گزارش نهایی:

تحلیل و پیش‌بینی بیماری قلبی بر اساس دیتاست Heart Disease Prediction

تهیه‌کننده: آرزو راستکار

تاریخ: ۱۳ دسامبر ۲۰۲۵

هدف پروژه: بررسی دیتاست ۲۷۰ نمونه‌ای بیماران قلبی، شناسایی عوامل خطر، و ساخت مدل classification برای پیش‌بینی Presence/Absence بیماری قلبی با دقت بالا (تقریبا ٪۸۵).

ابزارها: Python (Pandas, Scikit-learn, Seaborn, Matplotlib)

دیتاست: ۱۳ ویژگی دموگرافیک/بالینی

(Age, Sex, Chest pain type, BP, Cholesterol, FBS over 120, EKG results, Max HR, Exercise angina, ST depression, Slope of ST, Number of vessels fluro, Thallium) + هدف.(Heart Disease: Absence/Presence)

۱ .مقدمه و بررسی اولیه داده

دیتاست شامل ۲۷۰ بیمار (Absence ۵۶٪ ، Presence۴۴٪) هست – کمی imbalance ، اما باstratify مدیریت شد. هیچ missing value ای نداره (تمیز).

آمار توصیفی کلیدی : (از describe())

ویژگی	میانگین	انحراف استاندارد	حداقل	حداکثر
Age	۵۴/۴۳	۹/۱۱	۲۹	۷۷
Sex (۰=مرد،۱=زن)	۰/۶۸	۰/۴۷	۰	۱
Chest pain type	۳/۱۶	۰/۹۲	۱	۴
BP	۱۳۰/۳۴	۱۷/۰۶	۹۴	۲۰۰
Cholesterol	۲۴۹/۶۶	۵۱/۸۳	۱۴۹	۵۶۴
Max HR	۱۵۰/۱۸	۲۲/۹۰	۷۱	۲۰۲
ST depression	۱/۰۵۰	۱/۱۶	۰	۶/۲
Thallium	۴/۷۰	۱/۹۴	۳	۷

مشاهدات :مردان (Sex = 1) بیشتر درگیرن؛ کلسترول بیشتر از ۲۵۰ در ٪۵۰ Presence دیده می‌شه .

۲. تحلیل اکتشافی(EDA)

-توزیع کلاس: Absence۱۵۰ (٪۵۶)، Presence ۱۲۰ (٪۴۴)

-همبستگی با هدف (باینری): Thallium (۰,۵۲۵، قوی مثبت – عامل خطر عروقی) ، Max HR (-۰,۴۱۹، محافظ – ضربان پایین خطرناک).

- جدول همبستگی کلیدی:

ویژگی	همبستگی
Thallium	۰/۵۲۵
Number of vessels fluro	۰/۴۵۵
Exercise angina	۰/۴۱۹
ST depression	۰/۴۱۸
Chest pain type	۰/۴۱۷
Max HR	-۰/۴۱۹

-Heatmap: Visualization همبستگی رسم شد – هیچ multicollinearity شدید (<۰,۸) دیده نشد.

-تفسیر: تمرکز روی Thallium > 5 وST depression > 1 برای غربالگری.

۳ .پیش‌پردازش

- Split: ۲۰/۸۰ (Train: ۲۱۶، Test: ۵۴) با stratify (تعادل ٪۵۶ Absence در train).

-Scaling : StandardScaler روی train (میانگین ۰، واریانس ۱)- برای تعدیل واریانس (مثل Cholesterol)

۴ .مدل‌سازی و ارزیابی

دو مدل Logistic Regression : baseline (LR، با scaling) و Random Forest (RF، بدون).

-نتایج:Baseline

مدل	دقت (Test)	Recall Presence	F1 Presence
LR	۸۵/۱۹٪	۹۲٪	۰/۸۵
RF	۸۱/۴۸٪	۸۳٪	۰/۸۰

-اهمیت ویژگی‌ها (از RF):

رتبه	ویژگی	اهمیت
۱	Chest pain type	۰/۱۳۰
۲	Max HR	۰/۱۲۰
۳	ST depression	۰/۱۱۰
۴	Thallium	۰/۱۰۰

-تفسیر: LR برتره (ساده، interpretable) ؛ RF برای feature selection عالی. هر دو overfitting ندارن.

۵. Tuning مدل

- GridSearchCV برای LR (cv=5): بهترین {'C': 0.1, 'penalty': 'l2'} - قوی regularization.

-دقتCV : ٪۸۴,۲۶(stable)

-دقتTuned Test : ٪۸۵,۱۹ (بدون افت)

-جدولCV Scores

C	CV دقت l2
۰/۱	٪۸۴/۲۶
۱	٪۸۲/۸۸
۱۰	٪۸۱/۹۶

۶ .نتیجه‌گیری و توصیه‌ها

-بهترین مدل: Tuned LR (٪۸۵ دقت، ٪۹۲Recall Presence) – مناسب بیمارستان برای غربالگری.

-عوامل کلیدی خطر: Thallium بالا، Chest pain type = ۴، ST depression (شدید).

-محدودیت‌ها: دیتاست کوچک (۲۷۰) – نیاز به داده بیشتر؛ imbalance کم، اما SMOTE می‌تونه کمک کنه.

-توصیه‌ها:

- تمرکز روی بیماران >۵۵ سال با Max HR >۱۴۰ و کلسترول >۳۰۰

تحلیل جامع بر اساس نمودارهای رسم‌شده: پیش‌بینی و عوامل خطر بیماری قلبی

در این تحلیل، با استفاده از دیتاست Heart Disease Prediction شامل ۲۷۰ بیمار و ۱۴ ویژگی دموگرافیک و بالینی، روابط، توزیع‌ها و عملکرد مدل‌های پیش‌بینی را از طریق visualizationهای متنوع بررسی کردیم. ماتریس همبستگی (Heatmap) نشان می‌دهد که ویژگی‌هایی مانند Thallium (با همبستگی ۰,۵۲) و ST depression (۰,۴۲) بیشترین ارتباط مثبت با وجود بیماری قلبی (Presence) دارند، در حالی که حداکثر ضربان قلب (Max HR) با همبستگی منفی -۰,۴۲ به عنوان عامل محافظ عمل می‌کند؛ این الگوها حاکی از اهمیت مشکلات عروقی و الکتریکی قلب در پیش‌بینی است، بدون وجود multicollinearity شدید (همبستگی‌های بالای ۰,۸ نادر است).

Histogram کلی داده‌ها (data.hist) توزیع ویژگی‌های عددی را آشکار می‌سازد، جایی که سن (Age) توزیع نرمالی با میانگین ۵۴ سال دارد، اما کلسترول (Cholesterol) skew راست نشان می‌دهد (با مقادیر بالا در ۲۰٪ موارد)، که با boxplot BP و Cholesterol همخوانی دارد: در گروه Presence، میانه کلسترول به ۲۷۰ می‌رسد (بالاتر از ۲۴۰ در Absence) و outliers (خارج از IQR) در کلسترول ≈۲۰ مورد را تشکیل می‌دهند، که می‌تواند نشان‌دهنده بیماران پرخطر با سطوح غیرطبیعی باشد.

Histogram Age با KDE(کرنل density estimation) تأیید می‌کند که توزیع سن پیک در ۶۰-۵۰ سال دارد و KDE منحنی صاف نرمال را ترسیم می‌کند، که با 3D Scatter (Age-Cholesterol-BP) هم‌راستاست: در این visualization سه‌بعدی، نقاط قرمز (Presence) عمدتاً در ناحیه سن بالا، کلسترول >۳۰۰ و BP >۱۴۰ cluster می‌شوند، در حالی که نقاط آبی (Absence) پراکندگی گسترده‌تری نشان می‌دهند، و این الگو بر نقش سن و کلسترول به عنوان عوامل تجمعی خطر تأکید دارد.

Pairplot کلی روابط pairwise را برجسته می‌کند، جایی که scatterplot Max HR vs Thallium نشان‌دهنده جداسازی واضح کلاس‌ها است (Presence با Thallium بالا و Max HR پایین)، و این با scatterplot KMeans (۳ کلاستر، inertia ≈ k ۴۳۱-) همخوانی دارد: کلاسترها بر اساس سن و کلسترول جدا می‌شوند، با کلاستر قرمز (Presence) در مقادیر بالا، که پیشنهاد می‌کند KMeans می‌تواند برای unsupervised grouping بیماران پرخطر مفید باشد.

Bar-plot توزیع کلاس imbalance (countplot) را آشکار می‌کند (۱۵۰ Absence vs ۱۲۰ Presence، ٪۵۶ به ٪۴۴)، و countplot Heart Disease by Sex نشان می‌دهد که مردان (Sex=۱) ٪۱۷۰ موارد Presence را تشکیل می‌دهند (در مقابل ٪۵۰ در زنان)، که heatmap cross-tab (Sex vs Disease) با مقادیر ۵۰/۷۰ تأیید می‌کند و بر نابرابری جنسیتی در خطر بیماری تأکید دارد – این یافته‌ها حاکی از نیاز به غربالگری جنسیت‌محور است.

در بخش مدل‌سازی، line plot KNN scores(K=۱-۲۰) دقت را به عنوان تابعی از همسایگان نشان می‌دهد، با پیک ٪۸۹ در K=۱۸، که نشان‌دهنده بهینه‌سازی hyperparameter برای KNN است و overfitting در Kهای پایین (به دلیل نویز) را برجسته می‌کند. Bar plot SVC kernels مقایسه kernelها را ترسیم می‌کند، جایی که linear با ٪۹۱ دقت برتر است (poly٪۸۸، RBF٪۸۸، sigmoid٪۸۵)، و این برتری به دلیل روابط خطی قوی در داده (مانند همبستگی Thallium) است.

Line plot Decision Tree max_features دقت را بر اساس تعداد ویژگی‌ها نشان می‌دهد، با بهبود از ٪۷۰ در max_features=۱ به ٪۸۵ در max_features=۲، که حاکی از سادگی مدل DT در این دیتاست است و نیاز به pruning برای جلوگیری از overfitting را پیشنهاد می‌کند. Bar plot Random Forest estimators دقت را به عنوان تابعی از تعداد درخت‌ها (۱۰۰۰-۱۰) ترسیم می‌کند، با تثبیت در ٪۸۷ برای ۲۰۰+ درخت، که نشان‌دهنده ensemble learning به عنوان روش robust برای مدیریت نویز (مانند outliers کلسترول) است.

در مجموع، این visualizationها الگویی جامع را آشکار می‌سازند: عوامل خطر اصلی (Thallium بالا، کلسترول >۳۰۰، سن >۵۵، جنسیت مرد) با دقت ٪۹۱ توسط SVC linear پیش‌بینی می‌شوند، در حالی که clustering KMeans و imbalance جنسیتی بر لزوم مداخلات هدفمند (مانند غربالگری مردان بالای ۵۰ سال) تأکید دارند؛ این یافته‌ها نه تنها دقت مدل‌ها را (٪۹۱-۸۵) تأیید می‌کنند، بلکه کاربرد بالینی را برای تشخیص زودهنگام برجسته می‌سازند، با پتانسیل بهبود از طریق SMOTE برای تعادل کلاس‌ها.