دانشگاه صنعتی شریف

دانشکده مهندسی کامپیوتر

درس یادگیری ماشین

دکتر عباس حسینی

زمان انتشار: ۲۹ اسفند ۱۳۹۹

زمان تحویل: ۲۶ اردیبهشت ۱۴۰۰

۱- مقدمه

حالا بیش از یک سال است که با همه‌گیری ویروس کرونا از نزدیک دست و پنجه نرم کرده‌ایم. اگر کمی اخبار را دنبال کرده باشید احتمالاً به گوشتان خورده که یکی از بزرگترین چالش‌های بیمارستان‌ها در طی این دوران فراهم کردن **امکانات** لازم برای بیماران بوده است.

یکی از راه‌حل‌هایی که در برابر چالش «کمبود امکانات درمانی برای بیماران بدحال» وجود دارد، پیش‌بینی کردن حال بیماران فعلی در بیمارستان است. این که بتوانیم با استفاده از داده‌ها و علائم حیاتی بیمار در طی زمان بستری بودن روند سلامتی او را دنبال کنیم و بتوانیم در زودترین زمان ممکن پیش‌بینی کنیم که این بیمار نیاز به بستری شدن در بخش مراقبت‌های ویژه -ICU- و تبعاً امکانات بیشتر دارد.

قابلیت پیش‌بینی کردن نیاز بستری شدن بیماران در ICU از دو جهت به کارهای درمان کمک خواهد کرد: ۱- می‌توانیم روند درمان بیمار بسته به پیش‌بینی وضعیت او بهبود دهیم. ۲- نسبت به میزان امکانات مورد نیاز در بیمارستان پیش‌بینی خواهیم داشت؛ بدین ترتیب می‌توانیم امکانات را بین مراکز مختلف توزیع کنیم یا درخواست خرید/تولید امکانات جدید بدهیم.

۲- شرح مساله

ما در این پروژه با در دست داشتن اطّلاعات مختلف هر بیمار در طی زمان بستری بودنش قصد داریم تا در نهایت حدس بزنیم که **«آیا این بیمار در نهایت نیاز به انتقال به ICU خواهد داشت یا خیر؟»**.

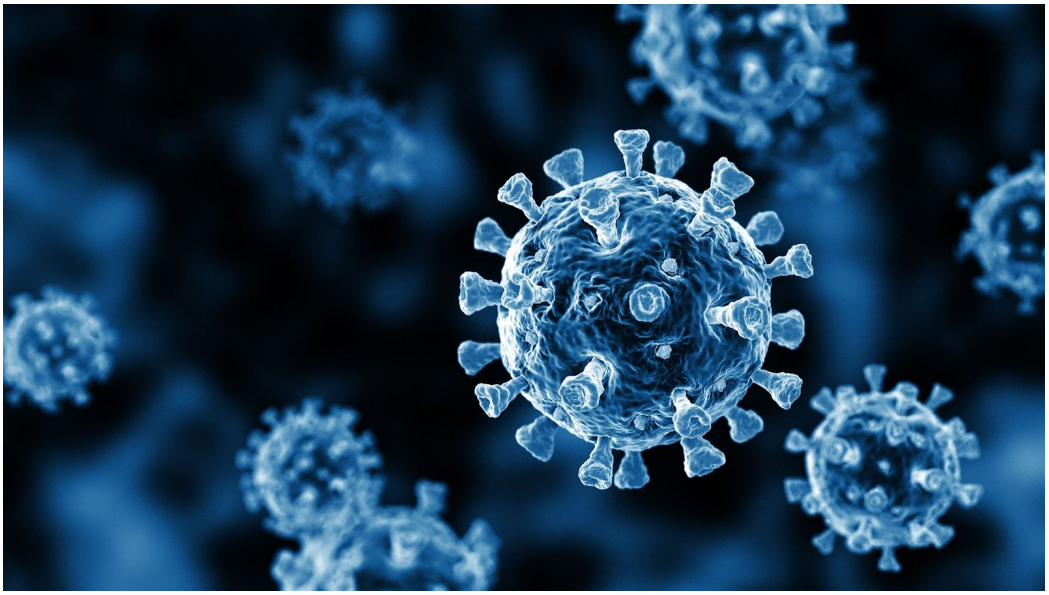
۳- توضیحات دیتاست

دیتاستی که در اختیارتان قرار داده‌ایم، مربوط به داده‌های یکی از بیمارستان‌های کشور ایتالیا است که در اوایل شروع همه‌گیری ویروس کووید جمع‌آوری‌شده و در دسترس عموم قرار داده شده‌است.

در این دیتاست هر ردیف متناظر با اطلاعات پزشکی یک بیمار است که در بیمارستان بستری شده‌است و در بازه‌ی خاصی پس از بستری بیمار، این اطلاعات از او ثبت شده‌است. این بازه‌ی خاص با ستون "window" مشخص شده‌است. به عنوان مثال $\text{window} = [2, 4]$ یعنی در بازه‌ی ۲ تا ۴ ساعت پس از بستری بیمار، این اطلاعات از او ثبت شده‌است.

در کل اطلاعات ۳۸۵ بیمار مختلف در این دیتاست جمع‌آوری شده‌است. همچنین اطلاعاتی که برای هم بیمار ثبت شده را می‌توان در ۴ دسته‌ی مختلف تقسیم کرد:

- اطلاعات جمعیتی بیمار
- بیماری‌های زمینه‌ای بیمار (با نام tags در دیتاست)
- نتایج آزمایش خون بیمار
- علائم حیاتی



۴- نحوه‌ی ارزشیابی

۴.۱- تحلیل اکتشافی داده[1] - (۲۵ + ۱۰ درصد)

تمیز کردن دیتا، کشف ویژگی‌ها، ترسیم نمودارها و به طور کلی «تحلیل کردن داده‌ها» یکی از مهم‌ترین (و شاید مهم‌ترین) بخش انجام پروژه‌های یادگیری ماشین است. در این قسمت از شما انتظار داریم تا جایی که ممکن است بتوانید شواهد مفیدی از دیتا برای مراحل بعدی ایجاد کنید تا بتوانید ویژگی‌های مهم را استخراج کنید و مدل بهتری را طراحی کنید.

در این قسمت همچنین مهم است که علاوه بر فراهم کردن شواهد دیتایی بتوانید برداشت خود را از شواهد بیان کنید و بگویید که این شواهد چه کمکی به شما در اخذ نتایج بهتر و شناخت داده کرده.

برای اینکه دقیقا متوجه شوید در این مرحله چه کارهایی می‌توانید انجام دهید و انجام این کار چگونه به شما کمک خواهد کرد لینک‌های شماره‌ی ۲ و ۳ را مطالعه کنید.

۴.۲- مهندسی ویژگی‌ها[2] - (۱۵ + ۵ درصد)

پس از بررسی کردن داده و به دست آوردن شواهد از روی دیتا و کاستی‌های آن (مانند وجود مقدار زیادی خانه‌ی خالی در داده) باید بتوانید با در نظر گرفتن اطلاعاتی که از قسمت قبل به دست آوردید، دیتاست اولیه را «ماشین فهم‌تر» کنید تا داده برای شروع الگوریتم یادگیری ماشین آماده شده باشد.

برای مثال می‌توانید ستون‌هایی را از داده حذف کنید (چرا؟)، خانه‌های بدون مقدار را با روش‌های مختلف مقداردهی کنید، مقادیر را نرمالایز کنید و...

برای اینکه متوجه شوید در این مرحله چه کارهایی می‌توانید انجام دهید و انجام این کار چگونه به شما کمک خواهد کرد می‌توانید لینک شماره‌ی ۴ را مطالعه کنید.

---

[1] Exploratory Data Analysis (EDA)
[2] Feature Engineering

۴.۳- تست مدل‌های مختلف - (۱۵ + ۵ درصد)

در نهایت قصد داریم تا برای حل مساله‌ی «پیش‌بینی نیاز به بستری شدن بیماران در ICU» یک مدل داشته باشیم، ولی پیش از آن بهتر است کارای چند مدل را برای حل این مساله به طور خاص سنجیده باشیم. در این مرحله از شما خواسته می‌شود تا مسیری که بعد از «نهایی کردن دیتاست» تا «انتخاب مدل نهایی» طی می‌کنید را گزارش کنید. نتایجی که روی مدل‌های مختلف می‌گیرید، دلایلی که آن‌ها را انتخاب نمی‌کنید در این بخش اهمیت دارند.

۴.۴- نتایج مدل نهایی - (۱۵ + ۱۰ درصد)

در آخرین گام حل مساله نیاز داریم تا یک مدل را به عنوان مدل نهایی انتخاب کرده و آن را به عنوان محصول خروجی در نظر بگیریم.

در این مرحله از شما انتظار داریم تا:

- دلیل انتخاب مدل را بیان کنید.
- در حد خلاصه نحوه‌ی کارکرد آن را توضیح بدهید.
- همچنین دقت کنید که متریک اصلی برای ارزیابی مدل، F1-Score می‌باشد. هرجا نیاز شد عملکرد مدل را ارزیابی کنید از این معیار استفاده کنید. اما دقت کنید که تنها متریک مجاز F1-Score نیست بلکه بر حسب چارچوب مسئله‌ای که در آن قرار دارید می‌توانید از متریک‌های دیگر نیز استفاده کنید. اما حتما دلایل لزوم استفاده از آن را در فایل مستند شرح دهید.
- همچنین توصیه می‌شود که که بررسی کنید مدل شما در چه سناریوهایی بهتر عمل می‌کند و در چه سناریوهایی ضعف دارد.
- بررسی میزان اهمیتی که مدل به هرکدام از فیچرهای ورودی می‌دهد نیز حائز امتیاز خواهد بود.
- همچنین دقت کنید که اگر مدل شما بتواند بیماری را که قرار است در نهایت در ICU بستری شود، زودتر پیش‌بینی کند کار بسیار ارزشمندی انجام داده‌است زیرا می‌توان از قبل آمادگی‌های لازم را کسب کرد و اقدامات مورد نیاز را انجام داد. بنابراین بخش زیادی از نمره‌ی امتیازی این قسمت این است که مدل شما توانایی پیش‌بینی زودهنگام (early prediction) داشته باشد. ایده‌هایی که برای حل این قسمت دارید را حتما در مستند توضیح دهید.

- همچنین دقت کنید که هر چه مدل ساده‌تر باشد ارزشمندتر است، به عنوان مثال دو مدلی که عملکرد یکسان دارند مدلی ارزشمندتر است که از پیچیدگی کمتری برخوردار باشد.

۴.۵- گزارش - (۲۰ + ۱۰ درصد)

ارائه کردن پروژه‌ی ماشین لرنینگ یکی از مهم‌ترین مهارت‌های نرمی است که یک متخصص ماشین لرنینگ باید بتواند به خوبی آن را انجام بدهد، چنان که توانایی Data Storytelling یکی از مهم‌ترین ویژگی‌هایی که یک دانشمند داده باید در سال ۲۰۲۱ داشته باشد تلقی شده.

لذا یکی از اصلی‌ترین قسمت‌های حل یک مسئله در حوزه‌ی علوم داده، نوشتن یک مستند مناسب پس از حل مسئله است. شما در این مستند باید به صورت ساده و البته خلاصه، مسیری که در روند حل مسئله طی‌کرده‌اید و چالش‌هایی که با آن‌ها روبرو شدید را توضیح دهید و ایده‌هایی که برای حل این چالش‌ها استفاده کرده‌اید را نیز توضیح دهید.

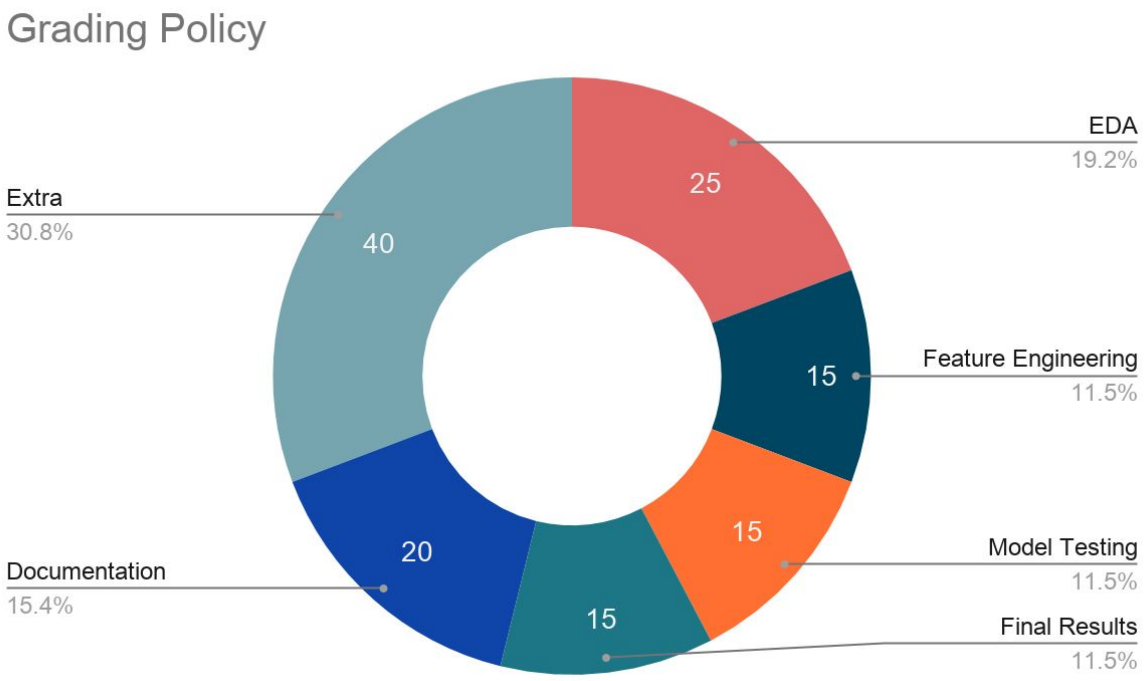
در فرایند توضیح دادن روش حل‌تان، هر چه از مصورسازی‌ها و مثال‌های مناسب‌تر استفاده کرده باشید ارزشمندتر است. به‌طور کلی روند Data Storytelling باید در مستند مشهود باشد.

اگر هر حرفی در مورد آماره‌ی دیتاست دارید حتما باید همراه با مصورسازی باشد و نوع نموداری که انتخاب می‌کنید باید با مسئله‌ی متناظر با آن بخواند به‌طور مثال اگر قصد دارید توزیع یک کمیت را بررسی کنید نمودار هیستوگرام برای منظور شما مناسب است.

مهم است که مقاله‌ها و منابعی که در طی مسیر حل مساله از آن‌ها استفاده می‌کنید را حتماً ذکر کنید.

همچنین دقت کنید که شاید لازم شود پروژه‌های خود را ارائه نیز بکنید. در مورد این موضوع بعدا کامل‌تر اطلاع‌رسانی خواهد شد.

نکته‌ی مهم: تحویل دادن کد یا صرف ارائه‌ی نتایج بدون ارائه کردن مستند نمره‌ای در بر نخواهد داشت.



شکل ۲: نحوه‌ی توزیع نمرات در نمودار بالا قابل مشاهده است. مجموع نمرات پروژه ۱۳۰ است و کسب ۱۰۰ نمره از آن به منزله‌ی نمره‌ی کامل خواهد بود. توجه کنید که ۴۰ نمره‌ی پروژه مربوط به بالاتر از حد انتظار انجام دادن بخش‌های مختلف پروژه خواهد بود و شما با در حد انتظار انجام دادن تمامی بخش‌ها ۹۰٪ نمره را دریافت خواهید کرد.

۵- چند نکته

- یکی از اهداف این پروژه تقویت مهارت یادگیری شما بسته به نیاز مساله است؛ بنابراین سرچ کردن در منابع مختلف و انجام مطالعات و دیدن مثال‌های مشابه بسیار مورد استقبال قرار خواهد گرفت. از شما می‌خواهیم تا منابعی که مورد مطالعه قرار می‌دهید را در مستند خود بیاورید.
- برای هر مساله در مورد زبان برنامه‌نویسی و ابزارهای مورد استفاده **هیچ محدودیتی** وجود ندارد. اگر چه استفاده از زبان پایتون توصیه می‌شود.
- فاز یک پروژه به صورت تک‌نفره است.
- مهلت تحویل پروژه تمدید نخواهد شد.
- خروجی‌های مورد نیاز پروژه: ۱- کد ۲- مستند توضیح
  - استفاده از خروجی jupyter notebook به دلیل اینکه خروجی‌های مورد نیاز را به صورت یک‌پارچه قابل ارائه می‌کند توصیه می‌شود.

- دقت کنید که تمامی خروجی‌های پروژه به عنوان دارایی معنوی[3] شما تلقی خواهد شد. تیم تدریس نتایج ارائه شده توسط شما را صحت‌سنجی خواهند کرد. لذا شباهت کد شما با دیگر دانشجویان، کپی صرف از منابع و برداشت بدون ذکر منبع پس از بررسی به عنوان تخلف آموزشی مورد پیگرد قرار خواهد گرفت و نمره‌ای در پی نخواهد داشت.
- در مورد پرسش و پاسخ:
  - در صورتی که مشکل یا سوالی در روند پروژه داشتید می‌توانید آن را در پیاتزای درس بپرسید. (این روش توصیه می‌شود.)
  - همچنین می‌توانید مستقیما با تیم پروژه‌ی درس: اسرا کاشانی‌نیا، پویا معینی، امین روانبخش و آرمین مرادی ارتباط بگیرید.
  - در صورتی که سوالاتتان مربوط به ابعاد غیرعلمی پروژه (مانند سیاست نمره‌دهی) است آن را از آرمین مرادی -مسئول پروژه- بپرسید.

۶- لینک‌های مفید

لینک ۱: مراحل انجام پروژه‌ی ML

لینک ۲: اهمیت EDA

لینک ۳: چگونگی EDA

لینک ۴: اهمیت و چرایی Feature Engineering

لینک ۵: اهمیت Data Storytelling

در ضمن امیدواریم سال جدید خیلی خوبی داشته باشید:)

[3] Intellectual Property