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

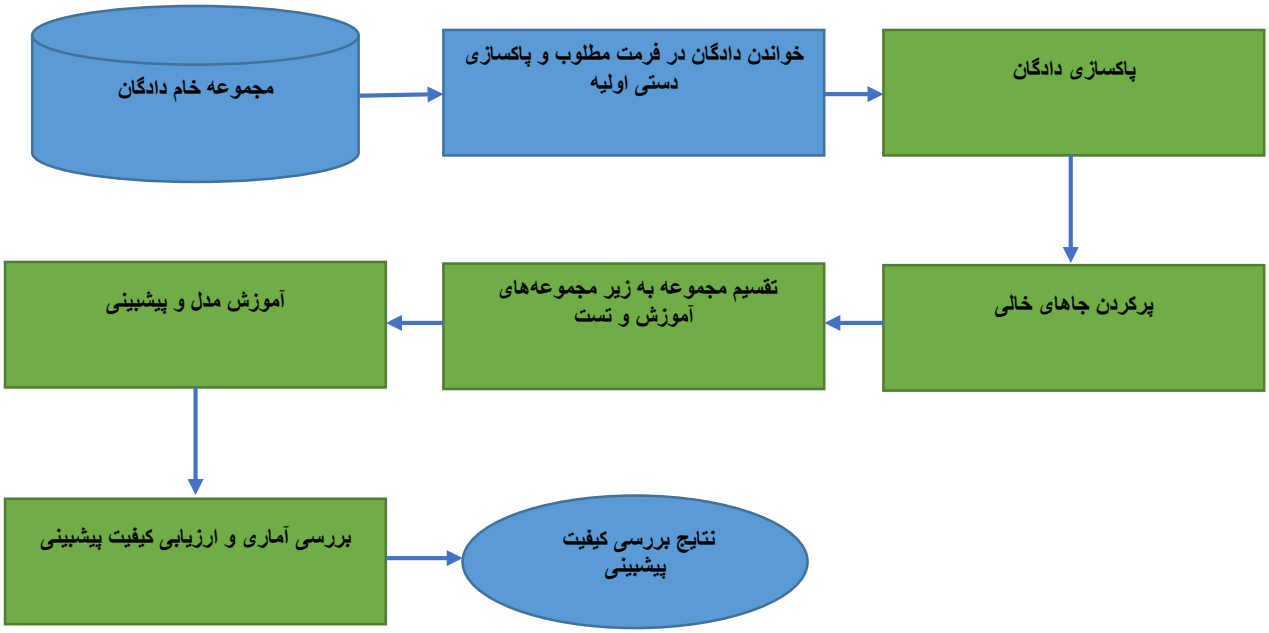
استاد راهنما: خانم دکتر مریم رستگارپور

گزارش پیشرفت: شماره ۱ –۱۴۰۱/۳/۵

۱– مقدمه:

در این گزارش ابتدا مجموعه دادگان با جزییات بیشتری نسبت به پروپوزال بررسی و معرفی شده است. در مرحله بعد، سعی شده است که یک نمونه ابتدایی از پایپ لاین پردازش دادگان پروژه طراحی و اجرا شود. به گونه‌ای که پیشبینی‌های این پایپ لاین به عنوان مرجعی برای صحه‌گذاری، کیفیت سنجی و توسعه تکنیک‌هایی که در گزارش‌های بعد ارائه خواهند شد به کار گرفته شود.

پایپ لاین اولیه پیشنهادی این تحقیق در شکل ۱ نمایش داده شده است.



شکل ۱: طرح کلی برای پایپ لاین اولیه پروژه. این پایپ لاین به عنوان مرجع در نظر گرفته شده و با پیشرفت مراحل تحقیق نتایج هر مرحله با نتایج اولیه مقایسه خواهد شد.

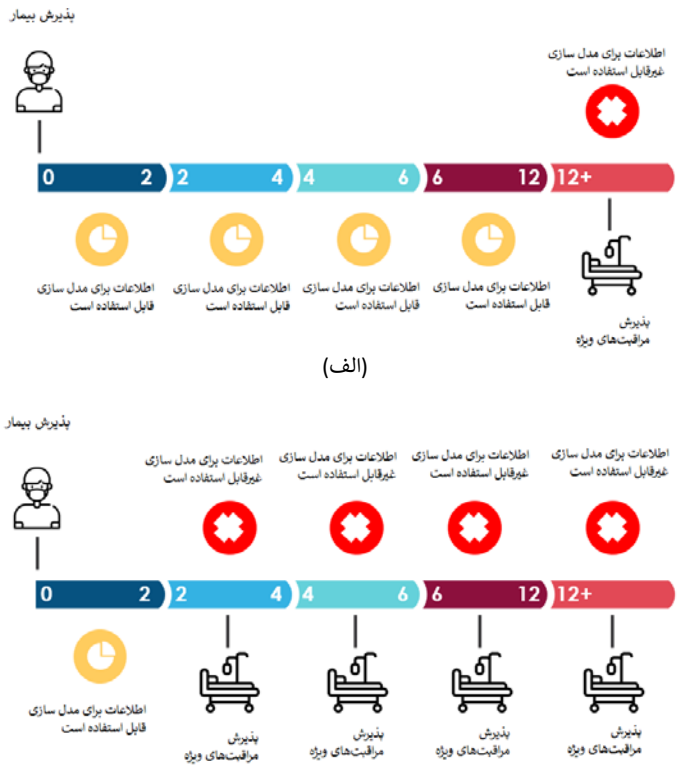
۲– مجموعه دادگان:

داده‌های مورد استفاده در این تحقیق از ۳۸۴ مراجعه به بیمارستان سیریولبانز در سائوپلو برزیل جمع‌آوری شده است و به صورت رایگان در اختیار عموم قرار گرفته و از پایگاه اینترنتی Kaggle قابل دسترسی است [۱].

متغیرهای حیاتی و نتایج آزمایش خون و غلظت اکسیژن خون برای بیماران مبتلا به کووید ۱۹ در پنجره‌های زمانی ۲ ساعته از لحظه پذیرش تا ۱۲ ساعت بعد از پذیرش اندازه‌گیری و ثبت شده‌اند.

این مجموعه داده‌ها در فرمت استاندارد مایکروسافت اکسل xlsx ذخیره شده است و کل مجموعه داده‌ها شامل یک جدول با تعداد ۱۹۲۶ سطر (۱۹۲۵ سطر اطلاعات بیماران و یک سطر عنوان ستون‌ها) و ۲۳۱ ستون است.

پایگاه Kaggle بر این موضوع تاکید کرده است که اطلاعات حیاتی بیمارانی که در دارای پرچم هدف هستند (در بخش مراقبت‌های ویژه بستری شده‌اند) نباید در مدل سازی برای پیش‌بینی استفاده شود. زیرا توالی حوادث مشخص نیست. مثلاً ممکن است که نیاز به بستری شدن در بخش مراقبت‌های ویژه قبل از جمع‌آوری اطلاعات اتفاق افتاده باشد (شکل ۲) [۱].



شکل ۲: اطلاعات حیاتی بیمارانی که در حال حاضر در بخش مراقبت‌های ویژه بستری شده‌اند دیگر نباید در مدل سازی برای پیش‌بینی استفاده شود.

ویژگی‌هایی که در هر مراجعه ثبت شده‌اند شامل ۳ ویژگی از اطلاعات جمعیتی بیمار، ۹ ویژگی از سوابق بیماری، ۳۶ ویژگی از نتایج آزمایش خون و ۶ ویژگی از اندازه‌گیری علایم حیاتی هستند.

مقادیر هر یک از این ۵۴ ویژگی براساس ماهیت اندازه‌گیری، یا به صورت یک عدد حقیقی و یا به صورت یک عدد دودویی ذخیره شده‌اند. در صورتی که مقدار ویژگی یک عدد حقیقی باشد، داده موجود در هر ستون بر اساس مقادیر بیشینه و کمینه در محدوده [1 ,1-] وزن‌گذاری شده‌اند. ضمناً این داده‌های حقیقی به صورت مقدار بیشینه، کمینه، میانه، میانگین و تغییرات دامنه، توسعه یافته‌اند که در مجموع تعداد ستون‌های مجموعه داده را به ۲۳۱ رسانده است.

یکی از چالش‌های کار با این مجموعه –و اغلب دادگان پزشکی- وجود جاهای خالی در اطلاعات جمع‌آوری شده است. در این وضعیت که اطلاعات بیمار در یک یا چند تا از پنجره‌های زمانی جمع‌آوری نشده است، پیشنهاد جمع‌آورندگان اطلاعات این است که جای خالی این اطلاعات می‌تواند با اطلاعات اندازه‌گیری شده در پنجره زمانی قبل پر شود [۱].

۳– خواندن دادگان:

مجموعه دادگان موجود در سایت Kaggle در فرمت xlsx مایکروسافت اکسل ارائه شده است. با بررسی اطلاعات موجود در مجموعه داده مشخص می‌شود که بدون نگرانی از از دست دادن اطلاعات می‌توان آن را به فرمت csv تبدیل کرد. فرمتی که برای پاکسازی و سپس تجزیه و تحلیل در محیط‌های برنامه‌نویسی مناسب‌تر است. بنابراین در اولین قدم این تبدیل انجام شد.

در حال حاضر طراحی پایپ لاین اولیه با زبان پایتون و در محیط گوگل کولب[1]انجام شده است. برای خواندن اطلاعات و عملیات پیش‌پردازش از کتابخانه پاندا[2](https://pandas.pydata.org/) استفاده شده است.

مخزن نرم‌افزاری مربوط به این گزارش در آدرس زیر قابل دسترسی است:

https://colab.research.google.com/drive/13mQD__bHDcGiwuTtcUVic-gPb5cB4-v7?usp=sharing

۴– پاکسازی دادگان:

اطلاعات حیاتی بیماران تنها در صورتی در مدل سازی برای پیش‌بینی استفاده می‌شود که این بیماران در حال حاضر در بخش مراقبت‌های ویژه بستری نشده باشند. با در نظر گرفتن این موضوع مراجعاتی که در بدو ورود در بخش مراقبت‌های ویژه بستری شده‌اند از فرآیند مدل‌سازی حذف می‌شوند و عملاً تعداد مراجعات قابل بررسی به ۳۵۳ مورد کاهش می‌یابد.

اطلاعات بیماران در بازه‌های ۲ ساعته و در پنج مرحله جمع‌آوری شده است. طبیعتاً در این بین برخی از بیماران در مرحله اول و برخی در مرحله دوم تا پنجم به بخش مراقبت‌های ویژه منتقل شده‌اند. برخی هم هرگز در بخش مراقبت‌های ویژه بستری نشده‌اند. از طرف دیگر اطلاعات کسانی که در مراقبت‌های ویژه بستری شده‌اند دیگر قابل استفاده نخواهد بود. این موضوع باعث می‌شود که سری زمانی در مورد هر بیمار طول متفاوتی داشته باشد و انتخاب و آموزش مدل پیچیده می‌شود. برای طراحی پایپ لاین اولیه تنها از اطلاعات بیمارانی استفاده شده است که شامل سری کامل زمانی هستند. (شکل ۲-الف). بدین معنی که این بیماران در چهار مرحله اولی در بخش عادی بستری بوده‌اند و تنها در مرحله آخر (پنجم) یا به بخش مراقبت‌های ویژه منتقل شده‌اند و یا خیر (مرخص شده‌اند؟).

با اعمال این محدودیت تعداد مراجعات قابل بررسی باز هم کاهش می‌یابد و به ۲۵۵ مورد می‌رسد.

طبیعتاً با پیشرفت تحقیق و در قدم‌های آینده به موضوع سری‌های زمانی نامساوی با جزییات پرداخته خواهد شد و نتایج بررسی‌ها با نتایج اولیه مقایسه خواهند شد.

۵– پرکردن جاهای خالی

در پایپ لاین اولیه، پیشنهاد جمع‌آورندگان مجموعه دادگان -که البته ساده‌ترین نحوه برخورد با چالش جاهای خالی است- پیاده‌سازی شده 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.

روش‌های پیشرفته‌تر پرکردن جاهای خالی دادگان در آینده پیاده‌سازی خواهد شد و نتایج با روش اولیه مقایسه می‌شوند.

باید به این نکته توجه داشت که از تعداد ۲۵۵ مورد مراجعه، تعداد ۶۵ مورد به بستری شدن در بخش مراقبت‌های ویژه منجر شده است و بقیه موارد عادی بوده است. این عدم تعادل نمونه‌برداری باعث کاهش کیفیت آموزش مدل و در نهایت کاهش دقت پیش‌بینی خواهد شد. با این وجود در پایپ لاین اولیه، اطلاعات به صورت موجود و بدون استفاده از روش‌های نمونه‌افزایی پردازش شده‌اند. بنابراین در آینده مقدار تاثیر احتمالی روش‌های نمونه‌افزایی در مقایسه با نتایج این طراحی اولیه قابل اندازه‌گیری خواهد بود.

[1] Google Colab
[2] Pandas

۶– زیرمجموعه‌های آموزش و تست

از آن‌جا که در طراحی پایپ لاین اولیه، تنظیم دقیق ابرپارامترها[۳] یکی از اهداف نیست، مجموعه اطلاعات به دو بخش آموزش[۴] و تست[۵] تقسیم شده‌اند و بخش اعتبارسنجی[۶] در نظر گرفته نشده است. در آینده تنظیم دقیق ابرپارامترها نیز برای مدل‌های مختلف اجرا خواهد شد. قابل ذکر است که در اولین طراحی ۵۰٪ از مجموعه اطلاعات به عنوان زیرمجموعه آموزش و ۵۰٪ به عنوان تست در نظر گرفته شده است.

در بررسی اولیه مجموعه دادگان ستون‌هایی با تعداد زیاد اطلاعات تکراری دیده می‌شود. بر این اساس و به طور موازی، جهت کاهش ابعاد، زیرمجموعه‌های آموزش و تست به وسیله یک تابع تحلیل مؤلفه‌های اصلی[۷] از کتابخانه sklearn [۲] بررسی شده و با حذف ویژگی‌های با واریانس کم، تعداد ستون‌های جدول دادگان به ۱۵ ستون کاهش یافته است.

۷– آموزش مدل و پیش‌بینی

در پایپ لاین اولیه، از دو مدل آماری کلاسیک به عنوان طبقه‌بندی کننده استفاده شده است. ابتدا یک طبقه‌بندی کننده رگرسیون لجستیک[۸][۳] و در ادامه دو مدل ماشین بردار پشتیبانی[۹][۴] به صورت خطی و غیرخطی پیاده سازی شده اند. برای پیاده‌سازی این مدل‌ها از کتابخانه sklearn استفاده شده است.

هر سه مدل در چهار حالت مختلف اجرا شده و نتایج با هم مقایسه شده‌اند:

جدول ۱: حالت‌های اجرای طبقه‌بندی کننده‌های پایپ‌لاین اولیه

بدون متعادل سازی وزن‌ها و بدون کاهش ابعاد	بدون متعادل سازی وزن‌ها و با کاهش ابعاد
با متعادل سازی وزن‌ها و بدون کاهش ابعاد	با متعادل سازی وزن‌ها و با کاهش ابعاد

ورودی این مدل‌ها، اطلاعات جمع‌آوری شده از بیماران در چهار پنجره زمانی است که به صورت یک بردار یک بعدی تغییر فرم داده شده است. پرچم هدف یک داده دوتایی است. عدد ۰ به معنی عدم بستری شدن بیمار در بخش مراقبت‌های ویژه است و عدد ۱ به معنی بستری شدن بیمار خواهد بود.

واضح است که این نوع پیش‌بینی به نحوی کم کیفیت‌ترین حالت پیش‌بینی است. حال ایده‌آل این است که مدل با دیدن اطلاعات بیمار در اولین پنجره زمانی –یعنی سریع‌ترین زمان- پیش‌بینی دقیقی درباره احتمال بستری شدن بیمار در آینده انجام دهد. با این وجود در این طراحی اولیه اطلاعات تا آخرین پنجره زمانی قبل از بستری شدن بیمار برای آموزش استفاده خواهد شد تا به عنوان مرجعی برای مقایسه‌های آینده استفاده شود.

۸– ارزیابی کیفیت پیش بینی

برای ارزیابی کیفیت پیش بینی، نتایج با شاخص‌های دقت، حساسیت و امتیاز اف۱ ارائه شده‌اند.

---

[۳] hyperparameters
[۴] train
[۵] test
[۶] validation
[۷] Principal Component Analysis - PCA
[۸] logistic regression
[۹] Support Vector Machine (SVM)

۹– نتایج آماری اجرای پایپ لاین اولیه

نتایج اولین پیش‌بینی با شرایط فوق اگرچه دارای دقتی متوسط به بالا (بیش از ۷۰٪) است، مقدار بسیار پایین امتیاز اف ۱ به وضوح نشان می‌دهد که عدم تعادل در تعداد نمونه‌های بستری و غیر بستری در بخش مراقبت‌های ویژه، منجر به سوگیری طبقه‌بندی کننده به سمت تشخیص همه نمونه‌ها به عنوان سالم شده است. وزن‌گذاری اطلاعات ورودی بر اساس تعداد موارد، نتایج پیش‌بینی را به طرز قابل توجهی بهبود بخشیده که تاییدی بر این عدم تعادل است. علاوه بر آن تاثیر کاهش ابعاد نیز قابل توجه است. ارزیابی کمی نتایج پیش‌بینی‌ها در جدول ۲ قابل مشاهده است.

جدول ۲ : خروجی پایپ‌لاین اولیه، ارزیابی آماری پیشبینی سه طبقه‌بندی کننده در چهار حالت مختلف

---- Logistic regression experiments ----	---- SVM with RBF kernel experiments ----	---- Nonlinear SVM experiments ----
No balancing weights, no PCA accuracy: 0.7142857142857143 f1_score: 0.14285714285714285 balanced accuracy: 0.5096409574468085 average precision: 0.25828373015873013	No balancing weights, no PCA accuracy: 0.746031746031746 f1_score: 0.0 balanced accuracy: 0.5 average precision: 0.25396825396825395	No balancing weights, no PCA accuracy: 0.7222222222222222 f1_score: 0.3396226415094339 balanced accuracy: 0.5767952127659575 average precision: 0.3030753968253968
No balancing weights, with PCA accuracy: 0.7222222222222222 f1_score: 0.22222222222222224 balanced accuracy: 0.5355718085106382 average precision: 0.2743818681318681	No balancing weights, with PCA accuracy: 0.7380952380952381 f1_score: 0.05714285714285714 balanced accuracy: 0.5049867021276595 average precision: 0.2564484126984127	No balancing weights, with PCA accuracy: 0.6904761904761905 f1_score: 0.3157894736842105 balanced accuracy: 0.5555186170212766 average precision: 0.2837896825396825
Weighted, no PCA accuracy: 0.6904761904761905 f1_score: 0.2909090909090909 balanced accuracy: 0.5452127659574468 average precision: 0.2774327122153209	Weighted, no PCA accuracy: 0.6984126984126984 f1_score: 0.3666666666666667 balanced accuracy: 0.5814494680851063 average precision: 0.30171130952380953	weighted, no PCA accuracy: 0.7222222222222222 f1_score: 0.3396226415094339 balanced accuracy: 0.5767952127659575 average precision: 0.3030753968253968
Weighted, with PCA accuracy: 0.6746031746031746 f1_score: 0.3278688524590164 balanced accuracy: 0.555186170212766 average precision: 0.2823617952928298	weighted, with PCA accuracy: 0.6746031746031746 f1_score: 0.3492063492063492 balanced accuracy: 0.5654920212765957 average precision: 0.28864247311827956	weighted, with PCA accuracy: 0.6904761904761905 f1_score: 0.3157894736842105 balanced accuracy: 0.5555186170212766 average precision: 0.2837896825396825

یک نتیجه اولیه از پیش‌بینی مدل مرجع این است که چالش عدم تعادل نمونه‌ها احتمالاً بیشتر و جدی‌تر از پرکردن جاهای خالی در مجموعه دادگان باشد.

۱۰– مراحل آینده

علیرغم آگاهی از مشکل عدم تعادل نمونه‌ها، هدف اصلی برای پیاده‌سازی در مرحله بعد روش پیشرفته‌تری برای پرکردن جاهای خالی (ایمپیوتیشن) مجموعه دادگان است. برای این منظور الگوریتم KNNImputer و همچنین روش‌های Dimension Reductionبرای کاهش بعد ویژگی‌ها پیاده‌سازی خواهد شد. بسته به این‌که نتایج آموزش و پیشبینی بعد از استفاده از این الگوریتم چگونه باشد، می‌توان به مراحل بعدی یعنی نمونه‌افزایی نیز فکر کرد.

۱۱–منابع

1. Sírio-Libanês, H. (2020, June 22). Covid-19 - clinical data to assess diagnosis. Kaggle. Retrieved December 24, 2021, from https://www.kaggle.com/S%C3%ADrio-Libanes/covid19
2. https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
3. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
4. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html