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

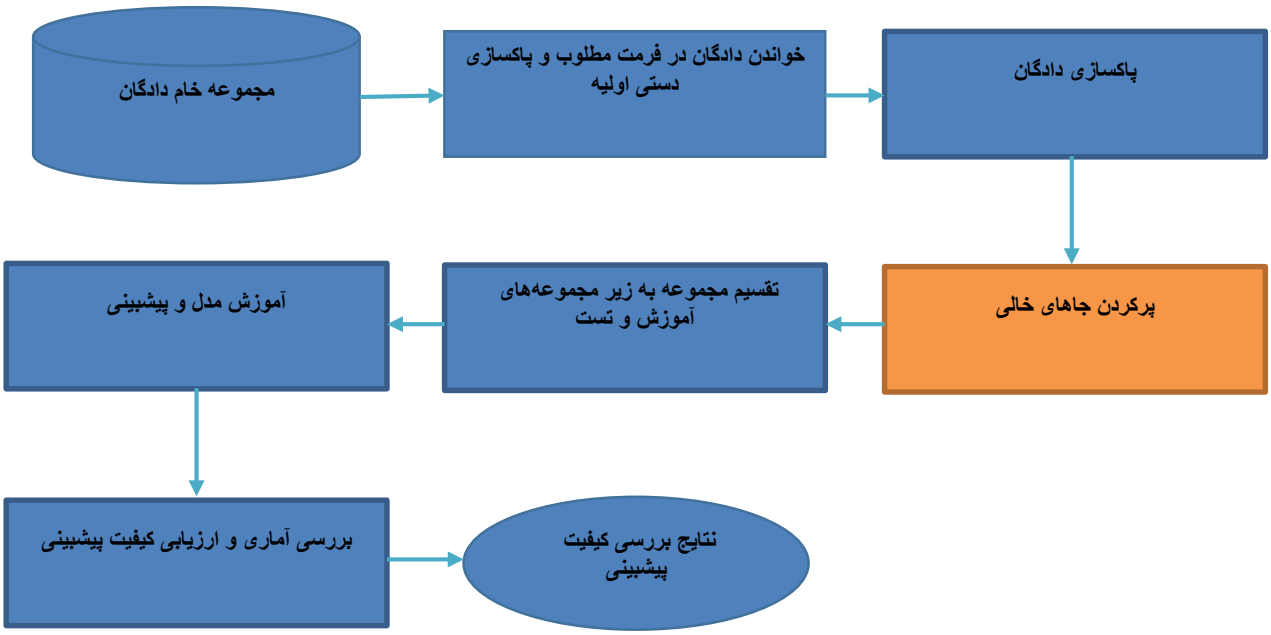
گروه تخصصی: مهندسی نرم‌افزار

استاد راهنما: خانم دکتر مریم رستگارپور

گزارش پیشرفت: شماره ۳ –۱۴۰۱/۵/۳۰

۱– مقدمه:

در گزارش قبل پایه‌گذاری مدل‌های مرجع و مقایسه آن‌ها را ادامه دادیم و روش Random Forest با و بدون اعمال bootstrapping و با و بدون وزن‌دار کردن الگوریتم پیاده سازی شد. همچنین نمایش دادگان به روش‌های خطی (PCA) و غیر خطی (t-SNE و UMAP) اجرا شد. در این گزارش با بازنگری در پایپ لاین اولیه، گزینه‌های مختلفی برای مرحله پر کردن جاهای خالی پیشنهاد و پیاده سازی شده و نتایج مدل‌های مرجع این بار با در نظر گرفتن این گزینه‌ها با هم مقایسه شده‌اند.



شکل ۱: در این گزارش با بازنگری در پایپ لاین اولیه، گزینه‌های مختلفی برای مرحله پر کردن جاهای خالی پیشنهاد و پیاده سازی شده

۲– روش Padding:

این روش در اصل همان روش اجرا شده در پایپ لاین اولیه است که به عنوان مرجع مقایسه برای روش‌های دیگر در نظر گرفته شده است. این روش همچنین پیشنهاد جمع‌آورندگان مجموعه دادگان بوده است. در این روش، در صورتی که اطلاعات حیاتی بیمار در یک مرحله از پنجره‌های زمانی اندازه‌گیری نشده باشد، فرض بر آن است وضعیت بیمار پایدار بوده و می‌توان جای خالی این اطلاعات را با اطلاعات اندازه‌گیری شده در پنجره زمانی قبل یا بعد پر کرد.

جدول ۱: نمونه‌ای از روش padding

PATIENT_VISIT_IDENTIFIER	ALBUMIN_MIN	BE_ARTERIAL_MIN	BE_VENOUS_MIN	BIC_ARTERIAL_MIN	BIC_VENOUS_MIN	BILLIRUBIN_MIN
7	0.60526316	-1	-1	-0.31707317	-0.31707317	-0.96650968
7	0.60526316	-1	-0.94764398	-0.31707317	-0.36585366	-0.93894994
7	0.60526316	-1	-0.94764398	-0.31707317	-0.36585366	-0.93894994
7	0.60526316	-1	-0.94764398	-0.31707317	-0.36585366	-0.93894994

copy from above

۳– روش‌های ساده آماری:

در این روش‌ها جاهای خالی با استفاده از یکی از ابزارهای آماری توصیفی (مثلاً میانگین، میانه یا متداول ترین) در امتداد هر ستون جایگزین می‌شوند.

۳,۱- **متداول‌ترین مقدار(**Most Frequent**) :** پر تکرار ترین مقدار در امتداد هر ستون در جای خالی نوشته می‌شود. در صورتی که بیش از یک مقدار واجد این شرایط باشند مقدار کمتر نوشته خواهد شد.

۳,۲- **میانگین (**mean**):** میانگین مقادیر موجود در هر ستون در جای خالی نوشته می‌شود.

۳,۳- **میانه (**median**):** میانه مقادیر موجود در هر ستون در جای خالی نوشته می‌شود.

۴– روش چند متغیره یا تکرارشونده:

رویکرد پیچیده‌تری برای پیشبینی مقادیر خالی استفاده از روش‌های چند متغیره است. به این صورت که ویژگی شامل جاهای خالی به صورت تابعی از دیگر ویژگی‌ها تخمین زده می‌شود و سپس این تابع تخمینی برای پر کردن جاهای خالی مورد استفاده قرار می‌گیرد. تخمین تابع به صورت حلقه‌ای تکرار شونده برای حصول دقت بالاتر اجرا می‌شود.

۵– روش K–نزدیکترین همسایگی:

در روش پیاده سازی از فاصله اقلیدسی برای تعیین نزدیکترین همسایه استفاده شده است. این فاصله طبیعتاً برای همسایگی در هر ویژگی به صور جداگانه حساب شده است و در روش پیاده‌سازی شده همه فاصله‌ها به صورت یکنواخت (بدون وزن دهی) محاسبه شده‌اند. در صورتی که در یک نمونه بیش از یک ویژگی مفقود شده باشد، به ازائ هر ویژگی ناموجود، همسایه‌های مختلفی برای این نمونه پیدا خواهد شد. در صورتی که تعداد همسایه‌ها از حد مشخصی کمتر باشد و فاصله با داده‌های تعلیم نامشخص، میانگین داده‌های تعلیم برای پر کردن ویژگی خالی استفاده می‌شود.

۶– نتایج بررسی کیفیت پیشبینی بعد از بازنگری در ماژول پر کردن جاهای خالی:

بعد از بازنگری در ماژول «پر کردن جاهای خالی»، پایپ لاین پروژه مطابق گزارش ۲ مجدداً و برای همه گزینه‌های پر کردن جاهای خالی اجرا شد. جهت اطمینان از یکسان بودن شرایط مقایسه از جمله **Seed** اتفاقی، هر حالت در ۱۰۰ آزمایش تکرار شده و میانگین و انحراف معیار محاسبه شده است.

برای مقایسه روش‌های پر کردن جاهای خالی، نتایج بهترین روش پیشبینی (تا به حال) یعنی **Random Forest** با معیار ضریب جینی و بدون پیاده‌سازی **Bootstrapping** و بدون وزن‌دار کردن الگوریتم استفاده شده است.

جدول ۲: ارزیابی نتایج الگوریتم Random Forest با معیار انتخابی جینی بعد از پر کردن جاهای خالی به روش‌های مختلف (تعداد درختان تصمیم = ۱۰۰) (انحراف معیار در پرانتز)

روش پر کردن جاهای خالی	accuracy	f1_score	balanced accuracy	average precision
padding	0.7909 (0.0210)	0.2827 (0.0920)	0.5742 (0.0379)	0.2767 (0.0446)
متداول‌ترین مقدار	0.7308 (0.0083)	0.2458 (0.0326)	0.5501 (0.0127)	0.3001 (0.0120)
میانگین	0.7319 (0.0081)	0.2460 (0.0285)	0.5506 (0.0113)	0.3009 (0.0110)

میانه	0.7315 (0.0089)	0.2470 (0.0339)	0.5508 (0.0134)	0.3010 (0.0128)
چند متغیره	0.7331 (0.0120)	0.2927 (0.036)	0.5654 (0.0161)	0.3127 (0.0167)
k-نزدیکترین همسایگی	0.7299 (0.0080)	0.2345 (0.0358)	0.5465 (0.0132)	0.2972 (0.0120)

همان‌گونه که از نتایج مشخص است هیچ یک از روش‌های پیاده‌سازی شده برای پر کردن جاهای خالی نتوانسته است تاثیر مثبتی برروی نتایج پیشبینی بگذارد. تنها انحراف معیار شاخص های ارزیابی تا حدی بهتر شده است که نشان می‌دهد که یکپارچگی دادگان و پیشبینی بهبود جزیی پیدا کرده است. البته همچنان می‌توان بررسی‌های بیشتری بر روی پر کردن جاهای خالی انجام داد. به عنوان مثال می‌توان ساختارهای ترکیبی از روش‌های ساده آماری و روش تکرارشونده پیاده‌سازی کرد و یا پارامترهای مختلف را در روش K- نزدیکترین همسایگی امتحان کرد ولی با توجه به نتایج بالا و همچنین ساختار دادگان به نظر نمی‌رسد که هیچکدام از این روش‌ها امیدبخش باشند و همچنان روش padding برای ادامه کار کافی است.

۷- مراحل آینده:

پس از این‌که مشخص شد روش‌های مختلف پر کردن جاهای خالی تاثیر قابل توجهی برروی نتایج نخواهند داشت همچنان یافتن راه‌حلی برای عدم تعادل میان دادگان به عنوان مهمترین علت پایین بودن دقت مدل، موضوع گزارش‌های بعدی خواهد بود. ضمناً پیاده‌سازی روش‌های یادگیری ترکیبی دیگر مانند XGBoost بر روی این دادگان نیز در دستور کار است.