	2	0.018850
287	PC02_VENOUS	4	0.018419
35	PC02_VENOUS	1	0.018056
298	TTPA	4	0.017663
214	TGP	3	0.016582
305	TEMPERATURE_MEAN	4	0.014756
60	OXYGEN_SATURATION_MEDIAN	2	0.012349
274	FFA	4	0.011369

تعلیم مدل، تنها بر روی این ۱۰ ویژگی مهم نتایج ارزیابی جدول ۳ را به دنبال داشت (آزمایش ۱۰۰ بار تکرار شده است).

جدول ۳: معیارهای ارزیابی بعد از تعلیم مدل بر روی موثرترین ۱۰ ویژگی

	Mean	Standard Deviation
Accuracy	0.7574	0.0158
F1_score	0.5108	0.0243
Balanced Accuracy	0.7001	0.0177
Average Precision	0.3516	0.0197

از مقایسه‌ی نتایج جدول ۳ با آنچه در جدول ۱ و شکل ۱ گزارش شده است می‌توان نتیجه گرفت که انتخاب ۱۰ ویژگی مهم، تاثیر چندانی در بهبود دقت مدل ندارد اما سایر پارامترهای ارزیابی را به طور چشمگیری افزایش می‌دهد. برای به دست آوردن درک بهتری از تعداد ویژگی‌های مهم، آزمایش برای اعداد مختلف بین ۵ تا ۷۵ ویژگی به تعداد ۱۰۰ بار تکرار شد که نتایج در شکل ۲ نشان داده شده است.



شکل ۲: معیارهای ارزیابی مدل با تغییر تعداد ویژگی‌های مهم

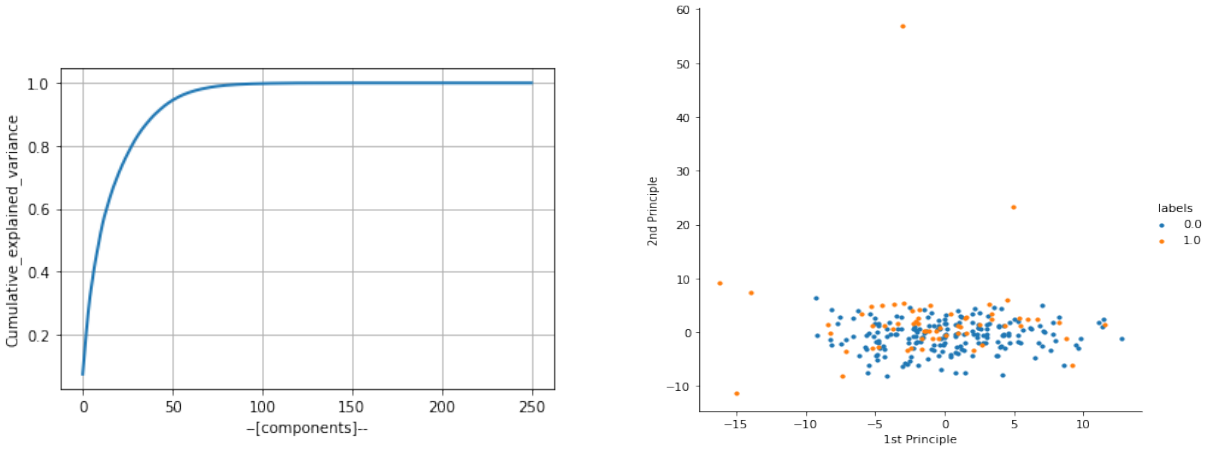
با توجه به این شکل و در نتیجه ی افزایش تعداد ویژگی های مهم، همه ی معیارهای ارزیابی به غیر از دقت، یک روند کاهشی را نشان می‌دهند و به ازای n=10 بالاترین مقادیر خود را دارند. روند تغییرات در دقت مدل، از این الگو پیروی نمیکند.

۳– کاهش بعد و نمایش دادگان:

با توجه به نتایج بخش قبل، به نظر می‌رسد استخراج ویژگی‌های مهم و کاهش بعد دادگان تأثیر خوبی بر بهبود عملکرد مدل دارد. با توجه به این موضوع، در این بخش با روش‌های PCA، t-SNE و UMAP بر روی دادگان اعمال می‌شود تا تصویر بهتری از توزیع دادگان در فضاهایی با بعد پایین‌تر به دست آید.

الف- تحلیل مؤلفه‌های اصلی (PCA):

برای نمایش دادگان در فضای با بعد پایین‌تر ابتدا آن‌ها را به گونه‌ای استانداردسازی (هنجارسازی) کردیم که میانگین هر ویژگی صفر شود. سپس با استفاده از روش محاسبه‌ی ماتریس کوواریانس، بردارها و مقادیر ویژه محاسبه شده، سپس دو مؤلفه اصلی استخراج گردیده و توزیع دادگان در فضا در شکل ۳-الف نشان داده شده است که عدم تعادل تعداد نمونه‌ها در آن به خوبی مشاهده می‌شود. به منظور بررسی تأثیر هر مؤلفه روی واریانس، مقدار تجمعی واریانس‌های تک تک مؤلفه ها محاسبه و در شکل ۳-ب رسم شده است. به خوبی مشاهده می‌شود که در صورت استفاده از یک روش خطی مانند PCA برای کاهش ابعاد ورودی، حداقل ۸۰ مؤلفه‌ی اصلی (تقریبا ٪۳۰ مؤلفه‌ها) باید لحاظ شوند.

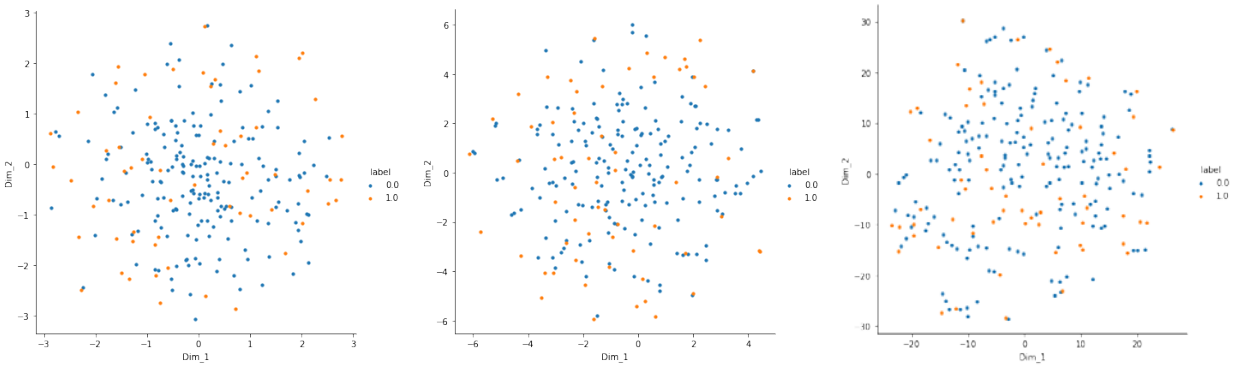


ب الف

شکل ۳: الف)توزیع نمونه‌ها در فضای دو بعدی متشکل از دو مولفه اصلی محاسبه شده به روش PCA، ب) مقدار تجمعی واریانس‌های تک تک مؤلفه ها

ب- روش t-SNE:

در ادامه برای نمایش داده‌ها در یک فضای ۲-بعدی از یک روش غیرخطی مبتنی بر تعریف همسایگی در فضای n-بعدی استفاده شده است. در این روش، ابتدا یک تابع توزیع احتمالات بر روی هر دو نقطه در فضای n-بعدی در نظر گرفته می‌شود به گونه‌ای که این تابع برای نقاط مشابه (بر اساس یک معیار خاص مشابهت) عدد بالاتری را دریافت می‌کند و برای نقاط غیرمشابه، عدد احتمال کمتری را در نظر می‌گیرد. سپس، روش t-SNE تابع توزیع احتمال مشابهی را در فضای با بعد پایین‌تر در نظر گرفته و سپس واگرایی این دو توزیع احتمالی را که با معیار کولبک-لیبلر (Kullback-Leibler) بر حسب موقعیت دو نقطه در فضا سنجیده می‌شود، کمینه می‌کند. در این پیاده‌سازی، از معیار فاصله‌ی اقلیدسی برای سنجش فاصله/شباهت میان دو نقطه استفاده می‌شود. برای رسم شکل ۴، تلاش شده که با تغییر پارامترهای این الگوریتم، توزیع نمونه‌ها در فضای دو بعدی به صورت خوشه‌های مجزا دیده شوند که ظاهراً این امر با استفاده از این روش و تغییر این پارامترها امکان‌پذیر نیست. دراین بررسی‌ها هم از داده‌های هنجار شده و هم غیر هنجار شده استفاده گردید اما تغییری در توزیع خوشه ها مشاهده نشد.



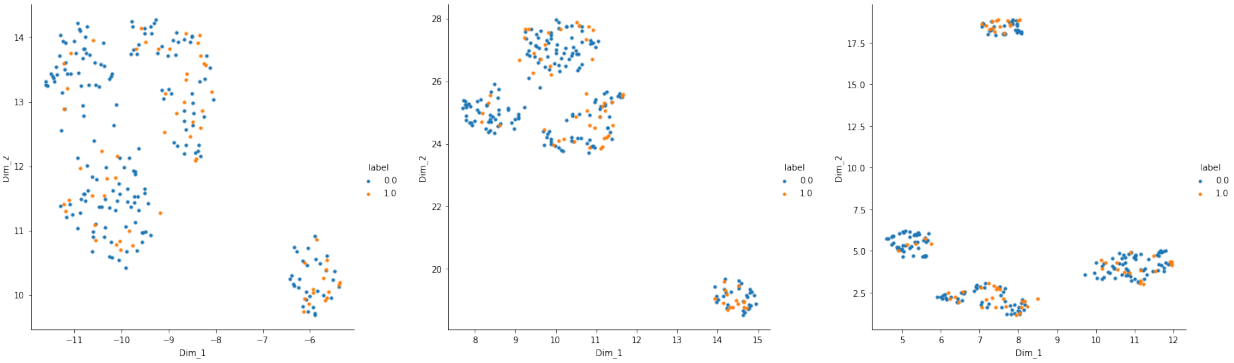
ج ب الف

شکل ۴: کاهش بعد مجموعه دادگان با استفاده از روش t-SNE با تغییر پارامتر پیچیدگی (Perplexity) الف) ۱۰، ب) ۵۰، ج) ۱۰۰

ج- روش UMAP:

یکی دیگر از روش‌های غیرخطی کاهش بعد دادگان، Uniform Manifold Approximation and Projection (UMAP) است که برخلاف t-SNE نه تنها برای نمایش دادگان به کار می‌رود بلکه کاربردهای خوشه‌بندی و کلاس‌بندی به روش با و بدون سرپرستی نیز دارد. ایده‌ی اصلی این الگوریتم نیز مانند t-SNE، تعریف همسایگی و تشابه در فضای n-بعدی است به این صورت که نقاط همسایه در فضای با بعد بالا، در فضای با بعد پایین نیز همسایه باقی می‌مانند. اما بر خلاف t-SNE، در این روش چینش اولیه‌ی نقاط در فضای با بعد پایین به صورت تصادفی صورت نمی‌گیرد، بلکه محل نقاط در این فضا متناسب با فاصله‌ی نسبی آن‌ها از یکدیگر در فضای اصلی تعیین می‌شود. این تفاوت باعث می‌شود که نتایج UMAP برخلاف نتایج t-SNE تکرارپذیر باشند. به علاوه، از آن‌جا که در هر تکرار فقط یک نقطه و یا مجموعه‌ی کوچکی از نقاط در فضای با بعد پایین جابه‌جا می‌شوند تا خوشه‌هایی از دادگان شکل بگیرند، پیاده‌سازی UMAP نیز سریع‌تر از t-SNE است که در هر تکرار، تمام نقاط جابه‌جا می‌شوند.

مهمترین پارامتر این الگوریتم تعداد نمونه‌های مورد نظر در همسایگی هر نمونه است. درحالیکه همسایگی بزرگتر تصویری کلی از چینش دادگان در فضا به دست می‌دهد، همسایگی‌های کوچکتر، جزییاتی از خوشه‌های داخلی موجود در هر خوشه‌ی اصلی را نشان می‌دهند. به علاوه، این روش را هم می‌توان به روش با سرپرستی و هم بدون سرپرستی تعلیم داد. شکل ۵ مثالی از خوشه‌بندی دادگان به این روش و به ازای ۳ عدد متفاوت (۱۵، ۵۰ و ۱۰۰) برای همسایگی را نشان می‌دهد.

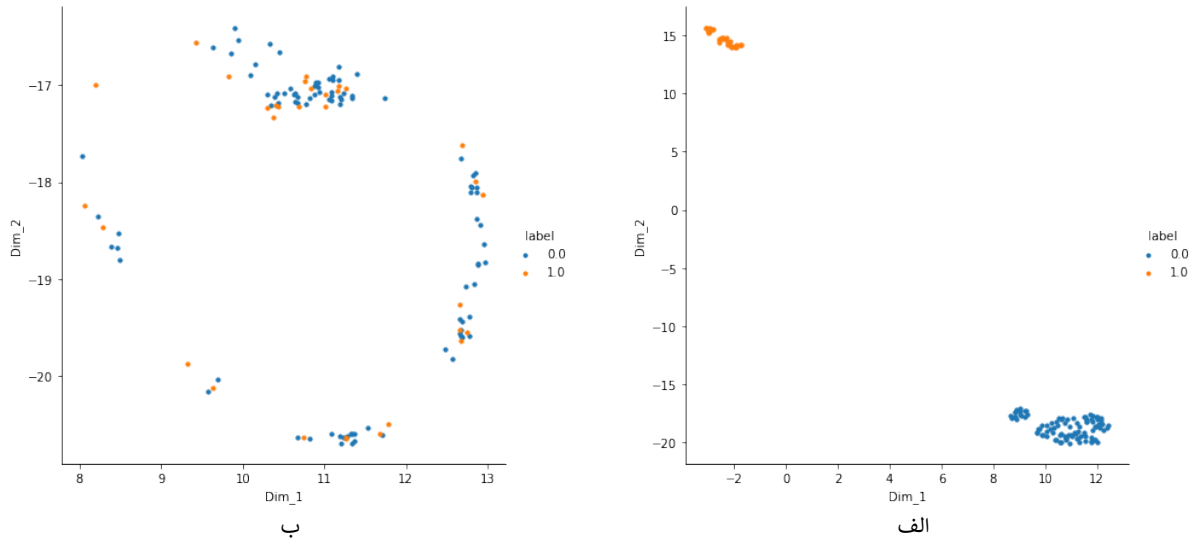


ج ب الف

شکل ۵: کاهش بعد مجموعه دادگان با استفاده از روش UMAP بدون سرپرستی و با تغییر پارامتر همسایگی. الف)۱۵، ب) ۵۰، ج) ۱۰۰

در حالت تعلیم با سرپرستی، همان‌گونه که در شکل ۶ دیده می‌شود، عدم تعادل تعداد نمونه‌ها و انتخاب پارامترها باعث شده که مدل روی دادگان تعلیم over fit شده و نتایج روی دادگان ارزیابی قابل قبول نباشد. برای حل این مشکل، به نظر می‌رسد که به جای کاهش

بعد دادگان به دو بعد، باید ابعاد بالاتری را نیز امتحان نمود و یا اینکه به طور سیستماتیک، پارامترها را تغییر داد و اثر آنها بر روی نتایج را مشاهده نمود.



شکل ۶: کاهش بعد مجموعه دادگان با استفاده از روش UMAP با سرپرستی. الف) مجموعه تعلیم که وجود دو خوشه را به خوبی نشان میدهد. ب)مجموعه ارزیابی که همه نمونه ها در خوشه ی سالم قرار گرفته اند.

۴– مراحل آینده:

همان‌گونه که در این گزارش دیده شد همچنان عدم تعادل میان دادگان مهمترین علت پایین بودن دقت مدل و سایر پارامترهای ارزیابی به نظر می‌رسد بنابراین در گزارش بعد روش‌های متعادل‌سازی تعداد نمونه‌ها برروی این مجموعه دادگان پیاده خواهد شد و اثر آن گزارش می‌شود. به علاوه در این گزارش تنها روش Random Forest از میان روش‌های یادگیری ترکیبی پیاده‌سازی شد. در گزارش های بعد روش‌هایی مانند XGBoost بر روی این دادگان پیاده خواهد شد. همچنین ممکن است روش‌های پیشرفته‌تر یادگیری بدون سرپرستی نیز برای ادامه تحقیق در نظر گرفته شود.

نباید فراموش کرد که موضوع پر کردن فضاهای خالی در دادگان همچنان باز بوده و باید مورد آدرس‌دهی قرار گیرد.