



پروژه یادگیری ماشین
کمپ کارآموزی دیتاساینس نشان

زمان انتشار پروژه: 06/24

زمان تحویل پروژه: 06/28

مقدمات

این پروژه بر اساس یکی از چالش‌های اصلی و حیاتی در دوران ویروس کرونا از ابتدای شیوع جهانی آن شکل گرفته است. این ویروس باعث شد که تعداد زیادی از افراد نیازمند بستری شدن در بیمارستان‌ها باشند. در دوران قرنطینه و محدودیت‌های اجتماعی، بیماران مبتلا به کرونا به مراقبت‌های ویژه و مداوم نیاز داشتند. یکی از راه‌های مؤثر برای کنترل و مدیریت این بیماری، پیش‌بینی نیاز داشتن/نداشتن بیماران مبتلا به کرونا به بستری شدن در بخش مراقبت‌های ویژه، با توجه به داده‌ها و علائم حیاتی بیمار در طول زمان بستری است. چنین پیش‌بینی‌هایی می‌توانند به کادر پزشکی کمک کنند تا تصمیمات بهتری در مورد تخصیص منابع و ارائه خدمات بهداشتی اتخاذ کنند و اطمینان حاصل شود که امکانات بیمارستانی به شکل بهینه مورد استفاده قرار می‌گیرند.

در این پروژه با استفاده از داده‌های بالینی بیماران مبتلا به کرونا که در بخش‌های مختلف بیمارستان بستری شده‌اند، تلاش می‌کنید تا مدلی برای پیش‌بینی نیاز به بستری شدن در بخش مراقبت‌های ویژه توسعه دهید. این مدل می‌تواند به کاهش فشار بر سیستم بهداشتی، بهبود مدیریت منابع و کاهش میزان مرگ‌ومیر ناشی از این بیماری کمک کند.

هدف

پاسخ به این سوال و یافتن راه حل آن: آیا می‌توان با استفاده از داده‌های بالینی بیماران مبتلا به کرونا نیاز آن‌ها به بستری شدن/نشدن در بخش مراقبت‌های ویژه را به‌دقت پیش‌بینی کرد؟

مجموعه داده‌ها

داده‌های مورد استفاده در این پروژه از منابع معتبر و مرتبط با بیماران مبتلا به کرونا جمع‌آوری شده‌اند. این داده‌ها به صورت ناشناس و با حفظ حریم خصوصی بیماران گردآوری شده و شامل اطلاعات جامعی از وضعیت بالینی آن‌ها هستند.

مجموعه داده‌ها شامل موارد زیر است:

- نتایج آزمایشات خون
- مواردی نظیر heart rate, blood pressure, temperature
- علائم حیاتی
- بیماری‌های زمینه‌ای
- اطلاعاتی نظیر سن و جنسیت
- زمان ثبت اطلاعات (observation window)
- و ...

روش ارزیابی

در این فرآیند همه موارد زیر در نظر گرفته می‌شوند:

- تحلیل اکتشافی داده‌ها (EDA)
- مهندسی ویژگی‌ها (Feature Engineering)
- بررسی امکان استفاده از مدل‌های مختلف و انتخاب مدل نهایی (Model Selection)
- نتایج مدل نهایی
- گزارش نهایی (Documentation)
- نوشتن سرویس Rest و ارائه در قالب ایمپج داکر
- کار با mlflow جهت track کردن experiment‌های مختلف (شامل نمره اضافه)
- Onnx کردن مدل و استفاده از تکنیک‌های optimization (شامل نمره اضافه)

- تحلیل اکتشافی داده‌ها (EDA)
 - در این مرحله، داده‌ها به دقت مورد بررسی و تحلیل قرار می‌گیرند تا الگوها، روابط و ناهنجاری‌های موجود شناسایی شوند. اقدامات انجام‌شده در این بخش عبارت‌اند از:
 - تمیزسازی داده‌ها:
 - مدیریت داده‌های Missing
 - تصحیح مقادیر نادرست
 - مقیاس‌بندی و نرمال‌سازی:
 - تنظیم مقیاس داده‌ها برای تضمین عملکرد بهتر الگوریتم‌های یادگیری ماشین. (در صورت نیاز)
 - تحلیل توصیفی:
 - بررسی توزیع متغیرها
 - میانگین‌ها، میانه‌ها و سایر شاخص‌های آماری
 - بررسی balanced یا imbalanced بود دیتاست در ویژگی‌های مختلف
 - شناسایی الگوها:
 - کشف روابط میان متغیرهای مختلف
 - تشخیص عوامل مؤثر بر نیاز به بستری شدن در بخش مراقبت‌های ویژه
 - بصری‌سازی داده‌ها:
 - استفاده از نمودارها برای درک بهتر ساختار و توزیع داده‌ها
 - کشف روابط میان ویژگی‌ها
 - (تحلیل دلیل انتخاب روش‌های مورد استفاده در همه بخش‌ها، در گزارش آورده شود)
- مهندسی ویژگی‌ها
 - در این مرحله، ویژگی‌های مهم و مرتبط با هدف تحقیق استخراج و ساخته می‌شوند تا عملکرد مدل بهبود یابد. اقدامات کلیدی شامل:
 - انتخاب ویژگی‌ها:
 - شناسایی و انتخاب متغیرهایی که بیشترین تاثیر را بر پیش‌بینی دارند.
 - ایجاد ویژگی‌های جدید
 - ترکیب و تبدیل متغیرهای موجود برای ساخت ویژگی‌های مفیدتر
 - بررسی داده‌های دسته‌ای:
 - تبدیل داده‌های کیفی به کمی
 - بررسی روش‌ها و نحوه درست تبدیل

- ساخت مدل اولیه (base model)

در این بخش، مدل‌های یادگیری ماشین برای پیش‌بینی نیاز به بستری شدن در بخش خاص توسعه و آموزش داده می‌شوند. مراحل انجام‌شده عبارت‌اند از:

- تقسیم‌بندی داده‌ها: جداسازی داده‌ها به مجموعه‌های آموزشی و آزمایشی برای ارزیابی عملکرد مدل.
- انتخاب الگوریتم مناسب: بررسی و انتخاب الگوریتم مناسب.
- آموزش مدل: استفاده از داده‌های آموزشی برای یادگیری الگوها و روابط موجود.
- پیش‌بینی اولیه: ارزیابی مدل بر روی داده‌های آزمایشی و بررسی دقت اولیه پیش‌بینی‌ها.

پس از ساخت مدل اولیه، فرآیند بهینه‌سازی برای بهبود دقت و کارایی مدل انجام می‌شود. اقدامات شامل:

- تنظیم هایپر پارامترها: استفاده از روش‌هایی مانند Grid Search و Random Search برای یافتن بهترین مقادیر هایپر پارامترها.
- اعتبارسنجی: به کمک روش‌هایی نظیر Cross-Validation و ارزیابی مدل در بخش‌های مختلف داده‌ها برای تضمین پایداری و قابلیت تعمیم.
- مقابله با overfitting
- ارزیابی مدل: استفاده از معیار F1-Score برای سنجش عملکرد مدل و انتخاب بهترین مدل. (متریک اصلی f1 هست، اما دقت کنید f1 تنها متریک مجاز نیست)
- Error Analysis: بررسی اینکه مدل در چه مواردی نقطه ضعف دارد و چرا؟

- گزارش نهایی

در نهایت، تمامی نتایج و یافته‌های پروژه در قالب یک گزارش جامع ارائه می‌شوند که شامل موارد زیر است:

- مقدمه و بیان مسئله
- روش‌شناسی و مراحل انجام کار
- نتایج تحلیل‌ها و مدل‌سازی
- بحث و تفسیر نتایج
- نتیجه‌گیری و پیشنهادات برای کارهای آینده
- منابع و مراجع استفاده‌شده

در نظر داشته باشید علت هر انتخاب و تحلیلی در طول انجام پروژه باید در مستندات آورده شود.

در صورت استفاده از روش‌های مختلف در هر مرحله از انجام پروژه مشخص شود که چرا این روش و این تکنیک مورد استفاده قرار گرفته؟

برخی از موارد در گزارش نهایی به عنوان مثال:

- تکنیک‌های مورد استفاده در پیش پردازش داده‌ها

- تحلیل نمودارهایی که در بخش EDA بررسی میکنید.
- راه حل های مورد استفاده در بخش انتخاب یا گسترش ویژگی ها
- نتایج مدل نهایی
- دلیل انتخاب مدل
- نحوه کارکرد مدل
- تاثیر هر فیچر در عملکرد مدل
- و ...

همچنین در این فرآیند ارائه پروژه به صورت داکيومنت، مطلوب است که چالش ها و مسائل و راه حل هایی که در طول مسیر با آن مواجه شدید را بیان کنید. استفاده هر چه بیشتر و بهتر از مصورسازی این فرآیند data storytelling را موفق تر میسازد.

****توجه (مهم):** موارد ذکر شده در صورت پروژه و در بخش "ارزیابی" به عنوان یک **baseline** و جهت دادن دید و ایده بهتر به شما و طبقه بندی بهتر ساختار ذهنی جهت پیشبرد پروژه آورده شده اند. در برخی قسمت ها شاید نیازی به انجام برخی مراحل به صلاحدید خود نداشته باشید و در سوی دیگر شاید نیاز باشد موارد و تکنیکهای دیگری نیز بررسی شوند. پس در نظر داشته باشید با توجه به روش انتخابی و نیاز خود مراحل را پیش برده و روش ها و تکنیک های مناسب را اتخاذ کنید و به کار ببرید.

به عنوان مثال شاید اصلا نیازی به ایجاد فیچر جدید نبینید و یا لزومی برای **encode** کردن فیچر **categorical** وجود نداشته باشد! یا اینکه نیازی به **grid search** نباشد. اما در نهایت انتخابات خود را با دلیل در داکيومنت نهایی ارائه دهید.

ایمیل های مرتبط جهت برقراری ارتباط و رفع ابهامات:

m.amin.ce.sbu@gmail.com

mehran.sfi16@gmail.com

ارتباط در تلگرام:

https://t.me/FR_MA1396

https://t.me/Mehran_Seyfi16

موفق باشید.