یکپارچه‌سازی داده‌های بیماری قلبی و طبقه‌بندی با الگوریتم KNN

تکلیف درس داده‌کاوی

محمود صدریان
شماره دانشجویی: ۴۰۴۳۵۳۳۶

Repository GitHub

۲۴ آبان ۱۴۰۴

چکیده

این گزارش یک مطالعهٔ داده‌کاوی بر روی مجموعه‌دادهٔ بیماری قلبی از مخزن UCI است و هدف آن پیش‌بینی وجود بیماری قلبی بر اساس ویژگی‌های بالینی بیماران با استفاده از الگوریتم KNN می‌باشد. داده‌ها پیش‌پردازش شده‌اند (جایگزینی مقادیر گمشده، نرمال‌سازی و تقسیم آموزش–آزمون) و تحلیل اکتشافی داده‌ها (EDA) شامل هیستوگرام‌ها، نمودارهای جعبه‌ای، نمودارهای احتمال–احتمال، ماتریس همبستگی، نمودارهای چندمتغیره (Pair Plot) و نمودارهای شمارشی انجام شده است. طبقه‌بند KNN با استفاده از تابع فاصلهٔ ترکیبی پیاده‌سازی شده و عملکرد آن با معیارهایی مانند دقت، بازخوانی، دقت مثبت، امتیاز F۱ و ماتریس اغتشاش ارزیابی گردیده است.

۱ مقدمه

بیماری‌های قلبی عروقی همچنان یکی از مهم‌ترین علل مرگ‌ومیر در سطح جهان به شمار می‌روند. تشخیص زودهنگام و برآورد صحیح ریسک ابتلا، نقش کلیدی در پیشگیری از عوارض شدید و بهبود کیفیت زندگی بیماران دارد. در سال‌های اخیر، روش‌های مبتنی بر داده و تکنیک‌های یادگیری ماشین به ابزارهایی بسیار مهم برای تحلیل داده‌های پزشکی و پشتیبانی از تصمیم‌گیری بالینی تبدیل شده‌اند.

در این پروژه از تکنیک‌های داده‌کاوی بر روی مجموعه‌دادهٔ شناخته‌شدهٔ بیماری قلبی از مخزن UCI Machine Learning Repository استفاده می‌شود. هدف اصلی این است که بررسی شود چگونه ویژگی‌های مختلف بیماران، مانند سن، جنسیت، نوع درد قفسهٔ سینه، سطح کلسترول، فشار خون استراحتی و حداکثر ضربان قلب، می‌توانند برای پیش‌بینی وجود بیماری قلبی به کار گرفته شوند.

تمرکز این مطالعه بر الگوریتم طبقه‌بندی K-Nearest Neighbors (KNN) است که یک روش ساده اما قدرتمند مبتنی بر فاصله محسوب می‌شود. مراحل اصلی پروژه شامل پیش‌پردازش داده‌ها، تحلیل اکتشافی داده‌ها (EDA)، پیاده‌سازی طبقه‌بند KNN برای ویژگی‌های ترکیبی عددی و دسته‌ای، و ارزیابی عملکرد مدل با استفاده از معیارهای استاندارد طبقه‌بندی است. نتایج این مطالعه، توان بالقوهٔ الگوریتم KNN در مسائل طبقه‌بندی پزشکی را نشان می‌دهد و اهمیت پیش‌پردازش مناسب و تحلیل دقیق ویژگی‌ها را در فرایند داده‌کاوی برجسته می‌سازد.

۲ توصیف مجموعه‌داده

مجموعه‌دادهٔ مورد استفاده در این پروژه، مجموعه‌دادهٔ بیماری قلبی از مخزن UCI Machine Learning Repository [۱] است. این داده‌ها شامل اطلاعات بیماران جمع‌آوری‌شده از چهار مرکز پزشکی مختلف شامل کلیولند، مجارستان، سوئیس و مرکز VA Long Beach می‌باشد. هر یک از این زیرمجموعه‌ها شامل ویژگی‌های بالینی مشابهی هستند که وضعیت سلامت بیماران و وجود یا عدم وجود بیماری قلبی را توصیف می‌کنند.

در این پروژه از نسخهٔ تجمیع‌شدهٔ مجموعه‌داده استفاده شده است که شامل حدود **۱۰۰۰** نمونه و **۱۳** ویژگی اصلی است. علاوه بر این، یک متغیر هدف دودویی وجود دارد که نشان می‌دهد آیا بیمار به بیماری قلبی مبتلا است (مقدار 1) یا خیر (مقدار 0). ویژگی‌های موجود را می‌توان در سه گروه اصلی طبقه‌بندی کرد:

- **ویژگی‌های عددی پیوسته**، مانند:
  - age: سن بر حسب سال
  - trestbps: فشار خون استراحت (بر حسب میلی‌متر جیوه)
  - chol: میزان کلسترول سرم (بر حسب میلی‌گرم بر دسی‌لیتر)
  - thalach: بیشینهٔ ضربان قلب ثبت‌شده
  - oldpeak: میزان افت قطعهٔ ST ناشی از ورزش
- **ویژگی‌های دودویی یا دسته‌ای رمزگذاری‌شده**:
  - sex: جنسیت (۰ = زن، ۱ = مرد)
  - cp: نوع درد قفسهٔ سینه (۰ تا ۳)
  - fbs: قند خون ناشتا بالاتر از ۱۲۰ (۱ = بله، ۰ = خیر)
  - restecg: نتیجهٔ نوار قلب استراحتی
  - exang: آنژین ناشی از ورزش (۱ = بله، ۰ = خیر)
  - slope: شیب قطعهٔ ST در اوج فعالیت
  - ca: تعداد عروق اصلی (۰ تا ۳) مشاهده‌شده در فلوروسکوپی
  - thal: وضعیت تالاسمی (مقادیر رمزگذاری‌شده ۰ تا ۳)

- **متغیر هدف**:
  - target: وجود بیماری قلبی (۱ = وجود بیماری، ۰ = عدم وجود بیماری)

در پیاده‌سازی، مجموعه‌داده به صورت فایل CSV و با استفاده از کتابخانهٔ pandas بارگذاری شده است. بررسی اولیهٔ داده‌ها شامل مشاهدهٔ ابعاد مجموعه‌داده، نوع ستون‌ها، و مقدار داده‌های گمشده انجام گرفته تا اطمینان حاصل شود که داده به‌درستی ادغام شده و برای مرحلهٔ پیش‌پردازش آماده است.

۳ پیش‌پردازش داده‌ها

پیش از به‌کارگیری هر الگوریتم یادگیری ماشین، ضروری است که مجموعه‌داده از نظر پاکیزگی، سازگاری و مقیاس‌بندی مناسب بررسی و آماده‌سازی شود. در این پروژه، فرایند پیش‌پردازش شامل سه مرحلهٔ اصلی است:

۱. رسیدگی به مقادیر گمشده،
۲. نرمال‌سازی ویژگی‌های عددی،
۳. تقسیم داده‌ها به مجموعه‌های آموزش و آزمون.

تمامی مراحل پیش‌پردازش با استفاده از کتابخانه‌های pandas و numpy در محیط Jupyter Notebook پیاده‌سازی شده‌اند. به منظور سهولت در نگه‌داری و استفادهٔ مجدد از کد، هر بخش در قالب توابع جداگانه و ماژولار تعریف شده است.

۱.۳ رسیدگی به مقادیر گمشده

پس از بارگذاری مجموعه‌دادهٔ تجمیع‌شدهٔ بیماری قلبی از فایل CSV، ابتدا با استفاده از فرمان‌های df.info() و df.isna().sum() وضعیت مقادیر گمشده و ناسازگاری‌های احتمالی بررسی شد. برای تضمین کامل بودن داده‌ها و جلوگیری از بروز خطا در مراحل بعدی، یک راهبرد جایگزینی ساده اما مؤثر اعمال گردید:

- برای ویژگی‌های **عددی** مانند age، trestbps، chol، thalach و oldpeak مقادیر گمشده با □□□□□ی ستون جایگزین شدند.
- برای ویژگی‌های **دسته‌ای یا دودویی** مانند sex، cp، fbs، restecg، exang، slope، ca و thal مقادیر گمشده با □□ (پرمقدارترین کلاس) جایگزین شدند.

استفاده از میانه برای داده‌های عددی باعث کاهش حساسیت نسبت به مقادیر پرت می‌شود و برای داده‌های پزشکی که ممکن است شامل مقادیر غیرعادی باشند گزینه‌ای مناسب است. همچنین، به‌کارگیری مد برای ویژگی‌های دسته‌ای باعث حفظ فراوان‌ترین مقدار و جلوگیری از ایجاد کلاس‌های غیرواقعی می‌شود.

این عملیات در قالب تابع impute_missing_values() پیاده‌سازی شده است. پس از اجرای آن، بررسی مجدد داده‌ها نشان داد که هیچ مقدار گمشده‌ای باقی نمانده است.

۲.۳ نرمال‌سازی ویژگی‌های عددی

از آن‌جا که الگوریتم KNN مبتنی بر فاصله است، بسیار مهم است که ویژگی‌های عددی در یک مقیاس مشابه قرار داشته باشند. در غیر این صورت، ویژگی‌هایی که دامنهٔ بزرگ‌تری دارند ممکن است بر محاسبات فاصله غالب شوند و منجر به نتایج نادرست گردند.

در این پروژه از روش **مقیاس‌گذاری مین-مکس** برای نرمال‌سازی ویژگی‌های عددی استفاده شده است. برای یک ویژگی عددی x مقدار نرمال‌شدهٔ آن به صورت زیر محاسبه می‌شود:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \tag{۱}$$

که $x_{\min}$ و $x_{\max}$ به ترتیب کوچک‌ترین و بزرگ‌ترین مقدار آن ویژگی در **مجموعهٔ آموزش** هستند.

برای جلوگیری از نشت اطلاعات (data leakage)، مقادیر حداقلی و حداکثری تنها از داده‌های آموزشی به دست آمده و همان مقادیر برای تبدیل داده‌های آزمون استفاده شده‌اند. این فرایند در تابع min_max_scale_train_test() پیاده‌سازی شده است که خروجی آن نسخهٔ نرمال‌شدهٔ داده‌های آموزش و آزمون است.

در صورتی که ستونی عددی مقدار ثابتی داشته باشد (یعنی $x_{\max} = x_{\min}$)، به منظور جلوگیری از تقسیم بر صفر، مقدار نرمال‌شدهٔ آن ستون برابر صفر قرار داده شده است.

۳.۳ تقسیم داده‌ها به آموزش و آزمون

برای ارزیابی توان تعمیم مدل، مجموعه‌داده به دو بخش آموزش و آزمون تقسیم شد. ابتدا داده‌ها به‌طور تصادفی به‌هم‌ریخته شدند و سپس با نسبت **۸۰% برای آموزش** و **۲۰% برای آزمون** تفکیک شدند:

- ۸۰ درصد داده‌ها برای آموزش مدل استفاده شدند،
- ۲۰ درصد داده‌ها برای آزمون نهایی کنار گذاشته شدند.

این تقسیم‌بندی توسط تابع train_test_split_df() انجام شد که ورودی آن مجموعه‌دادهٔ کامل و درصد داده‌های آزمون است. برای تضمین تکرارپذیری (reproducibility)، یک مقدار بذر ثابت (random seed) تنظیم شد.

پس از انجام نرمال‌سازی و تقسیم داده‌ها، ماتریس‌های X_train، y_train، X_test و y_test حاصل شدند که در مراحل بعدی برای آموزش و ارزیابی طبقه‌بند KNN مورد استفاده قرار گرفتند.

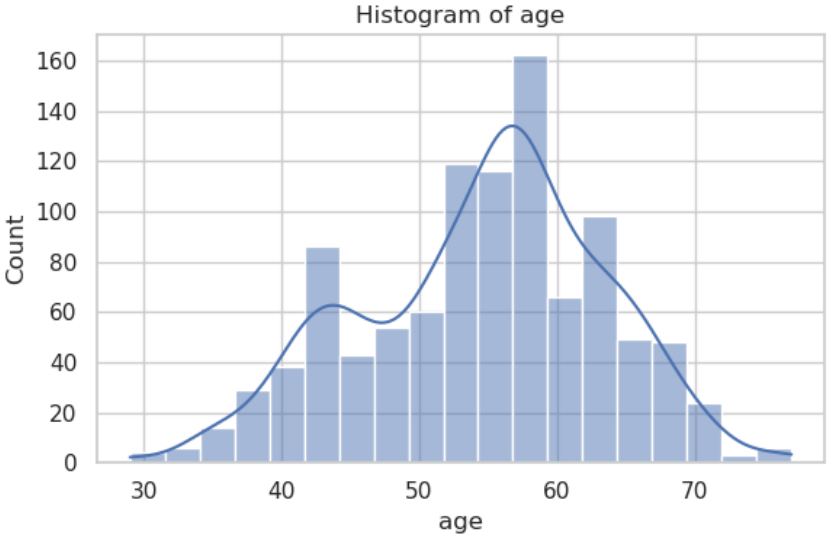
۴ تحلیل اکتشافی داده‌ها

پیش از آموزش طبقه‌بند، یک تحلیل اکتشافی داده‌ها (Exploratory Data Analysis, EDA) انجام شد تا درک بهتری از توزیع ویژگی‌ها، وجود مقادیر پرت و روابط بین متغیرها به دست آید. تمامی نمودارها و تصویرها با استفاده از کتابخانه‌های matplotlib و seaborn در پایتون تولید شده‌اند.

۱.۴ هیستوگرام ویژگی‌های عددی

برای بررسی اولیهٔ ویژگی‌های عددی، هیستوگرام‌هایی برای پنج ویژگی پیوستهٔ اصلی یعنی age، chol، oldpeak، thalach و trestbps رسم شد. این نمودارها شکل کلی توزیع، میزان کجی، و فراوانی مقادیر حدی را نشان می‌دهند. در ادامه، الگوهای مشاهده‌شده در هر ویژگی به‌طور خلاصه توصیف می‌شوند.

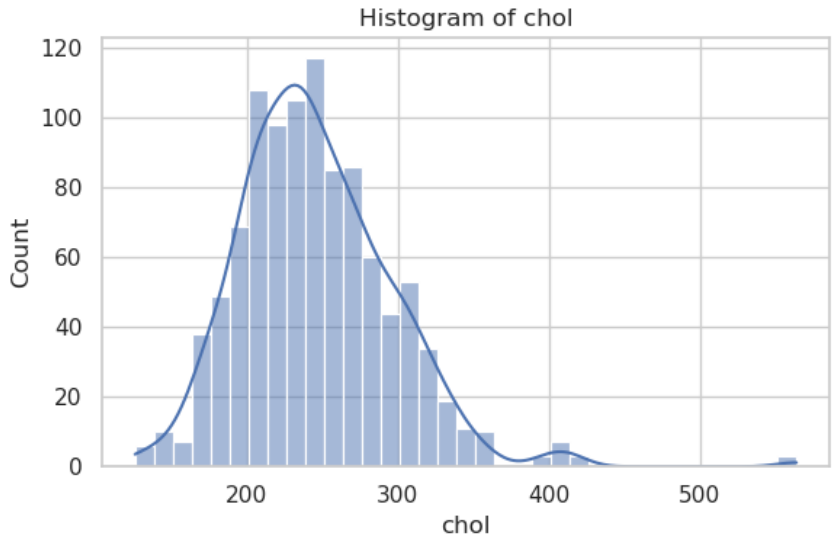
سن (age). نمودار شکل ۱ نشان می‌دهد که سن بیماران تقریباً در بازهٔ ۳۰ تا ۸۰ سال قرار دارد. توزیع سن شکلی نزدیک به نرمال با قله‌ای مشخص در حدود ۵۵ تا ۶۰ سال دارد. بخش عمدهٔ نمونه‌ها را بزرگسالان میانسال و سالمند تشکیل می‌دهند و تعداد افراد زیر ۴۰ سال بسیار کمتر است. این الگو از نظر بالینی طبیعی است، زیرا ریسک بیماری قلبی با افزایش سن به‌طور قابل توجهی بیشتر می‌شود. پس از سنین حدود ۶۵ تا ۷۰ سال، کاهش ملایمی در فراوانی نمونه‌ها دیده می‌شود که منعکس‌کنندهٔ تعداد کمتر بیماران بسیار سالمند در این مجموعه‌داده است.



شکل ۱: هیستوگرام ویژگی age.

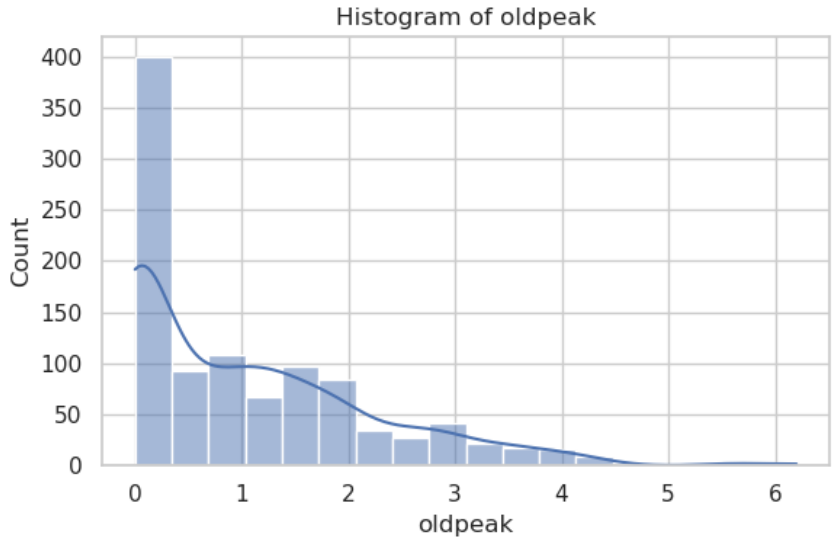
کلسترول (chol). هیستوگرام chol (شکل ۲) توزیعی با کجی ملایم به سمت راست را نشان می‌دهد. بیشتر مقادیر در بازهٔ ۲۰۰ تا ۲۶۰ میلی‌گرم بر دسی‌لیتر قرار دارند؛ این محدوده در عمل بالینی با سطوح مرزی یا بالا برای کلسترول سازگار است. دمی بلند در سمت راست

تا حدود ۴۰۰ تا ۵۵۰ میلی‌گرم بر دسی‌لیتر ادامه می‌یابد که نشان‌دهندهٔ تعداد اندکی از بیماران با اختلالات شدید چربی خون است. این مقادیر پرت از نظر پزشکی قابل قبول و با عوامل خطر شناخته‌شدهٔ بیماری قلبی هم‌خوان هستند.



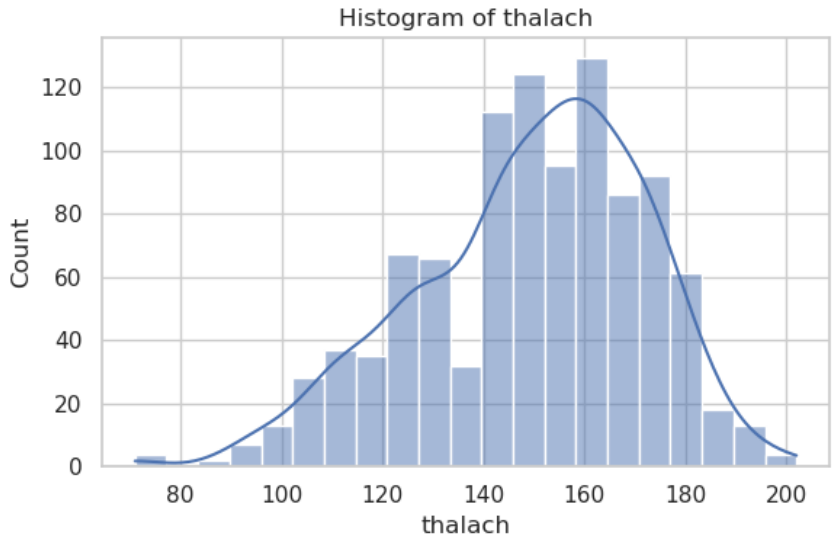
شکل ۲: هیستوگرام ویژگی chol.

oldpeak. ویژگی oldpeak که تغییر قطعهٔ ST ناشی از ورزش را نشان می‌دهد، در میان همهٔ ویژگی‌های عددی، بیشترین کجی به سمت راست را دارد (شکل ۳). بخش بزرگی از بیماران دارای مقداری نزدیک به صفر هستند که نشان‌دهندهٔ تغییر ناچیز در قطعهٔ ST است. با افزایش مقدار oldpeak، فراوانی به‌سرعت کاهش می‌یابد و تنها تعداد اندکی نمونه با مقادیر بالاتر از ۳ و مقادیر بسیار کمی با مقادیر حدود ۵ یا ۶ دیده می‌شود. این الگو نیز با انتظارات بالینی هم‌خوانی دارد، زیرا افت شدید ST در جمعیت کلی بیماران قلبی پدیده‌ای نسبتاً نادر است.



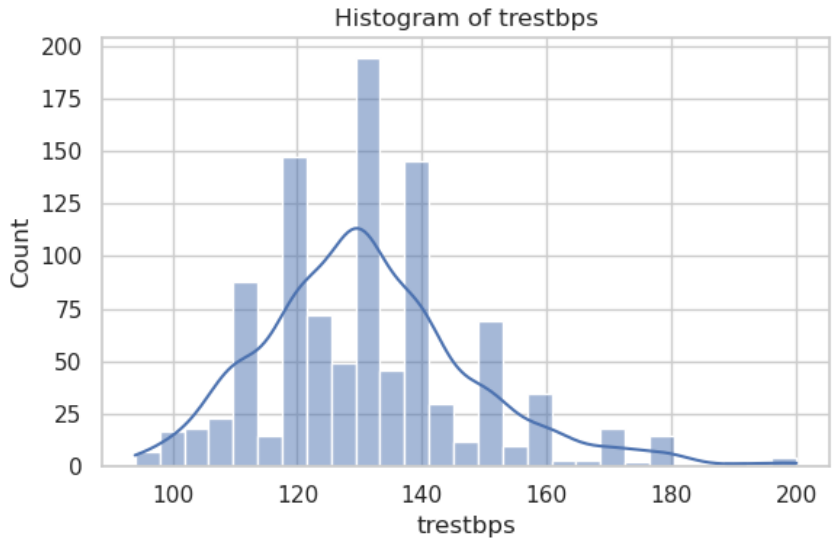
شکل ۳: هیستوگرام ویژگی oldpeak.

thalach. حداکثر ضربان قلب (thalach) توزیعی تقریباً متقارن و زنگوله‌ای‌شکل دارد (شکل ۴). بیشترین فراوانی در بازهٔ ۱۵۰ تا ۱۷۰ ضربان در دقیقه در طول تست ورزشی مشاهده می‌شود. مقادیر بسیار پایین (کمتر از ۱۰۰) بسیار نادر هستند و می‌توانند نشان‌دهندهٔ مشکلات شدید قلبی یا ناتمام‌ماندن تست ورزش باشند. این رفتار با مشاهدات بالینی روی جمعیت‌های مورد ارزیابی در تست‌های ورزش هم‌خوانی خوبی دارد.



شکل ۴: هیستوگرام ویژگی thalach.

فشار خون استراحت (trestbps). همان‌طور که در شکل ۵ دیده می‌شود، توزیع trestbps اندکی به سمت مقادیر بالاتر متمایل است. بیشتر مشاهدات در بازهٔ ۱۲۰ تا ۱۴۰ میلی‌متر جیوه قرار دارند که به فشار خون خفیف تا متوسط (پرفشاری خون) مربوط است و یک عامل خطر شایع برای بیماری قلبی محسوب می‌شود. مقادیر بالاتر، تا حدود ۱۸۰ تا ۲۰۰ میلی‌متر جیوه، با فراوانی کم‌تر مشاهده می‌شوند، اما از نظر بالینی برای بیماران پرخطر یا درمان‌نشده کاملاً واقع‌بینانه‌اند.



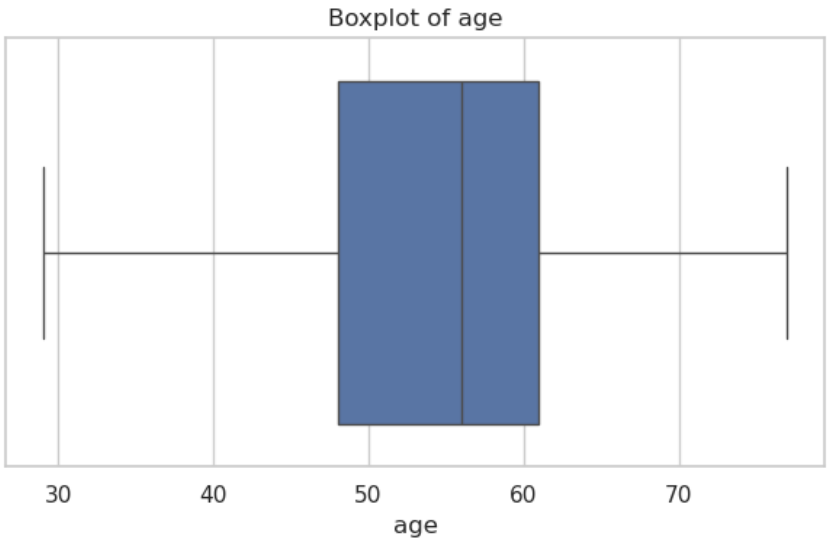
شکل ۵: هیستوگرام ویژگی trestbps.

در مجموع، هیستوگرام‌ها نشان می‌دهند که ویژگی‌های عددی ترکیبی از توزیع‌های تقریباً نرمال، با کجی ملایم و همچنین توزیع‌های شدیداً کج را تشکیل می‌دهند که همگی با رفتار واقعی داده‌های پزشکی سازگارند. این نتایج ضرورت نرمال‌سازی ویژگی‌ها را پیش از به‌کارگیری الگوریتم‌های مبتنی بر فاصله‌ای مانند KNN تقویت می‌کند.

۲.۴ نمودارهای جعبه‌ای و مقادیر پرت

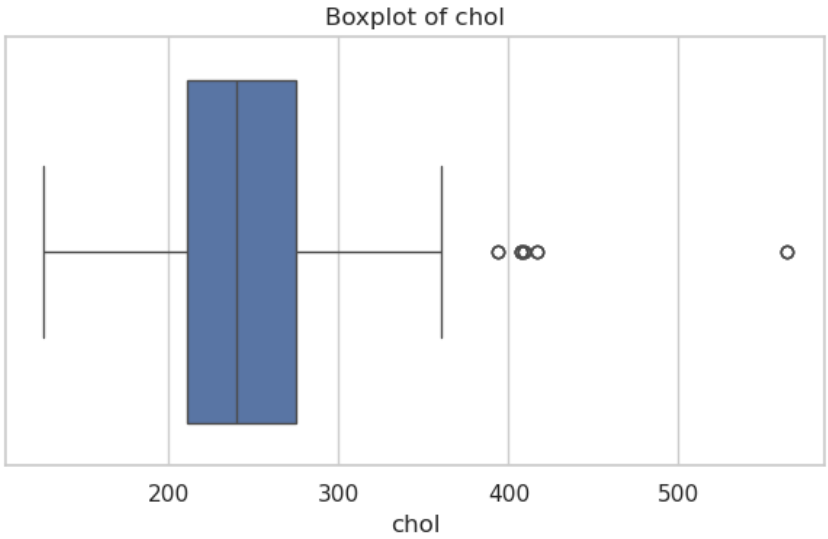
برای بررسی پراکندگی، گرایش مرکزی و مقادیر پرت، نمودارهای جعبه‌ای برای پنج ویژگی عددی اصلی رسم شد. این نمودارها دامنهٔ تغییرات هر ویژگی، موقعیت میانه و وجود مقادیر غیرعادی را به‌خوبی نمایش می‌دهند و درک بهتری از رفتار اندازه‌گیری‌های بالینی فراهم می‌کنند.

سن (age). نمودار جعبه‌ای age (شکل ۶) نشان می‌دهد که اغلب بیماران در بازهٔ تقریبی ۵۰ تا ۶۲ سال قرار دارند و میانه در حدود ۵۶ تا ۵۸ سال است. مقادیر پرت قابل‌توجهی مشاهده نمی‌شود و توزیع سنی نسبتاً یکنواخت و متمرکزی دارد. این الگو با انتظار بالینی مبنی بر شیوع بالاتر بیماری قلبی در بزرگسالان میانسال و سالمند سازگار است.



شکل ۶: نمودار جعبه‌ای ویژگی age.

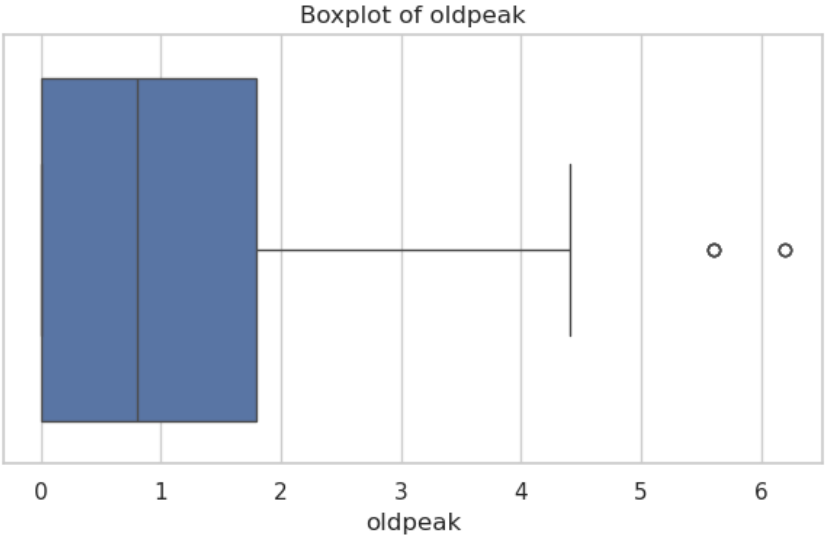
کلسترول (chol). نمودار جعبه‌ای chol (شکل ۷) وجود تعداد قابل‌توجهی مقادیر پرت در سمت بالاتر را نشان می‌دهد؛ مقادیری بالاتر از ۴۰۰ و حتی نزدیک ۵۵۰ میلی‌گرم بر دسی‌لیتر. دامنهٔ بین‌چارکی تقریباً بین ۲۰۰ تا ۲۷۵ میلی‌گرم بر دسی‌لیتر قرار دارد و میانه در حدود ۲۴۰ تا ۲۵۰ است. این مقادیر پرت از نظر بالینی منطقی بوده و می‌توانند بیانگر موارد شدید هیپرکلسترولمی باشند که از عوامل خطر مهم برای بیماری‌های قلبی-عروقی هستند.



شکل ۷: نمودار جعبه‌ای ویژگی chol.

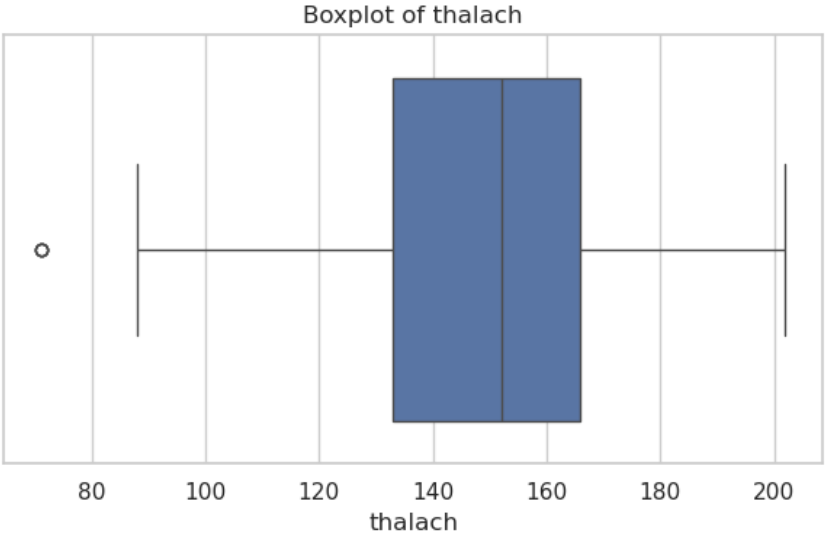
oldpeak. نمودار جعبه‌ای oldpeak (شکل ۸) توزیعی با کجی شدید به سمت راست را تأیید می‌کند؛ به‌طوری که بیشترین بخش مقادیر بین ۰ تا حدود ۵.۱ قرار دارد. چند مقدار پرت واضح با مقادیر بالاتر از ۵ نیز دیده می‌شود که بیانگر موارد نادر ولی از نظر بالینی بسیار قابل‌توجه افت شدید ST در حین تست ورزش هستند. این مقادیر به ساختار دنبالهٔ سنگین در هیستوگرام

این ویژگی کمک می‌کنند.



شکل ۸: نمودار جعبه‌ای ویژگی oldpeak.

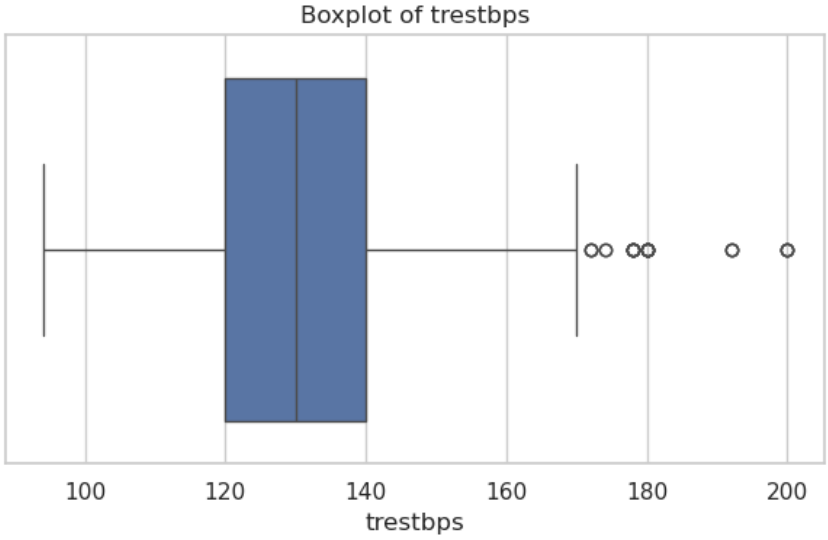
thalach. در نمودار جعبه‌ای thalach (شکل ۹) بیشتر داده‌ها بین حدود ۱۴۰ تا ۱۷۰ ضربان در دقیقه قرار دارند و میانه در حدود ۱۵۵ است. یک مقدار پرت پایین در حدود ۸۰ ضربان در دقیقه مشاهده می‌شود که ممکن است به محدودیت شدید عملکرد قلب یا ناقص‌بودن تست ورزش مربوط باشد. مقادیر بالاتر نیز تا حدود ۲۰۰ ضربان در دقیقه امتداد دارند که در زمینهٔ تست ورزش قابل‌قبول و از نظر فیزیولوژیک محتمل هستند.



شکل ۹: نمودار جعبه‌ای ویژگی thalach.

فشار خون استراحت (trestbps). نمودار جعبه‌ای trestbps (شکل ۱۰) توزیعی معمول در محیط‌های بالینی را نشان می‌دهد؛ به‌طوری که بیشتر مقادیر در بازهٔ ۱۲۰ تا ۱۴۰ میلی‌متر جیوه و میانه در حدود ۱۳۰ تا ۱۳۵ قرار دارد. چند مقدار پرت چشمگیر در محدودهٔ ۱۷۰ تا ۲۰۰

میلی‌متر جیوه وجود دارد که می‌تواند نمایانگر موارد پرفشاری خون شدید باشد. این الگو با الگوهای شناخته‌شدهٔ ریسک در بیماران قلبی سازگار است.



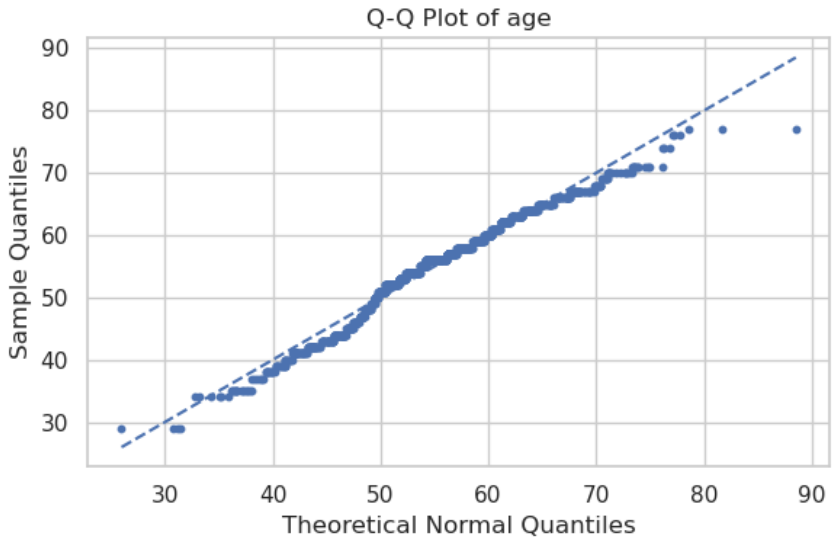
شکل ۱۰: نمودار جعبه‌ای ویژگی trestbps.

به‌طور کلی، نمودارهای جعبه‌ای حضور مقادیر پرت معنادار در ویژگی‌هایی مانند chol، oldpeak و trestbps را تأیید می‌کنند، در حالی که ویژگی‌هایی مانند age و thalach توزیع پایدارتر و متمرکزتری دارند. با توجه به این‌که الگوریتم KNN نسبت به تعداد کمی از مقادیر پرت جداافتاده (به‌ویژه پس از نرمال‌سازی) نسبتاً مقاوم است، این مقادیر در این پروژه حذف نشدند.

۳.۴ نمودارهای Q–Q

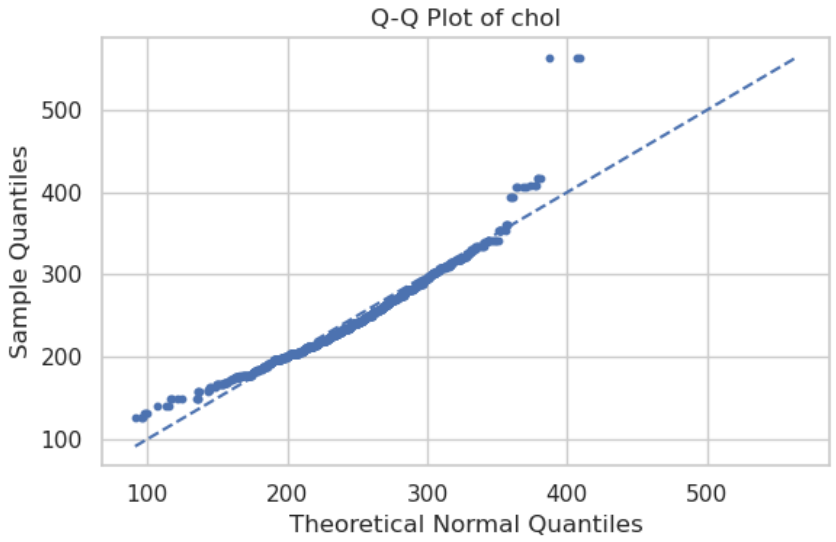
برای بررسی میزان نزدیکی ویژگی‌های عددی به توزیع نرمال، نمودارهای Q–Q برای سه ویژگی نماینده یعنی age، chol و thalach رسم شد. در هر نمودار، مقادیر مرتب‌شدهٔ نمونه با کوانتیل‌های یک توزیع نرمال با میانگین و واریانس مشابه مقایسه می‌شوند. انحراف از خط قطری، نشان‌دهندهٔ فاصله از نرمال بودن است.

سن (age). نمودار Q–Q مربوط به age (شکل ۱۱) نشان می‌دهد که بخش عمدهٔ نقاط در نزدیکی خط قطری قرار دارند؛ این موضوع نشان‌دهندهٔ نزدیک بودن بخش مرکزی توزیع به نرمال است. با این حال، در دو دم توزیع انحراف‌های قابل‌توجهی دیده می‌شود: سنین پایین‌تر اندکی زیر خط مرجع و سنین بالاتر (حدوداً بالای ۷۵ سال) کمی بالای آن قرار می‌گیرند. این الگو بیانگر دنباله‌های نسبتاً سنگین اما ساختار کلی نسبتاً متقارن است؛ رفتاری که برای متغیر سن در داده‌های بالینی تا حد زیادی معمول است.



شکل ۱۱: نمودار Q–Q برای ویژگی age.

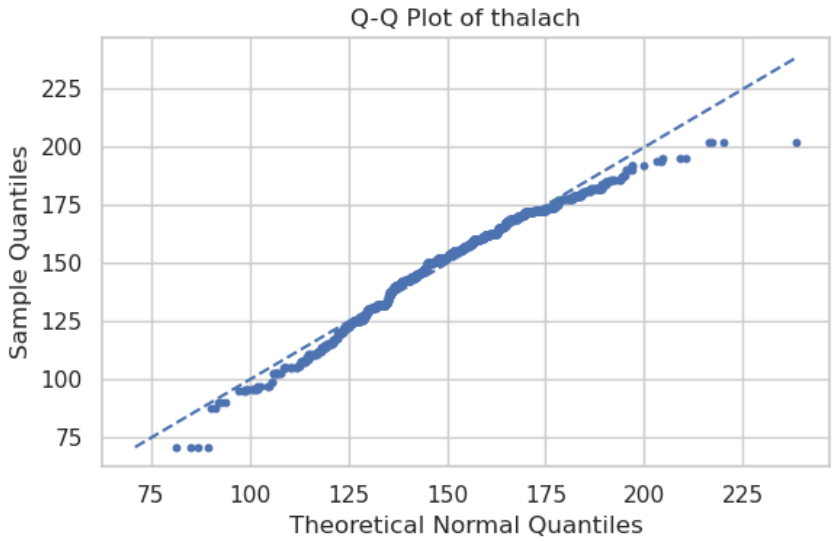
کلسترول (chol). نمودار Q–Q برای chol (شکل ۱۲) انحراف‌های بسیار شدیدتری را نشان می‌دهد، به‌ویژه در دم بالایی توزیع. در حالی که بخش مرکزی داده‌ها تا حدی با خط قطری هم‌راستا است، چندین مقدار بسیار بالا (بالاتر از ۴۰۰ میلی‌گرم بر دسی‌لیتر) به‌وضوح بالاتر از کوانتیل‌های نرمال متناظر قرار می‌گیرند. این نقاط متناظر با موارد شدید هیپرکلسترولمی هستند و رفتار کج‌به‌راست توزیع chol را تأیید می‌کنند. در نتیجه، این ویژگی با توزیع نرمال فاصلهٔ قابل‌توجهی دارد و حاوی مقادیر پرت بالینی مهم است.



شکل ۱۲: نمودار Q–Q برای ویژگی chol.

thalach. نمودار Q–Q ویژگی thalach (شکل ۱۳) نشان می‌دهد که بخش مرکزی توزیع تقریباً روی خط قطری قرار دارد و رفتاری نزدیک به نرمال را نمایش می‌دهد. با این حال، در دم پایین، چند مقدار بسیار کم (در حدود ۷۰ تا ۹۰ ضربان در دقیقه) به‌وضوح زیر خط نرمال

قرار گرفته‌اند و در دم بالا نیز برخی مقادیر بالاتر از ۲۰۰ ضربان در دقیقه کمی بالاتر از خط واقع شده‌اند. این انحراف‌ها نشان‌دهندهٔ کجی ملایم ناشی از مشاهدات بسیار کم یا بسیار بالا هستند که با تغییرات طبیعی پاسخ بیماران در تست ورزش سازگار است.

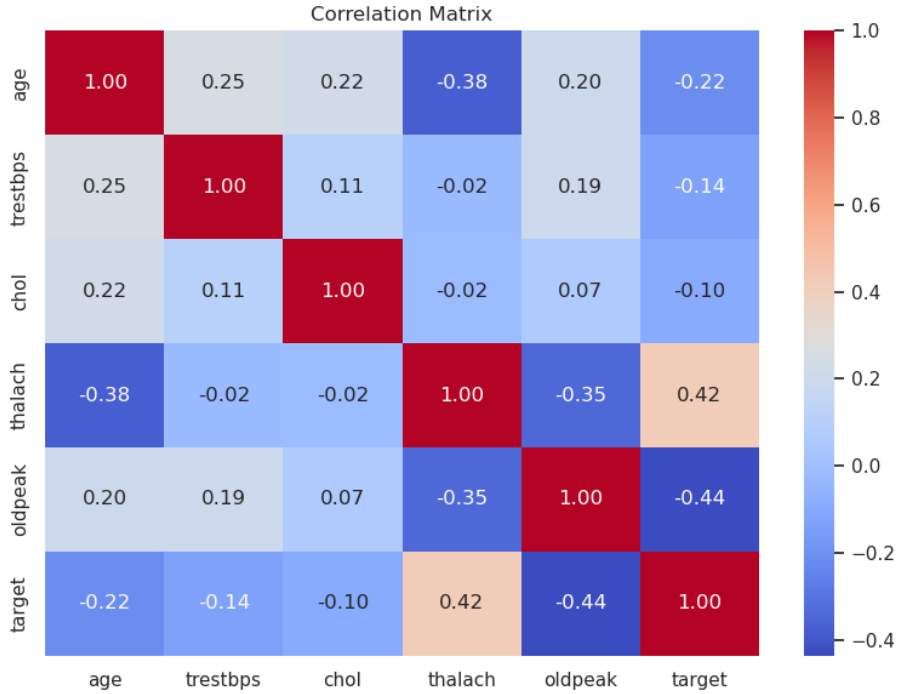


شکل ۱۳: نمودار Q–Q برای ویژگی thalach.

به‌طور کلی، نمودارهای Q–Q نشان می‌دهند که هیچ‌یک از ویژگی‌های بررسی‌شده کاملاً از توزیع نرمال پیروی نمی‌کنند. هر ویژگی در دم‌ها یا کلاً رفتاری با کجی و انحراف از نرمال دارد و ویژگی chol بیشترین انحراف را به دلیل وجود مقادیر بسیار بالای کلسترول نشان می‌دهد. این مسئله برای طبقه‌بند KNN مشکل‌زا نیست، زیرا این الگوریتم فرضی دربارهٔ نرمال بودن توزیع ویژگی‌ها ندارد.

۴.۴ ماتریس همبستگی

برای بررسی روابط خطی میان ویژگی‌های عددی و نیز ارتباط آن‌ها با متغیر هدف، ماتریس همبستگی پیرسون محاسبه و به صورت یک نقشهٔ حرارتی نمایش داده شد (شکل ۱۴). این ماتریس الگوهای بالینی معناداری را نشان می‌دهد و مشخص می‌کند کدام ویژگی‌ها برای تمایز بین بیماران دارای و فاقد بیماری قلبی مهم‌تر هستند.



شکل ۱۴: ماتریس همبستگی برای تعدادی از ویژگی‌های عددی و متغیر هدف.

قوی‌ترین همبستگی‌ها با متغیر هدف برای thalach و oldpeak مشاهده می‌شوند. ویژگی thalach (حداکثر ضربان قلب) همبستگی مثبت متوسطی با هدف دارد ($r = 0.42$) که نشان می‌دهد افرادی که در طول تست ورزش به ضربان قلب بالاتری می‌رسند، کمتر در معرض ابتلا به بیماری قلبی قرار دارند. در مقابل، oldpeak همبستگی منفی نسبتاً قوی با متغیر هدف دارد ($r = -0.44$) که با دانش بالینی موجود هم‌خوان است؛ یعنی افت بیشتر قطعهٔ ST معمولاً با ایسکمی میوکارد و احتمال بیشتر بیماری قلبی همراه است. سن نیز همبستگی منفی ضعیف‌تری با هدف دارد ($r = -0.22$) که بیانگر افزایش شیوع بیماری قلبی با افزایش سن است، هرچند این رابطه کاملاً خطی نیست.

سایر ویژگی‌ها مانند trestbps و chol تنها همبستگی‌های ضعیفی با متغیر هدف نشان می‌دهند؛ این موضوع حاکی از آن است که در چارچوب این مجموعه‌داده، این متغیرها به‌تنهایی قدرت پیش‌بینی محدودی دارند، اما ممکن است در ترکیب با سایر ویژگی‌ها در یک فضای چندمتغیره نقش مؤثری ایفا کنند.

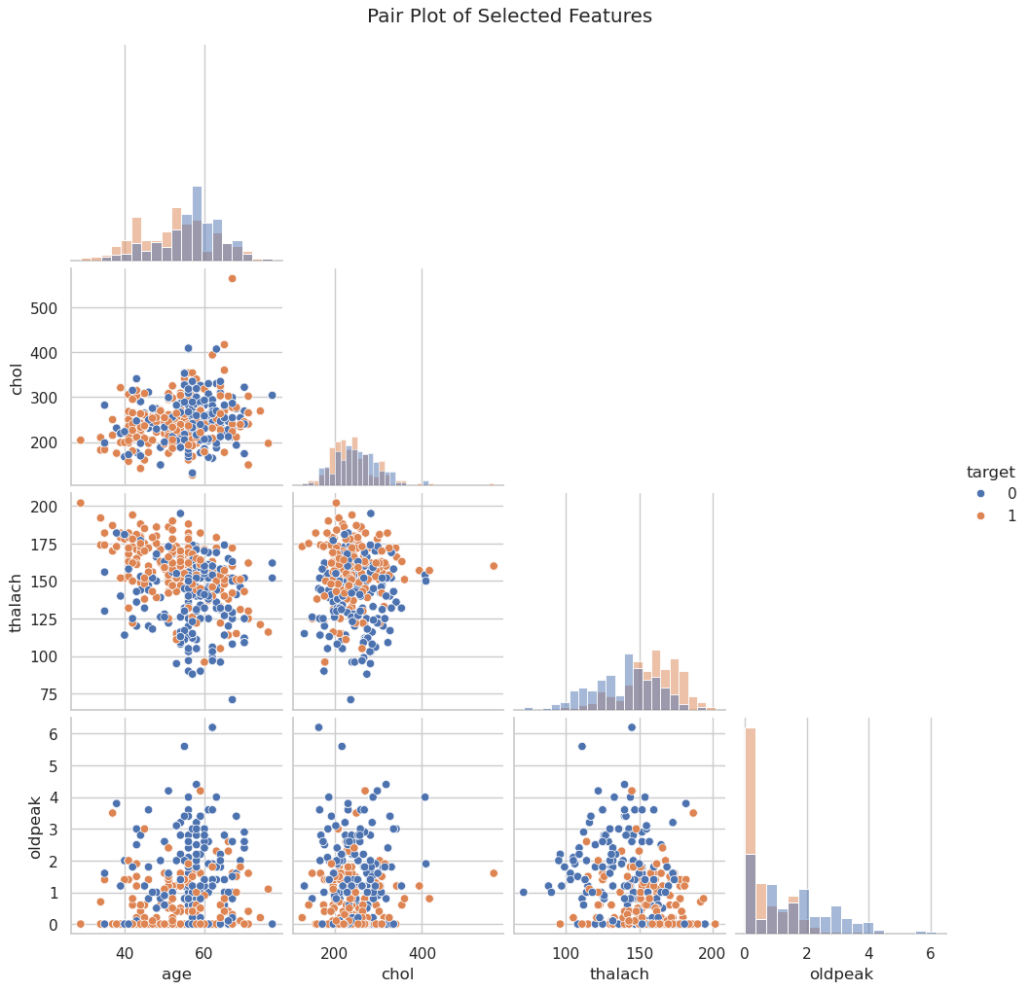
برخی روابط میان ویژگی‌ها نیز قابل توجه‌اند. سن و thalach همبستگی منفی متوسطی دارند ($r = -0.38$) که نشان می‌دهد افراد مسن‌تر معمولاً در تست ورزش به حداکثر ضربان قلب پایین‌تری می‌رسند. همچنین، oldpeak و thalach نیز همبستگی منفی ($r = -0.35$) نشان می‌دهند که کاهش تحمل ورزش و عملکرد قلبی ضعیف‌تر در بیماران با پاسخ ایسکمیک را منعکس می‌کند.

در مجموع، ماتریس همبستگی نشان می‌دهد که در این مجموعه‌داده، بیماری قلبی بیش از همه با ویژگی‌های مرتبط با عملکرد در تست ورزش مانند thalach و oldpeak مرتبط است، در حالی که اندازه‌گیری‌های استراحتی مانند trestbps و chol وابستگی خطی ضعیف‌تری با متغیر هدف دارند. این یافته‌ها همچنین توضیح می‌دهند که چرا مدل‌های مبتنی بر فاصله مانند KNN، که تعامل هم‌زمان چند ویژگی را در نظر می‌گیرند، می‌توانند علیرغم ضعف همبستگی

خطی تک‌ویژگی، عملکرد مناسبی داشته باشند.

۵.۴ نمودارهای جفتی (Pair Plots)

برای بررسی روابط چندمتغیره بین ویژگی‌های عددی کلیدی، یک pair plot برای ویژگی‌های age، chol، thalach و oldpeak رسم شد که در آن نقاط بر اساس متغیر هدف دودویی رنگ‌آمیزی شده‌اند. این نمودار نمایی هم‌زمان از توزیع حاشیه‌ای هر ویژگی و روابط دو به دوی آن‌ها ارائه می‌دهد (شکل ۱۵).



شکل ۱۵: نمودار جفتی برای چند ویژگی عددی منتخب، رنگ‌آمیزی‌شده بر اساس کلاس هدف.

هیستوگرام‌های روی قطر نشان می‌دهند که age و thalach توزیع‌هایی نسبتاً متقارن دارند، در حالی که chol و oldpeak به‌طور قابل توجهی به سمت راست کج شده‌اند. از نظر توزیع کلاس‌ها، بیماران دارای بیماری قلبی (هدف = ۱) اندکی بیشتر در محدودهٔ سن‌های بالاتر دیده می‌شوند، هرچند این اثر نسبتاً ضعیف است. در مقابل، تفاوت‌های واضح‌تری در توزیع thalach و oldpeak مشاهده می‌شود: افرادی که بیماری قلبی ندارند معمولاً به حداکثر ضربان قلب بالاتر می‌رسند، در حالی که بیماران قلبی بیش‌تر oldpeakهای بزرگ‌تر و thalachهای کم‌تر دارند.

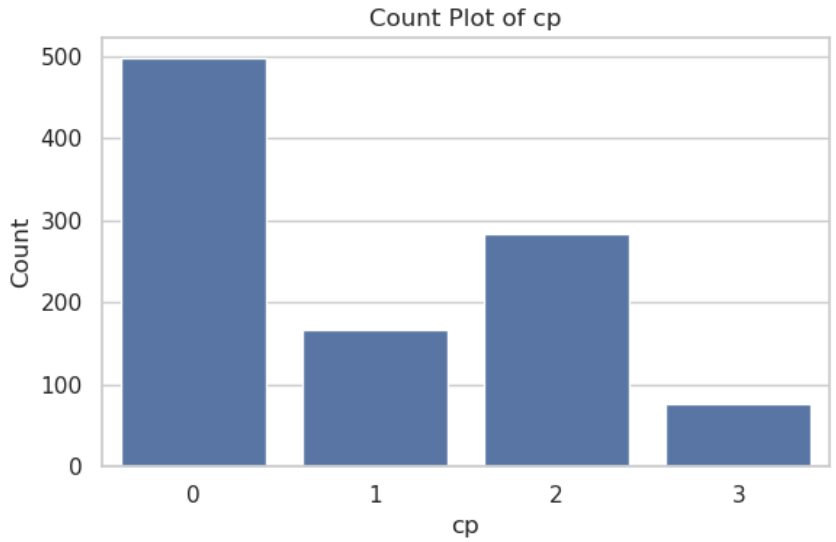
نمودارهای پراکندگی دو به دو الگوهای مهمی را آشکار می‌کنند. رابطهٔ بین age و thalach شیب منفی مشخصی دارد که کاهش ظرفیت ورزش با افزایش سن را نشان می‌دهد. همچنین، ترکیب thalach و oldpeak بهترین جداسازی بصری بین دو کلاس را فراهم می‌کند: افراد بدون بیماری قلبی عمدتاً در ناحیهٔ thalach بالا و oldpeak نزدیک صفر خوشه‌بندی می‌شوند، در حالی که بیماران قلبی در مقادیر پایین‌تر thalach و مقادیر بالاتر oldpeak متمرکز هستند. در مقابل، ویژگی chol در تمام نمایش‌های دو به دو هم‌پوشانی قابل‌توجهی بین دو کلاس نشان می‌دهد و حاکی از آن است که این ویژگی به تنهایی توان تفکیک‌کنندگی ضعیف‌تری دارد.

در مجموع، نمودارهای جفتی تأیید می‌کنند که بیماری قلبی در این مجموعه‌داده بیش از همه با شاخص‌های عملکردی مرتبط با تست ورزش (thalach و oldpeak) پیوند دارد، در حالی که ویژگی‌هایی مانند chol و age جداسازی ضعیف‌تر یا کم‌ثبات‌تری ارائه می‌دهند. این الگوهای چندمتغیره استفاده از طبقه‌بند مبتنی بر فاصله‌ای مانند KNN را توجیه می‌کنند، زیرا چنین مدلی می‌تواند از این روابط حتی در صورت غیرخطی بودن یا نبود جداسازی واضح در امتداد یک ویژگی منفرد بهره‌برداری کند.

۶.۴ نمودارهای شمارشی ویژگی‌های دسته‌ای

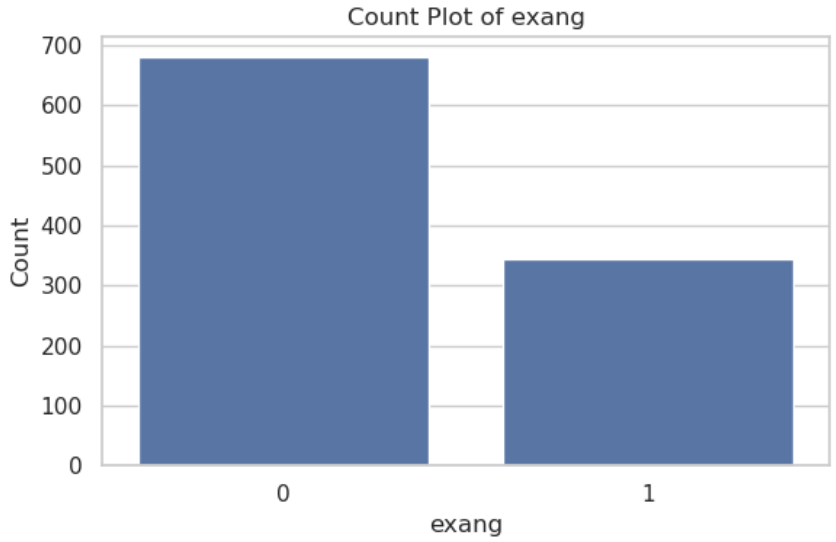
برای درک بهتر توزیع ویژگی‌های دسته‌ای در مجموعه‌داده، نمودارهای شمارشی برای cp (نوع درد قفسهٔ سینه)، exang (آنژین ناشی از ورزش)، sex و متغیر هدف target رسم شد. این نمودارها فراوانی هر دسته را نشان می‌دهند و به شناسایی عدم توازن‌های احتمالی در داده کمک می‌کنند.

نوع درد قفسهٔ سینه (cp). شکل ۱۶ نشان می‌دهد که توزیع نوع درد قفسهٔ سینه یکنواخت نیست. دستهٔ cp = 0 (آنژین تیپیک) با نزدیک به ۵۰۰ مشاهده، رایج‌ترین حالت است، در حالی که cp = 3 (بدون علامت) با کمتر از ۱۰۰ نمونه کم‌ترین فراوانی را دارد. دو دستهٔ دیگر از نظر فراوانی بین این دو قرار می‌گیرند. این الگو با مشاهدات بالینی واقعی سازگار است، زیرا آنژین تیپیک اغلب شکایت اصلی در بیماران قلبی محسوب می‌شود.



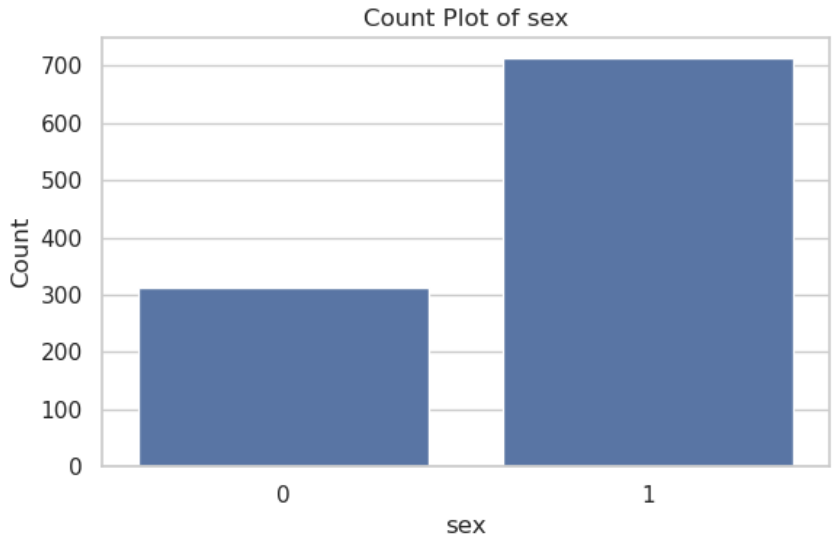
شکل ۱۶: نمودار شمارشی ویژگی cp (نوع درد قفسهٔ سینه).

آنژین ناشی از ورزش (exang). همان‌طور که در شکل ۱۷ دیده می‌شود، اکثریت بیماران (exang = 0) در طول تست ورزش دچار آنژین نمی‌شوند. زیرمجموعهٔ کوچک‌تری، تقریباً حدود یک‌سوم داده‌ها، شامل بیمارانی است که در طول تست دچار آنژین می‌شوند (exang = 1). این عدم توازن تا حدی طبیعی است، زیرا همهٔ افراد با وجود بیماری زمینه‌ای الزاماً در تست ورزش دچار درد قفسهٔ سینه نمی‌شوند.



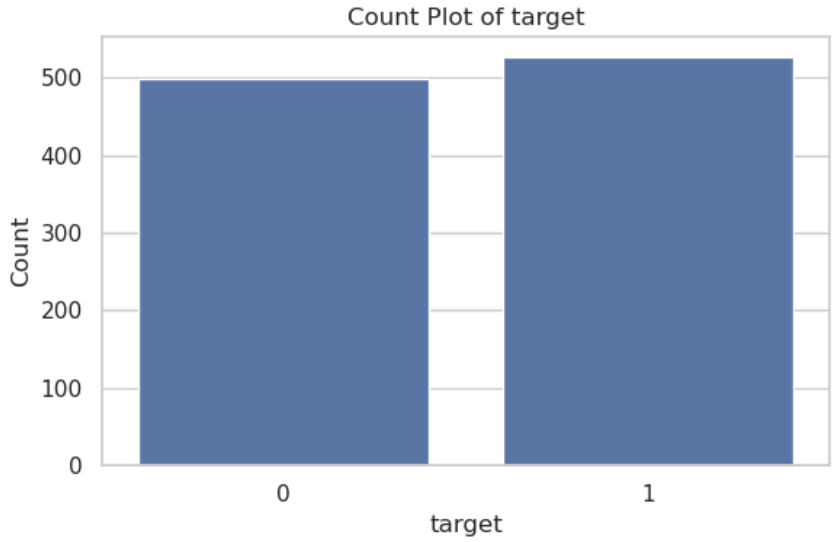
شکل ۱۷: نمودار شمارشی ویژگی exang.

جنسیت (sex). شکل ۱۸ نشان می‌دهد که در این مجموعه‌داده تعداد بیماران مرد (sex = 1) به‌طور قابل‌توجهی بیشتر از بیماران زن (sex = 0) است و نسبت تقریبی ۲ به ۱ را نشان می‌دهد. این الگو با شیوع بالاتر بیماری قلبی در مردان در بسیاری از جمعیت‌های بالینی هم‌خوانی دارد.



شکل ۱۸: نمودار شمارشی ویژگی sex.

متغیر هدف (target). همان‌طور که در شکل ۱۹ مشاهده می‌شود، توزیع کلاس در متغیر هدف تقریباً متعادل است؛ به این معنا که تعداد نمونه‌های کلاس target = 0 (بدون بیماری قلبی) و target = 1 (وجود بیماری قلبی) تقریباً برابر است. این تعادل برای آموزش مدل‌های طبقه‌بندی مفید است، زیرا ریسک سوگیری مدل به سمت یکی از کلاس‌ها را کاهش می‌دهد و نیاز به روش‌های متعادل‌سازی مانند oversampling یا وزن‌دهی به کلاس‌ها را برطرف می‌کند.



شکل ۱۹: نمودار شمارشی متغیر هدف target.

در مجموع، نمودارهای شمارشی نشان می‌دهند که توزیع ویژگی‌های دسته‌ای در این مجموعه‌داده با الگوهای کلینیکی واقعی سازگار است. اگرچه برخی دسته‌ها، مانند نوع درد قفسهٔ سینه و جنسیت، عدم توازن قابل‌توجهی دارند، اما توازن مناسب در متغیر هدف باعث می‌شود مدل‌هایی مانند KNN بدون نیاز به تنظیمات اضافی برای عدم توازن کلاس‌ها، به‌طور مؤثر آموزش داده شوند.

۵ طبقه‌بندی با KNN

در این بخش، الگوریتم K-Nearest Neighbors (KNN) برای طبقه‌بندی بیماران بر اساس اندازه‌گیری‌های بالینی آن‌ها به کار گرفته می‌شود. KNN یک روش یادگیری مبتنی بر نمونه و ناپارامتری است و برای این مجموعه‌داده مناسب است؛ زیرا در تحلیل اکتشافی داده‌ها تفکیک‌پذیری خطی قوی میان کلاس‌ها مشاهده نشد و چندین ویژگی، به‌ویژه thalach و oldpeak، روابط غیرخطی با وجود بیماری قلبی نشان می‌دهند. از آن‌جا که KNN مستقیماً بر فاصله بین بردارهای ویژگی تکیه دارد، پیش‌پردازش مناسب و مقیاس‌بندی درست ویژگی‌ها اجزای کلیدی زنجیرهٔ طبقه‌بندی محسوب می‌شوند.

۱.۵ مروری بر الگوریتم KNN

برای یک نمونهٔ آزمون x، طبقه‌بند KNN به صورت زیر عمل می‌کند:

۱. محاسبهٔ فاصله بین x و تمامی نمونه‌های مجموعهٔ آموزش،

۲. انتخاب k نمونهٔ آموزشیِ نزدیک‌تر به x،

۳. تعیین کلاس غالب در میان این k همسایه،

۴. انتساب همان برچسب کلاس به نمونهٔ x.

از آن‌جا که KNN هیچ فرض مشخصی دربارهٔ شکل تابعی یا توزیع احتمالاتی داده‌ها ندارد، برای مجموعه‌داده‌های پزشکی که در آن‌ها برهم‌کنش‌های پیچیده و غیرخطی میان متغیرها وجود دارد، بسیار مناسب است؛ موضوعی که در نمودارهای جفتی و الگوهای همبستگی مشاهده‌شده در بخش تحلیل اکتشافی نیز تأیید شد.

۲.۵ تابع فاصلهٔ ترکیبی برای ویژگی‌های عددی و دسته‌ای

مجموعه‌دادهٔ مورد استفاده شامل ترکیبی از ویژگی‌های عددی (مانند age، trestbps، chol، thalach و oldpeak) و ویژگی‌های دسته‌ای یا دودویی (مانند sex، cp، exang، restecg) است. استفاده از یک معیار فاصلهٔ صرفاً عددی (مثلاً فاصلهٔ اقلیدسی روی همهٔ ستون‌ها) به‌تنهایی مناسب نیست و می‌تواند به تفسیر نادرست ویژگی‌های دسته‌ای منجر شود.

برای رفع این مشکل، یک تابع فاصلهٔ ترکیبی به کار گرفته شد که در آن برای ویژگی‌های عددی از فاصلهٔ اقلیدسی و برای ویژگی‌های دسته‌ای از فاصلهٔ Hamming استفاده می‌شود:

$$d(x, y) = w_{\text{num}} \cdot d_{\text{num}}(x, y) + w_{\text{cat}} \cdot d_{\text{cat}}(x, y),$$

که در آن، هر دو وزن در این مطالعه برابر 0.5 در نظر گرفته شدند. این وزن‌دهی باعث می‌شود ویژگی‌های عددی—به‌خصوص ویژگی‌های تأثیرگذاری مانند thalach و oldpeak—در محاسبهٔ شباهت نقش پررنگی داشته باشند، بدون آن‌که اطلاعات موجود در ویژگی‌های دسته‌ای نادیده گرفته شود.

۳.۵ آموزش مدل و انتخاب مقدار k

پس از انجام پیش‌پردازش و نرمال‌سازی Min–Max، مجموعه‌داده با نسبت ۸۰/۲۰ به آموزش و آزمون تقسیم شد. طبقه‌بند KNN روی ماتریس X_{train} آموزش داده شد و روی X_{test} مورد ارزیابی قرار گرفت.

برای بررسی تأثیر اندازهٔ همسایگی، الگوریتم KNN برای مقادیر مختلف k ارزیابی شد:

$$k \in \{3, 5, 7\}.$$

برای هر مقدار k، عملکرد مدل با استفاده از معیارهای دقت (Accuracy)، دقت مثبت (Precision)، بازخوانی (Recall) و امتیاز F1 سنجیده شد؛ معیارهایی که در کاربردهای بالینی اهمیت ویژه‌ای دارند، زیرا هم خطاهای نوع اول (تشخیص اشتباه بیماری) و هم خطاهای نوع دوم (عدم تشخیص بیماری) می‌توانند پیامدهای جدی داشته باشند.

۴.۵ نتایج طبقه‌بندی

جدول ۱: عملکرد طبقه‌بند KNN برای مقادیر مختلف k.

F۱-score	Recall	Precision	Accuracy	k
۹۳۳۳.۰	۹۵۱۵.۰	۹۱۵۹.۰	۹۳۱۷.۰	۳
۸۲۷۹.۰	۸۶۴۱.۰	۷۹۴۶.۰	۸۱۹۵.۰	۵
۸۴۲۶.۰	۸۸۳۵.۰	۸۰۵۳.۰	۸۳۴۱.۰	۷

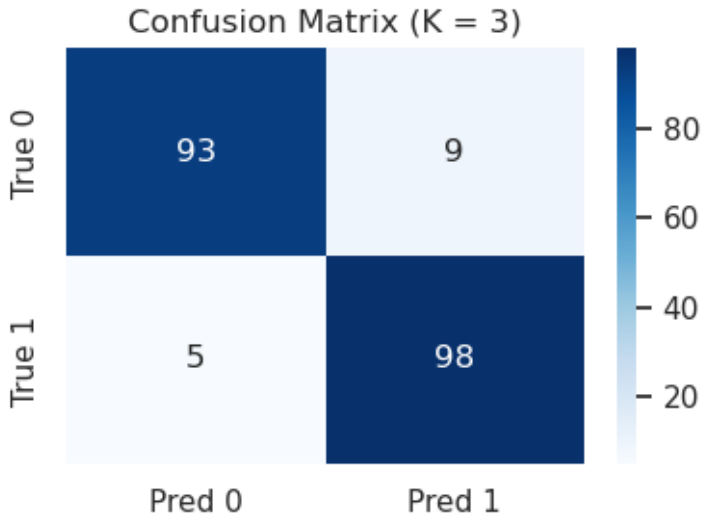
بیشترین مقدار دقت در $k = 3$ به دست آمد. این نتیجه با سطح متوسط هم‌پوشانی کلاس‌ها که در نمودارهای پراکندگی و نمودارهای جفتی مشاهده شد سازگار است؛ مقادیر کوچک‌تر k به طبقه‌بند اجازه می‌دهند الگوهای محلی را که عمدتاً توسط دو ویژگی بسیار اطلاع‌بخش thalach و oldpeak شکل می‌گیرند، بهتر شناسایی کند. با افزایش k، این تمایزهای محلی بیش از حد هموار شده و کارایی مدل کاهش می‌یابد.

در نتیجه، اندازهٔ همسایگی بهینه برای این مجموعه‌داده برابر است با:

$$k_{\text{best}} = 3.$$

۵.۵ ماتریس اغتشاش

ماتریس اغتشاش برای بهترین مدل (با $k = 3$) در شکل ۲۰ نشان داده شده است.



شکل ۲۰: ماتریس اغتشاش طبقه‌بند KNN برای $k = 3$.

همان‌طور که مشاهده می‌شود، مدل تعادلی مناسب بین حساسیت و ویژگی (Sensitivity/Specificity) برقرار کرده است. مقدار نسبتاً بالای بازخوانی نشان می‌دهد که بیشتر موارد بیماری قلبی به‌درستی شناسایی شده‌اند. در عین حال، تعداد کمی خطای نوع اول (مثبت کاذب) و نوع دوم (منفی کاذب) باقی می‌ماند که با توجه به هم‌پوشانی توزیع‌ها در تحلیل اکتشافی، به‌ویژه در ویژگی‌هایی مانند age و chol که جداسازی ضعیف‌تری دارند، قابل انتظار است. با وجود این، مدل در مجموع عملکرد پیش‌بینی‌کنندهٔ قدرتمندی را با توجه به ماهیت ترکیبی داده‌ها و تعداد محدود ویژگی‌های واقعاً تفکیک‌کننده نشان می‌دهد.

۶ جمع‌بندی

در این پروژه، یک تحلیل داده‌کاوی بر روی مجموعه‌دادهٔ تجمیع‌شدهٔ بیماری قلبی از مخزن UCI Machine Learning Repository انجام شد که هدف آن پیش‌بینی وجود بیماری قلبی با استفاده از طبقه‌بند K-Nearest Neighbors (KNN) بود. مسیر انجام کار شامل پیش‌پردازش داده‌ها، تحلیل اکتشافی جامع (EDA) و پیاده‌سازی یک نسخهٔ ترکیبی از الگوریتم KNN بود که توانایی کار با مجموعه‌ای شامل ویژگی‌های عددی و دسته‌ای را دارد.

در مرحلهٔ پیش‌پردازش، داده‌ها برای استفاده در الگوریتم‌های مبتنی بر فاصله آماده‌سازی شدند. مقادیر گمشده برای ویژگی‌های عددی با میانگین میانه و برای ویژگی‌های دسته‌ای با مد جایگزین شدند. سپس ویژگی‌های عددی با استفاده از روش مقیاس‌گذاری Min–Max نرمال‌سازی شدند تا از غلبهٔ ویژگی‌هایی با دامنهٔ عددی بزرگ‌تر در محاسبهٔ فاصله جلوگیری شود. در نهایت، مجموعه‌داده با نسبت ۸۰/۲۰ به آموزش و آزمون تقسیم شد تا ارزیابی مدل روی داده‌های دیده‌نشده به‌درستی انجام گیرد.

نتایج تحلیل اکتشافی الگوهای بالینی قابل‌توجهی را آشکار کردند. در میان تمام ویژگی‌های عددی، thalach (حداکثر ضربان قلب) و oldpeak (افت قطعهٔ ST ناشی از ورزش) بیشترین ارتباط را با وجود بیماری قلبی نشان دادند. بیماران مبتلا به بیماری قلبی معمولاً مقدار thalach پایین‌تر و مقدار oldpeak بالاتری داشتند و این موضوع منجر به تفکیک‌پذیری بهتر میان دو کلاس

در نمودارهای پراکندگی و نمودارهای جفتی شد. در مقابل، ویژگی‌هایی مانند chol و trestbps ارتباط‌های ضعیف‌تر و پراکنده‌تری داشتند و شامل چند مقدار پرت بالینی قابل‌قبول بودند. همچنین، توزیع نسبتاً متعادل متغیر هدف شرایط مناسبی را برای یادگیری مدل بدون نیاز به روش‌های متوازن‌سازی فراهم کرد.

طبقه‌بند KNN با بهره‌گیری از یک تابع فاصلهٔ ترکیبی اقلیدسی–همینگ به‌صورت کامل پیاده‌سازی شد. عملکرد مدل برای مقادیر مختلف k مورد ارزیابی قرار گرفت و بهترین نتایج در مقدار $k = 3$ حاصل شد. در این مقدار، مدل بالاترین دقت، دقت مثبت، بازخوانی و امتیاز F1 را به دست آورد. این نتیجه با الگوهای یافت‌شده در تحلیل اکتشافی هم‌خوان است؛ زیرا ساختار داده‌ها به‌واسطهٔ چند ویژگی بسیار تفکیک‌کننده ساختاری محلی دارند و مقدارهای کوچک‌تر k بهتر می‌توانند این الگوهای محلی را بدون بیش‌ازحد هموار کردن مرزها ثبت کنند. ماتریس اغتشاش نیز نشان داد که مدل توانایی خوبی در تشخیص صحیح هر دو کلاس دارد و بیشتر موارد بیماری قلبی به‌درستی شناسایی شده‌اند.

به‌طور کلی، این مطالعه نشان می‌دهد که الگوریتم KNN — با وجود سادگی آن — در صورت انجام پیش‌پردازش مناسب و انتخاب تابع فاصلهٔ صحیح، می‌تواند عملکرد پیش‌بینی‌کنندهٔ بسیار خوبی روی داده‌های پزشکی داشته باشد. نتایج همچنین اهمیت ویژگی‌های مرتبط با عملکرد ورزشی مانند thalach و oldpeak را به‌عنوان شاخص‌های کلیدی بیماری قلبی برجسته می‌کند و نشان می‌دهد که ویژگی‌هایی همچون chol و trestbps به‌تنهایی قدرت محدودتری در تمایز بین کلاس‌ها دارند.

برای آینده، پیشنهاد می‌شود از روش‌هایی مانند اعتبارسنجی متقابل (Cross-Validation) برای برآورد پایدارتر عملکرد مدل استفاده شود، وزن‌دهی بین فاصلهٔ عددی و دسته‌ای بهینه‌سازی گردد یا حتی مدل با طبقه‌بندهای پیشرفته‌تر مانند رگرسیون لجستیک، درخت تصمیم، SVM یا روش‌های ensemble مقایسه شود. چنین توسعه‌هایی می‌توانند به بهبود دقت طبقه‌بندی و ارائهٔ بینش‌های عمیق‌تر دربارهٔ روابط میان ویژگی‌ها در پیش‌بینی ریسک بیماری قلبی کمک کنند.

مراجع

[۱] مجموعه‌دادهٔ بیماری قلبی – مخزن داده‌های UCI.
https://archive.ics.uci.edu/ml/datasets/heart+disease