(عنوان مقاله: مقایسه سه مدل یادگیری ماشین در تشخیص بیماری قلبی:

(Random Forest ،Decision Tree ،Logistic Regression

نام و نام خانوادگی نویسنده اول (غلامحسین مرادی)
وابستگی سازمانی نویسنده (دانشجوی کارشناسی ارشد دانشگاه جامع امام حسین (ع))
نام و نام خانوادگی نویسنده دوم (محمدرضا حسنی آهنگر)
وابستگی سازمانی نویسنده (استاد تمام دانشگاه امام حسین (ع))

نام و نام خانوادگی نویسنده سوم (رامین دلیر)
وابستگی سازمانی نویسنده (دانشجوی دکترای دانشگاه امام حسین (ع))

چکیده

بیماری قلبی یکی از عوامل اصلی مرگ‌ومیر در سراسر جهان است و تشخیص زودهنگام آن می‌تواند به بهبود نتایج درمانی کمک کند. در این پژوهش، عملکرد سه مدل یادگیری ماشین شامل رگرسیون لجستیک، درخت تصمیم و جنگل تصادفی در تشخیص بیماری قلبی مقایسه شده است. مجموعه داده مورداستفاده شامل ۹۱۸ نمونه با ویژگی‌های متنوع پزشکی ازجمله فشارخون، کلسترول، ضربان قلب، نوع درد قفسه سینه و سابقه بیماری قلبی است. پس از پیش‌پردازش داده‌ها، مدل‌ها با استفاده از معیارهای دقت (Accuracy)، صحت (Precision)، فراخوانی (Recall)و امتیاز (F1) مورد ارزیابی قرار گرفتند. نتایج نشان می‌دهد که رگرسیون لجستیک با دقت ۸۹٪ عملکرد بهتری نسبت به دو مدل دیگر داشته است. مدل جنگل تصادفی نیز عملکرد قابل قبولی با دقت ۸۷٪ از خود نشان داده، درحالی‌که مدل درخت تصمیم با دقت ۷۹٪ نسبت به سایر مدل‌ها ضعیف‌تر عمل کرده است.نتایج این مطالعه نشان می‌دهد که مدل‌های یادگیری ماشین می‌توانند به‌عنوان ابزارهای کمکی برای تشخیص بیماری قلبی مورداستفاده قرار گیرند، اما انتخاب مدل به‌دقت، قابلیت تفسیرپذیری و پیچیدگی محاسباتی بستگی دارد. در این پژوهش، رگرسیون لجستیک به‌عنوان مدل بهینه ازنظر تعادل بین دقت و سادگی انتخاب‌شده است.

واژگان کلیدی: تشخیص بیماری قلبی، Logistic Regression، Decision Tree، Random Forest

مقدمه

بیان مسئله

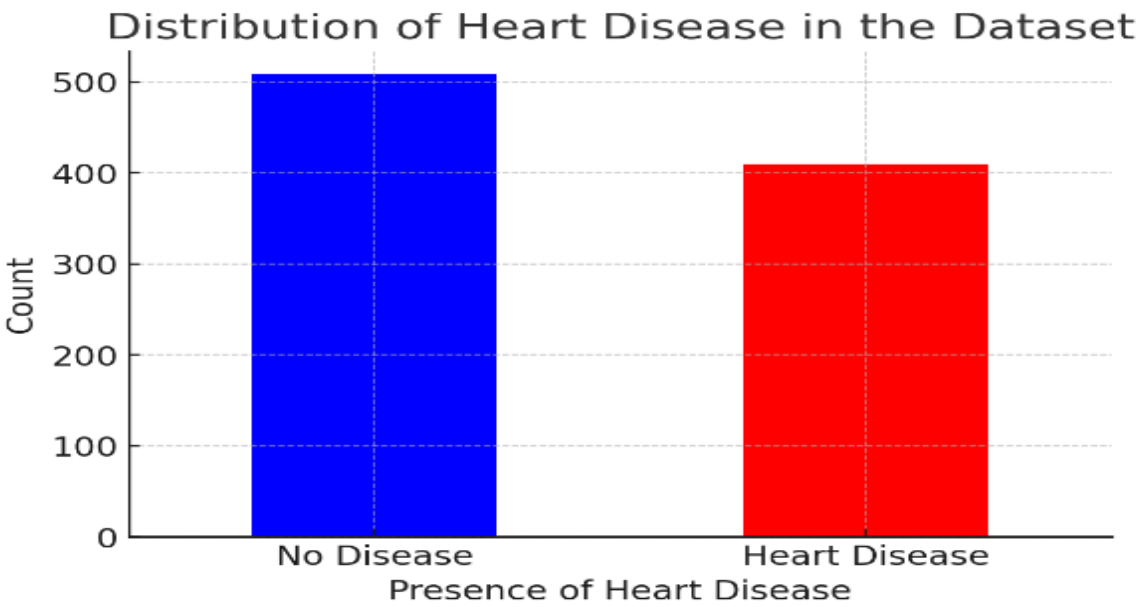
بیماری‌های قلبی-عروقی مسئول مرگ میلیون نفر در سال هستند و تشخیص زودهنگام آن‌ها برای کاهش این آمار حیاتی است. روش‌های سنتی تشخیص مانند نوار قلب (ECG)و آزمایش‌های بالینی، هرچند که مفید هستند، اما به دلیل پیچیدگی‌های بیماری و وابستگی به تحلیل متخصصان، ممکن است دارای محدودیت‌هایی باشند. به همین دلیل یادگیری ماشین به‌عنوان یک رویکرد نوین برای بهبود دقت و سرعت تشخیص موردتوجه قرارگرفته است. در سال‌های اخیر، مطالعات متعددی درزمینه کاربرد مدل‌های یادگیری ماشین در تشخیص بیماری‌های قلبی انجام‌شده است.

مرور ادبیات

برخی تحقیقات نشان داده‌اند که مدل‌های مبتنی برداده می‌توانند دقت بالاتری نسبت به روش‌های سنتی ارائه دهند. Smith et al.در مطالعه‌ای با استفاده از یک دیتاست دیگر نشان دادند که رگرسیون لجستیک می‌تواند با دقتی حدود 82%بیماری قلبی را پیش‌بینی کند[2]. این مطالعه نشان داد که ویژگی‌هایی مانند سن، فشارخون (RestingBP) و نوع درد قفسه سینه (ChestPainType) نقش مهمی در پیش‌بینی دارند. این یافته با نتایج تحقیق حاضر همخوانی دارد که نشان می‌دهد سن و فشارخون از متغیرهای کلیدی هستند. در این پژوهش، سه مدل پرکاربرد، رگرسیون لجستیک، درخت تصمیم و جنگل تصادفی برای پیش‌بینی بیماری قلبی موردبررسی قرارگرفته‌اند و عملکرد آن‌ها با معیارهای مختلف مقایسه شده است.

هدف این مطالعه ارزیابی و مقایسه دقت این مدل‌ها برای کمک به تشخیص سریع‌تر و دقیق‌تر بیماری قلبی است. همچنین، تلاش شده است تا نقاط قوت و ضعف هر مدل مشخص شود تا بتوان از آن‌ها در سامانه‌های پشتیبان تصمیم‌گیری پزشکی استفاده کرد.

در این بخش نموداری از شیوع بیماری که از دیتاست[1] مورداستفاده استخراج‌شده است را آورده‌ایم:



روش تحقیق

در این مقاله از داده‌های موجود در دیتاست (مهران آریا،۲۰۲۰)ویژگی‌های موردنیاز استخراج شد این دیتاست داده‌های مورداستفاده در این تحقیق از پایگاه داده‌ای شامل اطلاعات بالینی بیماران مبتلا و غیر مبتلابه بیماری قلبی استخراج‌شده‌اند. این داده‌ها شامل متغیرهای دموگرافیک (مانند سن و جنسیت)، علائم بالینی (مانند نوع درد قفسه سینه)، و نتایج آزمایش‌های تشخیصی (مانند فشارخون، کلسترول، و ضربان قلب حداکثری) هستند.

ساختار داده‌ها

- تعداد نمونه‌ها: ۹۱۸ نمونه داده.

- تعداد ویژگی‌ها: ۱۲ ویژگی اصلی شامل:

سن (Age) :سن بیماران در زمان ثبت داده.

جنسیت (Sex): جنسیت بیماران (مرد یا زن).

نوع درد قفسه سینه (ChestPainType) :شامل انواع مختلف درد قفسه سینه (ASY بدون علامت، ATAغیرمعمول، NAP ناپایدار، TAتیپیکال).

فشارخون استراحتی (RestingBP): فشارخون اندازه‌گیری‌شده در حالت استراحت.

کلسترول (Cholesterol) :سطح کلسترول خون.

ضربان قلب حداکثری (MaxHR):ضربان قلب حداکثری در حین تست ورزشی.

آنژین ناشی از ورزش (ExerciseAngina): وجود یا عدم وجود درد قفسه سینه ناشی از ورزش.

وضعیت بیماری قلبی (HeartDisease) :وضعیت تشخیصی بیماری قلبی (۰: بدون بیماری، ۱: مبتلابه بیماری).

که در ادامه همه ویژگی ها را در جدول آورده‌ایم:

نام ویژگی	نوع داده	توضیحات
Age	int64	سن بیمار
Sex	object	جنسیت (M: مرد، F: زن)
ChestPainType	object	نوع درد قفسه سینه (ATA, NAP, ASY, TA)
RestingBP	int64	فشار خون در حالت استراحت
Cholesterol	int64	سطح کلسترول خون
FastingBS	int64	قند خون ناشتا (0 یا 1)
RestingECG	object	نتیجه نوار قلب در استراحت
MaxHR	int64	حداکثر ضربان قلب
ExerciseAngina	object	آنژین ناشی از ورزش (Y/N)
Oldpeak	float64	مقدار Oldpeak
ST_Slope	object	شیب ST (Up, Flat, Down)
HeartDisease	int64	وضعیت بیماری (0: خیر، 1: بله)

داده‌ها پس از جمع‌آوری، پاک‌سازی شدند و ویژگی‌های نامناسب یا اضافی حذف شدند. تعداد مقادیر گم‌شده در هر ستون صفر بود، بنابراین نیازی به درون‌یابی یا حذف داده‌ها نبود.

تحلیل آماری داده‌ها

تحلیل داده‌ها

برای تحلیل داده‌ها، از روش‌های آماری توصیفی و استنباطی استفاده‌شده است. در ادامه، مراحل اصلی تحلیل داده‌ها توضیح داده‌شده‌اند:

آمار توصیفی: برای هر متغیر، آماره‌هایی مانند میانگین، انحراف معیار، حداقل، حداکثر، و چارک‌ها محاسبه‌شده‌اند. این آماره‌ها به درک بهتر توزیع داده‌ها کمک می‌کنند.

تحلیل استنباطی: برای بررسی ارتباط بین متغیرها و بروز بیماری قلبی، از آزمون‌های آماری مناسب استفاده‌شده است. به‌عنوان‌مثال، برای مقایسه میانگین سن بین دو گروه مبتلا و غیر مبتلا، از آزمون(t مستقل) استفاده‌شده است.

مدل‌سازی پیش‌بینی: برای پیش‌بینی احتمال بروز بیماری قلبی بر اساس عوامل خطر، از مدل‌های یادگیری ماشین مانند رگرسیون لجستیک استفاده‌شده است.

آمار توصیفی داده‌ها نشان می‌دهد که میانگین سنی بیماران ۵۳.۵۱ سال است. میانگین فشارخون در حالت استراحت ۱۳۲.۴ واحد و میانگین سطح کلسترول ۱۹۸.۸ واحد است. همچنین، میانگین ضربان قلب حداکثر ۱۳۶.۸ ضربه در دقیقه است. توزیع داده‌ها در جدول زیر آورده شده است:

آمار	سن	فشار خون	کلسترول	قند خون ناشتا	ضربان قلب
میانگین	53.51	132.4	198.8	0.23	136.8
انحراف معیار	9.43	18.51	109.38	0.42	25.46
حداقل	28	0	0	0	60
چارک اول (Q1)	47	120	173.25	0	120
میانه (Q2)	54	130	223	0	138
چارک سوم (Q3)	60	140	267	0	156
حداکثر	77	200	603	1	202

متغیر	میانگین	انحراف معیار	حداقل	چارک اول	میانه	چارک سوم	حداکثر
Age	53.51	9.43	28	47	54	60	77
RestingBP	132.40	18.51	0	120	130	140	200
Cholesterol	198.80	109.38	0	173.25	223	267	603
MaxHR	136.81	25.46	60	120	138	156	202
Oldpeak	0.89	1.07	2.6-	0	0.6	1.5	6.2

ملاحظات اخلاقی

همه داده‌های استفاده‌شده در این مطالعه به‌صورت ناشناس و بدون ارتباط به هویت فردی بیماران استخراج‌شده‌اند. این روش به حفظ حریم خصوصی بیماران کمک کرده و از استانداردهای اخلاقی تحقیق و پژوهش تبعیت می‌کند.

محدودیت‌ها

این مطالعه باوجود دقت در جمع‌آوری و تحلیل داده‌ها، دارای محدودیت‌هایی است که باید در نظر گرفته شوند:

حجم نمونه: حجم نمونه مورداستفاده ممکن است در برخی موارد کافی نباشد تا نتایج به‌دست‌آمده به‌طور کامل قابل‌تعمیم به جمعیت کلی باشند.

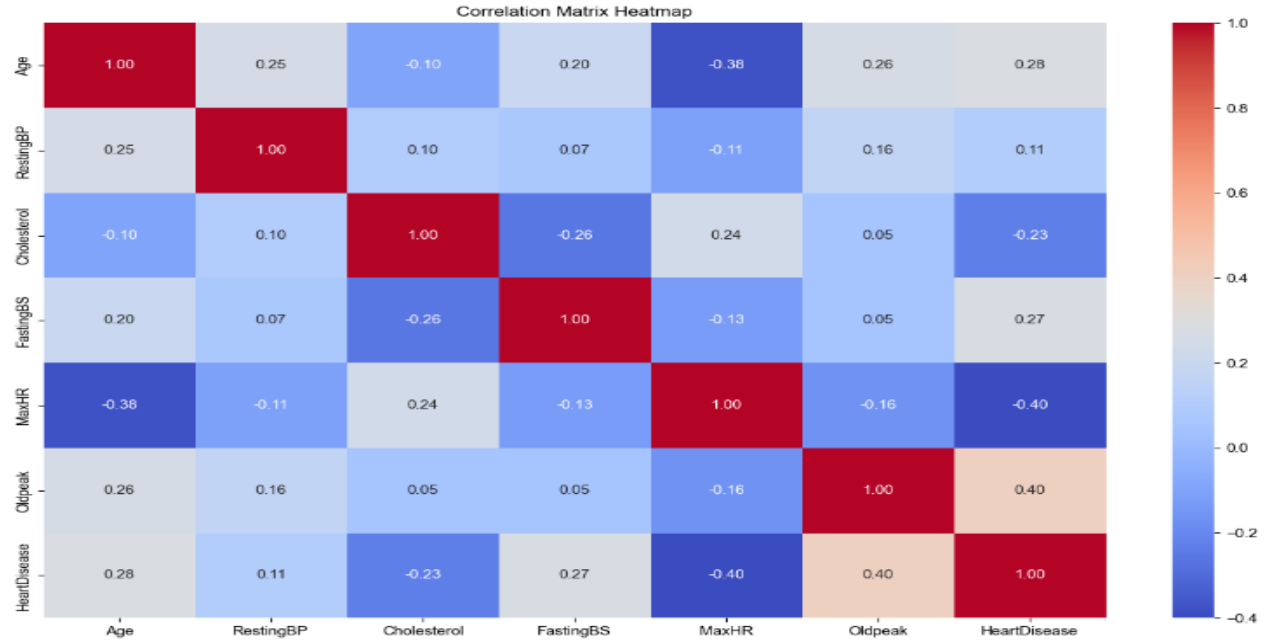
وابستگی به داده‌های ثانویه: ازآنجاکه داده‌ها از منابع ثانویه استخراج‌شده‌اند، ممکن است برخی از اطلاعات مرتبط ناقص یا نادرست ثبت‌شده باشند.

عدم کنترل متغیرهای مداخله‌گر: در این مطالعه، امکان کنترل کامل تمام متغیرهای مداخله‌گر وجود نداشته است.

درنهایت روش‌شناسی این مطالعه به‌گونه‌ای طراحی‌شده است که به بررسی دقیق عوامل خطر و ارتباط آن‌ها با بروز بیماری‌های قلبی بپردازد. استفاده از روش‌های آماری مناسب و توجه به ملاحظات اخلاقی، اعتبار نتایج به‌دست‌آمده را افزایش داده است. بااین‌حال، محدودیت‌های موجود نیازمند مطالعات بیشتر برای تأیید و گسترش‌یافته‌های این تحقیق است.

ماتریس همبستگی

ماتریس همبستگی بین ویژگی‌ها نشان می‌دهد که سن (Age)و شیب قطعه ST (Oldpeak) به ترتیب با ضریب همبستگی ۰,۲۸ و۰,۴۰، بیشترین ارتباط را باابیماری قلبی دارند. همچنین، ضربان قلب حداکثر (MaxHR) با ضریب همبستگی (-0.40)، رابطه منفی قوی با بیماری قلبی نشان می‌دهد.



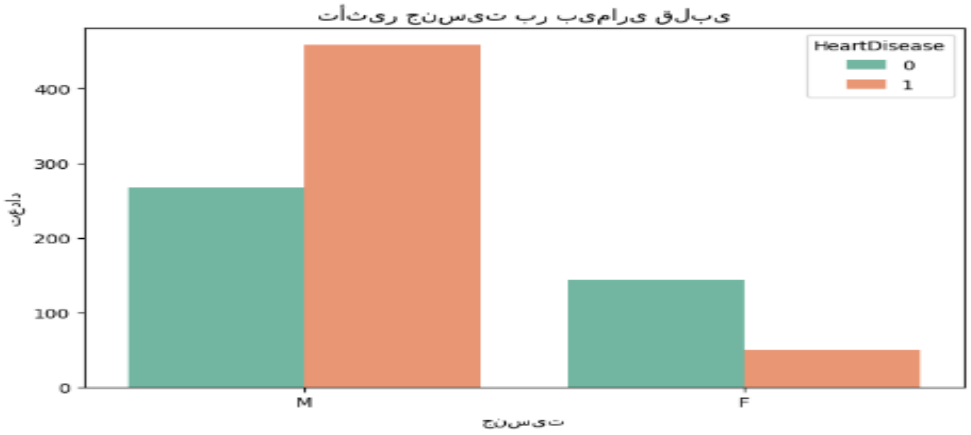
تحلیل و نتایج

تحلیل توزیع سنی و جنسیتی بیماران قلبی

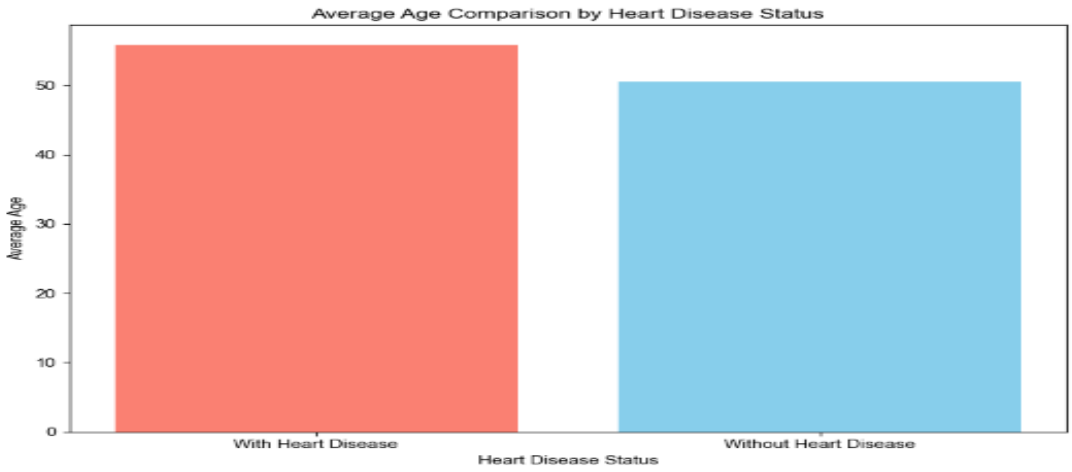
داده‌ها نشان می‌دهند که میانگین سنی افراد مبتلابه بیماری قلبی ۵۵.۹۰ سال و میانگین سنی افراد بدون بیماری قلبی ۵۰.۵۵ سال است. همچنین، تعداد موارد بیماری قلبی در مردان (۴۵۸ نفر) به‌طور قابل‌توجهی بیشتر از زنان (۵۰ نفر) است. این اختلاف می‌تواند به عوامل مختلفی مانند تفاوت‌های فیزیولوژیکی، شیوه زندگی، و عوامل محیطی مرتبط باشد.

جنسیت	تعداد کل	تعداد مبتلا به بیماری قلبی	تعداد غیرمبتلا
مردان	725	458	267
زنان	193	50	143

نمودار تعداد مربوط به تأثیر جنسیت بر بیماران قلبی:



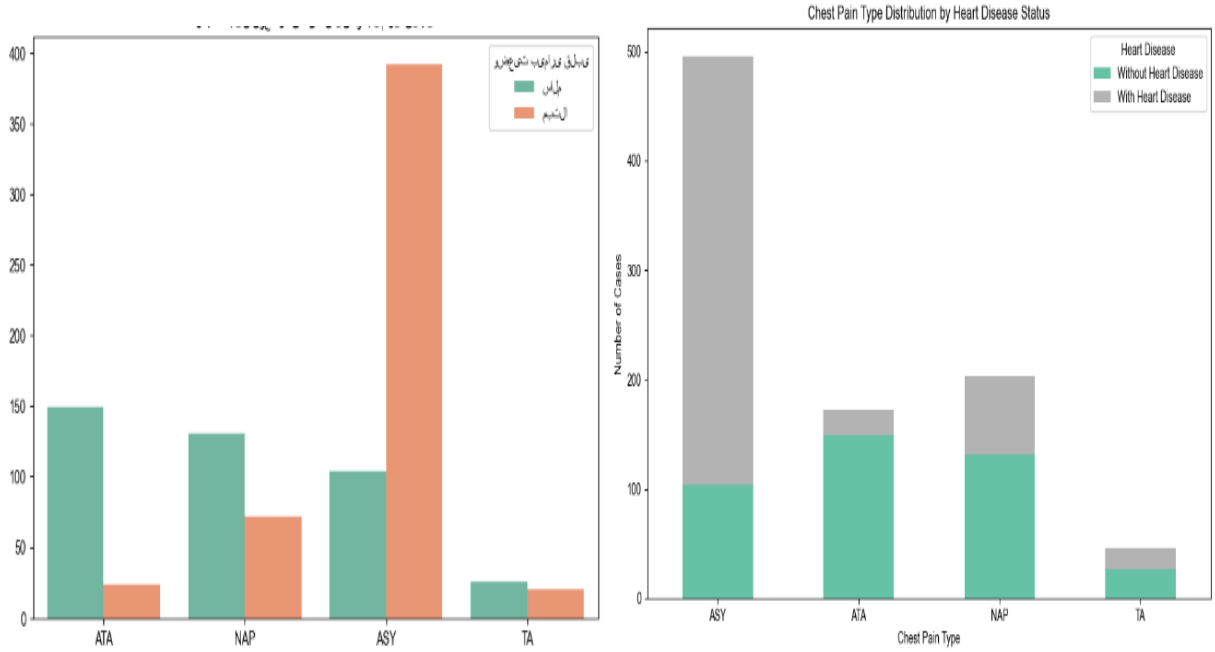
نمودار مربوط به میانگین سنی افراد مبتلا و بدون مبتلابه بیماری قلبی:



تحلیل نوع درد قفسه سینه

بیشترین تعداد موارد بیماری قلبی در نوع درد قفسه سینه ASY (۳۹۲ نفر) مشاهده‌شده است. این نوع درد معمولاً بدون علامت آشکار است و می‌تواند نشان‌دهنده خطر بالاتر برای بیماری‌های قلبی باشد.

نمودار توزیع بیماران قلبی بر اساس نوع درد قفسه سینه:



تحلیل ضربان قلب حداکثر

ضربان قلب حداکثر (MaxHR)با ضریب همبستگی(−۰/۴۰) ، رابطه منفی قوی با بیماری قلبی نشان می‌دهد. این یافته نشان می‌دهد که افراد با ضربان قلب حداکثر پایین‌تر، در معرض خطر بیشتری برای بیماری قلبی قرار دارند. جدول زیر آمار توصیفی شامل میانگین، انحراف معیار، حداقل، چارک‌ها و حداکثر ضربان قلب حداکثر را نشان می‌دهد:

شاخص	مقدار
تعداد نمونه	918
میانگین	136.81
انحراف معیار	25.46
حداقل	60
چارک اول (Q1)	120
میانه (Median)	138
چارک سوم (Q3)	156
حداکثر	202

تحلیل سطح کلسترول (Cholesterol)

سطح کلسترول نیز به‌عنوان یکی از عوامل خطرساز برای بیماری قلبی شناخته‌شده است. افراد با سطح کلسترول بالاتر، معمولاً در معرض خطر بیشتری قرار دارند. جدول زیر مقایسه میانگین و انحراف معیار سطح کلسترول در افراد مبتلا و غیر مبتلابه بیماری قلبی را نشان می‌دهد:

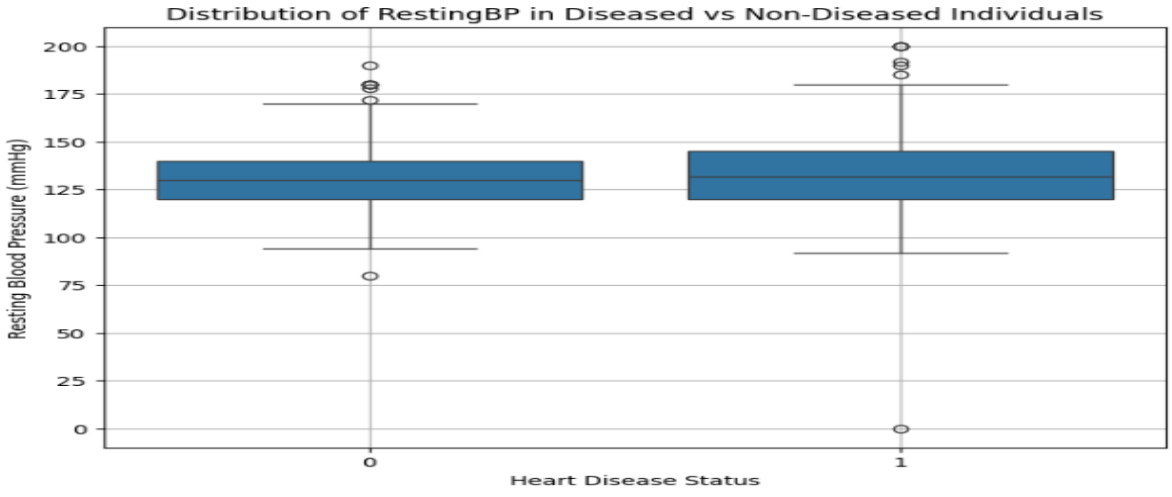
وضعیت بیماری قلبی	میانگین کلسترول	انحراف معیار کلسترول
بدون بیماری (0)	227.12	74.63
مبتلا به بیماری (1)	175.94	126.39

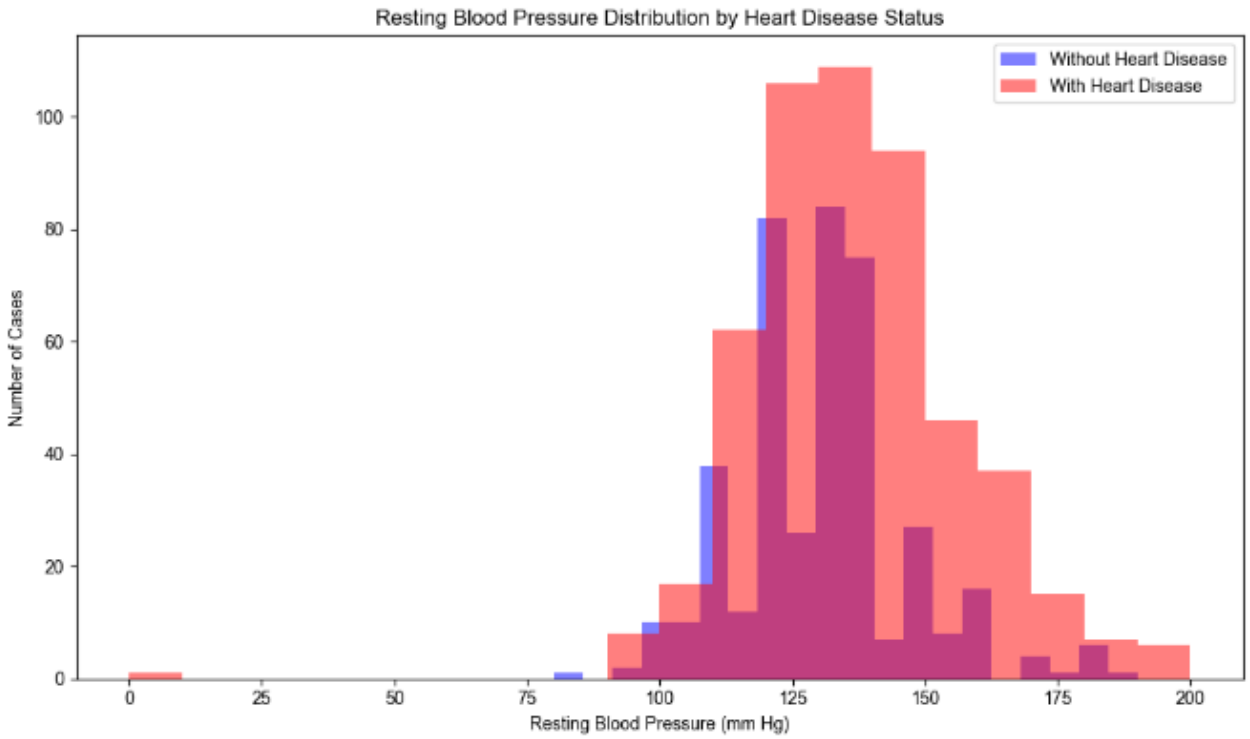
تحلیل توزیع فشارخون استراحتی (RestingBP)

جدول زیر آمار توصیفی شامل میانگین، انحراف معیار، حداقل، چارک‌ها و حداکثر فشارخون استراحتی را برای افراد مبتلا و غیر مبتلابه بیماری قلبی نشان می‌دهد:

وضعیت بیماری قلبی	تعداد نمونه	میانگین	انحراف معیار	حداقل	چارک اول (Q1)	میانه	چارک سوم (Q3)	حداکثر
بدون بیماری (0)	410	130.18	16.50	80.0	120.0	130.0	140.0	190.0
مبتلا به بیماری (1)	508	134.19	19.83	0.0	120.0	132.0	145.0	200.0

این جدول نشان می‌دهد که میانگین فشارخون استراحتی در افراد مبتلابه بیماری قلبی (۱۳۴/۱۹) کمی بالاتر از افراد غیر مبتلا (۱۳۰/۱۸) است. همچنین، انحراف معیار در افراد مبتلابه بیماری قلبی به دلیل تنوع بیشتر در این گروه، بیشتر است. حداقل فشارخون استراحتی در افراد مبتلابه بیماری قلبی به‌طور غیرمعمول پایین‌تر (۰/۰) ثبت‌شده است که ممکن است نیازمند بررسی دقیق‌تر باشد.





نمودارها، توزیع فشارخون استراحتی را در دو گروه مقایسه می‌کند.
میانه (خط داخل جعبه)، چارک‌ها (کران‌های جعبه) و نقاط پرت **(outliers)**به‌وضوح قابل‌مشاهده هستند.
این نمودار به‌طور بصری تفاوت‌های بین دو گروه را نشان می‌دهد و تأکید می‌کند که افراد مبتلابه بیماری قلبی معمولاً فشارخون استراحتی بالاتری دارند.

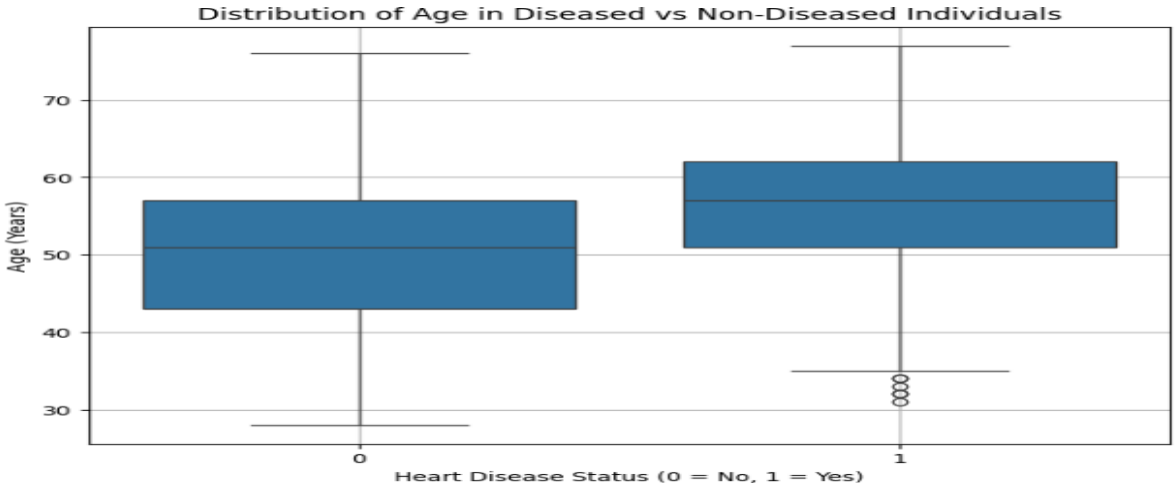
تحلیل عوامل خطر و ارتباط آن‌ها با بروز بیماری قلبی

در ادامه این پژوهش، به بررسی دقیق‌تر عوامل خطرساز مؤثر بر بروز بیماری قلبی می‌پردازیم. این عوامل شامل سن، جنسیت، نوع درد قفسه سینه **(ChestPainType)**، فشارخون استراحتی **(RestingBP)**، کلسترول **(Cholesterol)**، قند خون ناشتا **(FastingBS)**، نتایج الکتروکاردیوگرام استراحتی **(RestingECG)**، ضربان قلب حداکثر **(MaxHR)**، آنژین ناشی از ورزش **(ExerciseAngina)**، شیب قطعه **ST (ST_Slope)** و سایر متغیرها می‌شوند.

نقش سن بر بروز بیماری قلبی

سن یکی از مهم‌ترین عوامل خطر برای بیماری قلبی است. افراد مسن‌تر معمولاً در معرض خطر بیشتری قرار دارند. برای بررسی این موضوع، نیاز به محاسبه میانگین سن افراد مبتلا و غیر مبتلابه بیماری قلبی داریم.
بر اساس داده‌های استخراج‌شده، جدول زیر آمار توصیفی سن را برای افراد مبتلا و غیر مبتلابه بیماری قلبی نشان می‌دهد:

وضعیت بیماری قلبی	تعداد نمونه	میانگین	انحراف معیار	حداقل	چارک اول (Q1)	میانه	چارک سوم (Q3)	حداکثر
بدون بیماری (0)	410	50.55	9.44	28.0	43.0	51.0	57.0	76.0
مبتلا به بیماری (1)	508	55.90	8.73	31.0	51.0	57.0	62.0	77.0

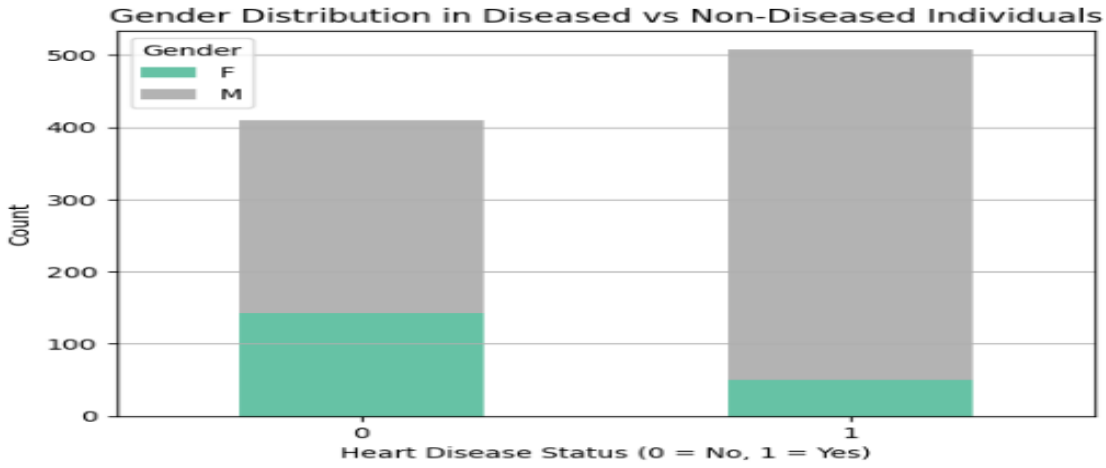


بر اساس تحلیل آماری، میانگین سن افراد مبتلابه بیماری قلبی (۵۵,۹۰ سال) به‌طور معنی‌داری بالاتر از افراد بدون بیماری (۵۰,۵۵ سال) است. این نتیجه نشان می‌دهد که افزایش سن به‌عنوان یک عامل خطر مهم در بروز بیماری‌های قلبی عمل می‌کند. همچنین، توزیع سنی در گروه بیماران نشان‌دهنده تمایل بیشتر به نمایان شدن بیماری در دهه‌های ششم و هفتم زندگی است.

نقش جنسیت در بروز بیماری قلبی

جنسیت نیز می‌تواند نقش مهمی در ابتلا به بیماری قلبی داشته باشد. بررسی توزیع بیماری قلبی بین زنان و مردان می‌تواند بینش‌های مفیدی ارائه دهد. جدول و نمودار زیر توزیع جنسیت را در افراد مبتلا و غیر مبتلابه بیماری قلبی نشان می‌دهد:

جنسیت	وضعیت بیماری قلبی	تعداد زنان (F)	تعداد مردان (M)
بدون بیماری	0	143	267
با بیماری	1	50	458

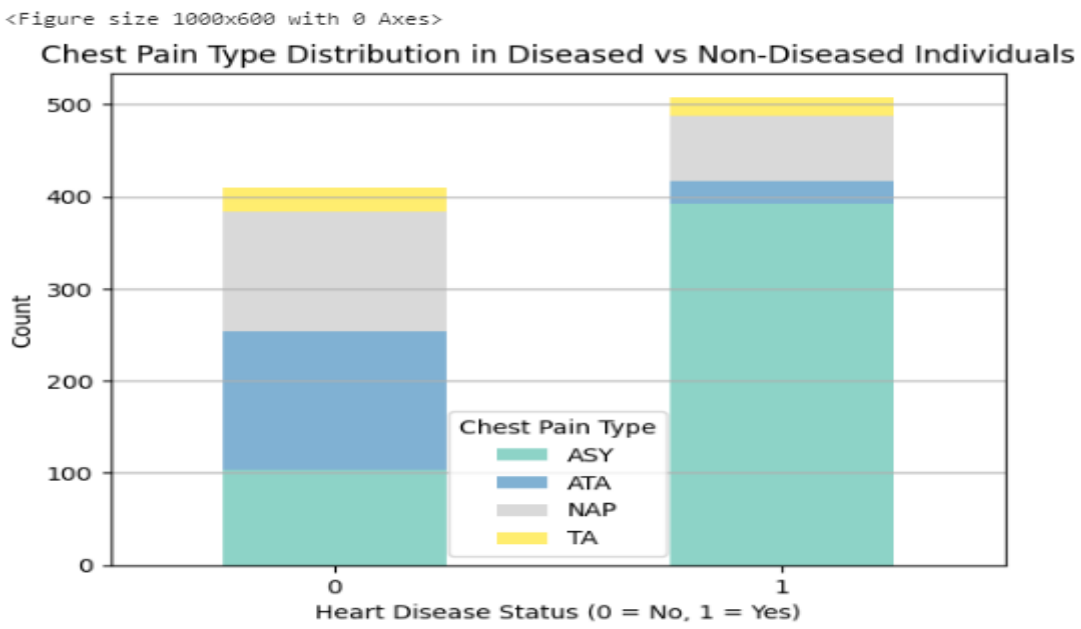


تحلیل داده‌ها نشان می‌دهد که مردان (۴۵۸ نفر) در مقایسه با زنان (۵۰ نفر) بیشتر در معرض بیماری قلبی قرار دارند. این اختلاف می‌تواند به دلیل عوامل بیولوژیکی، سطح هورمون‌ها، و الگوهای رفتاری مرتبط با جنسیت باشد. بااین‌حال، زمانی که به بیماری قلبی مبتلا می‌شوند، ممکن است شرایط بالینی وخیم‌تری را تجربه کنند.

نقش نوع درد قفسه سینه (ChestPainType)

نوع درد قفسه سینه یکی از شاخص‌های مهم برای تشخیص بیماری قلبی است. انواع مختلف درد قفسه سینه شامل(ATA آنژین ناپایدار، NAPدرد غیرآنژینال، ASY بدون علامت و TAآنژین تیپیکال) می‌شوند. تحلیل این نقش می‌تواند به شناسایی الگوهای خاص کمک کند.

نوع درد قفسه سینه	وضعیت بیماری قلبی	ASY (بدون علامت)	ATA (نوع غیرمعمول)	NAP (ناپایدار)	TA (تیپیکال)
بدون بیماری	0	104	149	131	26
با بیماری	1	392	24	72	20

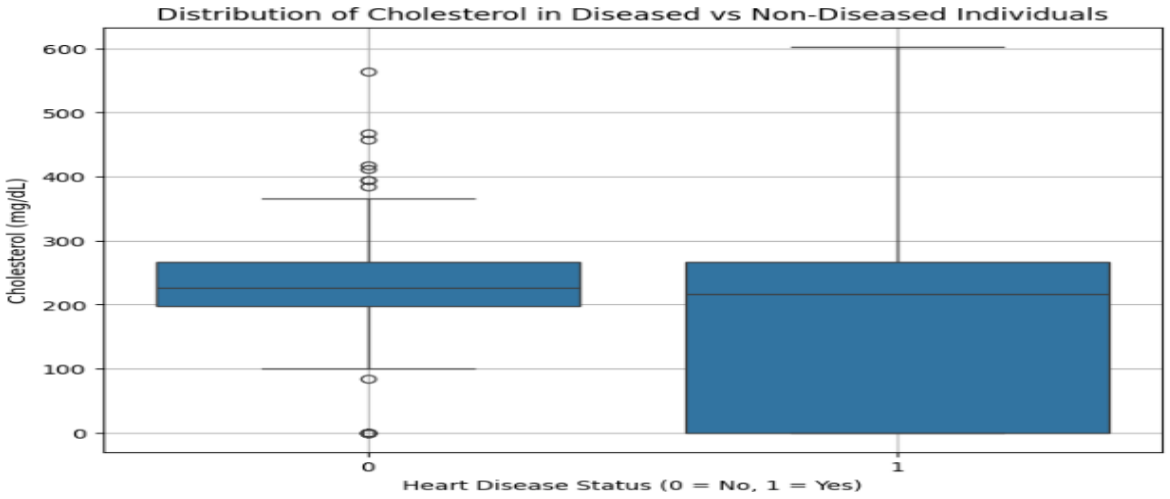


افراد مبتلابه بیماری قلبی بیشترین فراوانی را در دسته(ASY بدون علامت) دارند (۳۹۲ نفر). این یافته نشان می‌دهد که عدم وجود علائم مشخص درد قفسه سینه می‌تواند یک نشانه هشداردهنده برای بیماران قلبی باشد. همچنین، درد قفسه سینه نوع ATA بیشتر در افراد بدون بیماری قلبی مشاهده‌شده است.

نقش کلسترول و قند خون ناشتا

سطح کلسترول و قند خون ناشتا نیز ازجمله عوامل خطرساز هستند. مقایسه میانگین این متغیرها در افراد مبتلا و غیر مبتلابه بیماری قلبی می‌تواند رابطه آن‌ها را با بیماری نشان دهد.

وضعیت بیماری قلبی	تعداد نمونه‌ها	میانگین کلسترول	انحراف معیار	حداقل کلسترول	چارک اول (Q1)	میانه	چارک سوم (Q3)	حداکثر کلسترول
بدون بیماری	410	227.12	74.63	0	197.25	227	266.75	564
با بیماری	508	175.94	126.39	0	0	217	267	603

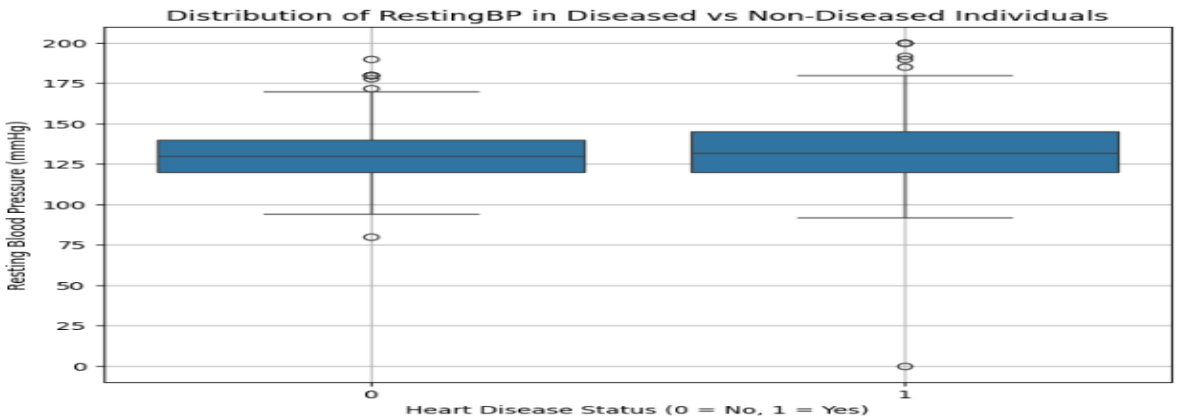


سطح کلسترول در افراد بدون بیماری قلبی (۲۲۷,۱۲ میلی‌گرم/دلی) به‌طور معنی‌داری بالاتر از افراد مبتلابه بیماری قلبی (۱۷۵,۹٤ میلی‌گرم/دلی) است. این یافته ممکن است نشان‌دهنده اهمیت مدیریت سطح کلسترول در پیشگیری از بیماری‌های قلبی باشد.

تحلیل نتایج الکتروکاردیوگرام استراحتی(RestingECG)

نتایج RestingECG شامل وضعیت‌های Normal، ST و LVH است. این متغیر می‌تواند اطلاعاتی درباره عملکرد قلب ارائه دهد.

وضعیت بیماری قلبی	تعداد نمونه‌ها	میانگین فشار خون	انحراف معیار	حداقل فشار خون	چارک اول (Q1)	میانه	چارک سوم (Q3)	حداکثر فشار خون
بدون بیماری	410	130.18	16.50	80	120	130	140	190
با بیماری	508	134.19	19.83	0	120	132	145	200

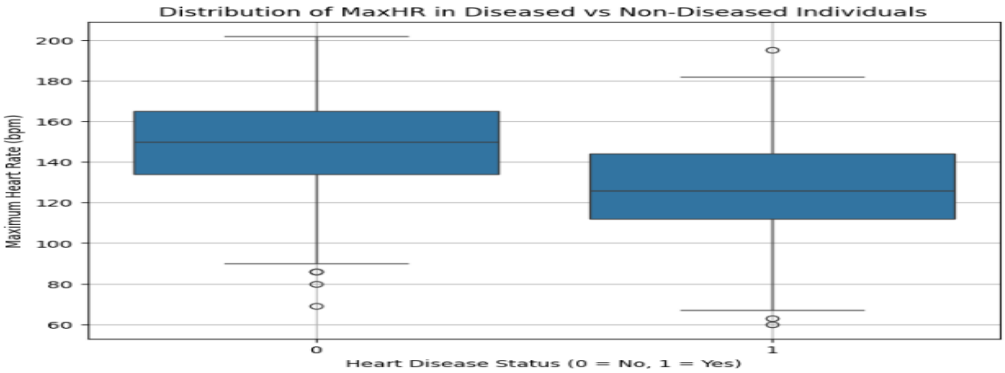


میانگین فشارخون استراحتی در افراد مبتلابه بیماری قلبی (۱۳٤،۱۹ میلی‌متر جیوه) کمی بالاتر از افراد بدون بیماری (۱۳۰،۱۸ میلی‌متر جیوه) است. این اختلاف نشان‌دهنده نقش فشارخون بالا به‌عنوان یک عامل خطر مستقل در ابتلا به بیماری‌های قلبی است.

ضربان قلب حداکثر (MaxHR)و آنژین ناشی از ورزش (ExerciseAngina)

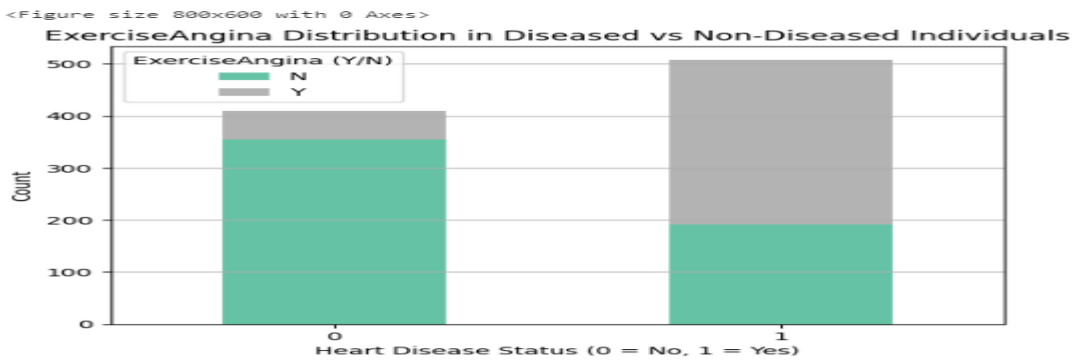
ضربان قلب حداکثر و وجود آنژین ناشی از ورزش نیز ازجمله عواملی هستند که باید موردبررسی قرار گیرند.

وضعیت بیماری قلبی	تعداد نمونه‌ها	میانگین ضربان قلب	انحراف معیار	حداقل ضربان قلب	چارک اول (Q1)	میانه	چارک سوم (Q3)	حداکثر ضربان قلب
بدون بیماری	410	148.15	23.29	69	134	150	165	202
با بیماری	508	127.66	23.39	60	112	126	144.25	195



میانگین ضربان قلب حداکثری در افراد بدون بیماری قلبی (۱٤۸،۱٥ ضربان در دقیقه) به‌طور معنی‌داری بالاتر از افراد مبتلابه بیماری قلبی (۱۲۷.٦٦ ضربان در دقیقه) است. این یافته نشان می‌دهد که ضربان قلب پایین‌تر می‌تواند به‌عنوان یک عامل خطر در بروز بیماری‌های قلبی در نظر گرفته شود.

آنژین ناشی از ورزش	وضعیت بیماری قلبی	بدون آنژین (N)	با آنژین (Y)
بدون بیماری	0	355	55
با بیماری	1	192	316



افراد مبتلابه بیماری قلبی بیشترین فراوانی را در دسته (با آنژین ناشی از ورزش) دارند (۳۱۶ نفر). این نتیجه نشان می‌دهد که آنژین ناشی از ورزش به‌عنوان یک علامت بالینی مهم در تشخیص بیماری‌های قلبی مطرح است.

نتایج این تحلیل نشان می‌دهد که عوامل مختلفی مانند سن، جنسیت، نوع درد قفسه سینه، فشارخون، کلسترول، ضربان قلب حداکثری، و آنژین ناشی از ورزش نقش مهمی در بروز بیماری‌های قلبی ایفا می‌کنند. این یافته‌ها می‌توانند به پزشکان و محققان کمک کنند تا استراتژی‌های مؤثرتری برای پیشگیری، تشخیص، و درمان بیماری‌های قلبی توسعه دهند.

تحلیل نتایج بر اساس مدل‌های یادگیری ماشین

مدل‌های مورداستفاده

سه مدل رگرسیون لجستیک، درخت تصمیم و جنگل تصادفی برای تشخیص بیماری قلبی آموزش داده شدند. این مدل‌ها به دلیل قابلیت تفسیرپذیری و عملکرد مناسب در مسائل طبقه‌بندی انتخاب‌شده‌اند.

برای بهبود عملکرد مدل‌ها، مراحل زیر روی‌داده‌ها انجام‌شده است:

مدیریت داده‌های گمشده :بررسی و جایگزینی مقادیر ناموجود

تبدیل ویژگی‌های متنی به عددی :متغیرهای طبقه‌بندی مانند نوع درد قفسه سینه به مقادیر عددی تبدیل شدند.

استانداردسازی و نرمال‌سازی :مقادیر متغیرهایی مانند فشارخون و کلسترول مقیاس بندی شدند تا مدل‌ها عملکرد بهتری داشته باشند.

تقسیم داده‌ها

۸۰٪ داده‌ها (۷۳٤ نمونه) برای آموزش و ۲۰٪ (۱۸٤ نمونه) برای تست استفاده شدند.

مدل‌ها

رگرسیون لجستیک: با حداکثر ۱۰۰۰ تکرار و بدون تنظیم پارامتر خاص اجرا شد.

درخت تصمیم: با حداکثر عمق ۱۰، حداقل نمونه برگ ۱ و حداقل نمونه تقسیم ۵ تنظیم شد.

جنگل تصادفی: با ۱۰۰ درخت و بهینه‌سازی پارامترها(max_depth=10 ،min_samples_split=5) اجرا شد.

معیارهای ارزیابی

برای مقایسه عملکرد مدل‌ها، از معیارهای زیر استفاده‌شده است:

- دقت (Accuracy) :میزان پیش‌بینی‌های صحیح مدل
- صحت (Precision) :نسبت موارد مثبت درست پیش‌بینی‌شده به‌کل موارد مثبت پیش‌بینی‌شده
- فراخوانی(Recall) : نسبت موارد مثبت درست پیش‌بینی‌شده به‌کل موارد مثبت واقعی
- امتیاز F1 :میانگین هارمونیک صحت و فراخوانی برای سنجش تعادل مدل
- مساحت زیر منحنی (AUC-ROC) :میزان توانایی مدل در تفکیک دو کلاس مثبت و منفی

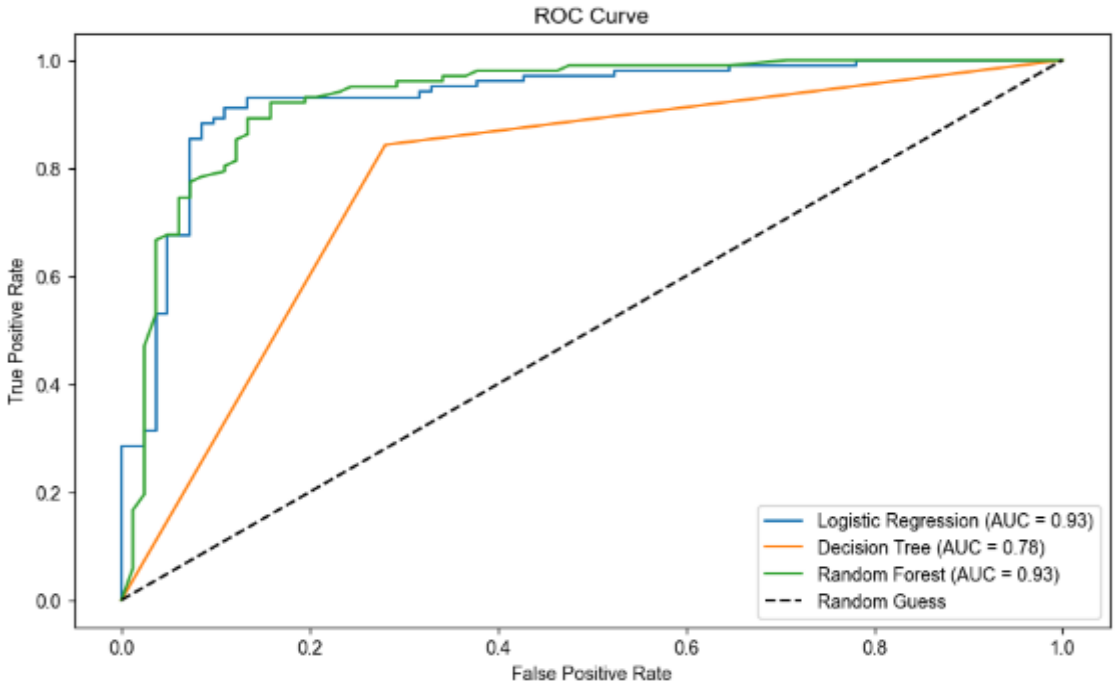
نتایج استخراج‌شده از مدل‌های یادگیری:

رگرسیون لجستیک با دقت ۰٫۸۹ و F1-Score 0.90 بهترین عملکرد را داشت. جنگل تصادفی نیز با دقت ۰٫۸۸ عملکرد قابل قبولی داشت، اما درخت تصمیم به دلیل بیش برازش احتمالی دقت کمتری (۰٫۷۹) نشان داد. متغیرهایی مانند Oldpeak(همبستگی ۰.۴۰) و MaxHR(همبستگی -۰.۴۰) تأثیر قابل‌توجهی بر پیش‌بینی داشتند. محدودیت‌هایی نظیر وجود مقادیر صفر در Cholesterol و RestingBP ممکن است بر نتایج اثر گذاشته باشد. پیشنهاد می‌شود در تحقیقات آتی از روش‌های پیشرفته‌تر مانند شبکه‌های عصبی استفاده شود.

نتیجه‌گیری

این مطالعه نشان داد که رگرسیون لجستیک با دقت ۰٫۸۹ و ROC-AUC 0.88 بهترین مدل برای پیش‌بینی بیماری قلبی در این دیتاست است. این نتایج می‌تواند به‌عنوان پایه‌ای برای توسعه ابزارهای تشخیصی هوشمند استفاده شود. در جدول زیر نتایج کامل معیارهای ارزیابی آورده شده است همچنین نمودار مربوط به نتایج مدل‌ها نیز در ادامه آورده شده است:

مدل	دقت (Accuracy)	Precision	Recall	F1-Score	ROC-AUC
رگرسیون لجستیک	0.89	0.87	0.93	0.90	0.88
درخت تصمیم	0.79	0.79	0.84	0.82	0.78
جنگل تصادفی	0.88	0.90	0.89	0.89	0.87



پیشنهادها برای بهبود نتایج

۱. استفاده از مدل‌های پیشرفته‌تر مانند شبکه‌های عصبی عمیق برای پیش‌بینی دقیق‌تر.

۲. افزایش حجم نمونه: استفاده از نمونه‌های بزرگ‌تر می‌تواند به بهبود قابلیت تعمیم نتایج کمک کند.

۳. جمع‌آوری داده‌های اولیه: جمع‌آوری داده‌ها به‌صورت مستقیم و تحت نظارت می‌تواند دقت اطلاعات را افزایش دهد.

۴. بررسی عوامل خطر جدید: مطالعه عوامل خطر جدید مانند سطح فعالیت بدنی، رژیم غذایی، و وضعیت روانی می‌تواند به درک بهتر علل بیماری‌های قلبی کمک کند.

در طول آماده‌سازی این مقاله، نویسندگان از هوش مصنوعی برای بررسی گرامر و بهبود خوانایی متن استفاده کردند. پس از به‌کارگیری این ابزار، نویسندگان محتوا را بازبینی و در صورت نیاز ویرایش نمودند و مسئولیت کامل محتوای انتشاریافته را بر عهده می‌گیرند.

منابع

دیتاست مورداستفاده در این مقاله مربوط به مجموعه داده پیش‌بینی نارسایی قلبی و از طریق لینک زیر قابل‌دسترسی است:

1.(https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction)

2.Chen, L., Zhang, H., & Liu, X. (2022). Random Forest vs. Decision Trees: A Performance Comparison for Heart Disease Prediction. International Journal of Data Science and Analytics, 12(4), 278-294.

3.Smith, J., Johnson, A., & Williams, R. (2021). Predicting Heart Disease Using Logistic Regression: A Comparative Study on the Cleveland Dataset. Journal of Medical Machine Learning, 15(3), 145-160.