دوره جامع علم داده دانشگاه تهران

پروژه تحلیل داده با پایتون

پیش بینی سکته قلبی براساس مجموعه داده Heart Data

استاد : جناب آقای دکتر امیررضا تجلی

مجری : مریم نوری

پائیز ۱۴۰۳

مقدمه

بیماری‌های قلبی و عروقی یکی از عوامل اصلی مرگ‌ومیر در سراسر جهان هستند. سکته قلبی به‌عنوان یکی از خطرناک‌ترین انواع این بیماری‌ها، به شناسایی و پیشگیری زودهنگام نیاز دارد. روش‌های سنتی تشخیص، مانند آزمایش‌های پزشکی و ارزیابی بالینی، هرچند موثرند، اما گاهی اوقات به دلیل زمان‌بر بودن یا محدودیت در پیش‌بینی خطر، ناکافی می‌شوند.

در این میان، یادگیری ماشین به‌عنوان ابزاری قدرتمند برای تحلیل داده‌های پیچیده و چندبعدی، توانایی شگفت‌انگیزی در شناسایی الگوها و پیش‌بینی بیماری‌ها از خود نشان داده است. این تکنیک‌ها می‌توانند به پزشکان در تشخیص زودهنگام و تصمیم‌گیری‌های درمانی کمک کنند و خطرات احتمالی را کاهش دهند.

هدف این پروژه، بررسی کارایی این الگوریتم‌ها در تحلیل داده‌ها و ارائه مدلی دقیق برای پیش‌بینی سکته قلبی است. برای پیش‌بینی خطر سکته قلبی در بخش مدلسازی، از الگوریتم‌های مختلف یادگیری ماشین نظیر رگرسیون لجستیک، درخت‌های تصمیم، جنگل‌های تصادفی، و الگوریتم‌های مبتنی بر شبکه عصبی استفاده شد. فرآیند مدیریت داده‌ها بر اساس متد CRISP-DM (Cross Industry Standard Process for Data Mining)، انجام شد.

فرآیند مدیریت داده‌ها و مدلسازی بر اساس متد CRISP-DM (Cross Industry Standard Process for Data Mining)، مطابق مراحل زیر انجام شد:

۱. درک مجموعه داده
۲. پاکسازی داده که شامل مراحل زیر است:
   - تشخیص داده‌های گمشده و مدیریت آن‌ها.
   - شناسایی داده‌های تکراری و حذف آن‌ها.
   - تشخیص داده‌های نویز (Noise) و مدیریت آن‌ها.
   - تشخیص داده‌های پرت (Outliers) و حذف آن‌ها.
   - تشخیص داده های نامتقارن و مدیریت آنها.
۳. آماده سازی داده که شامل مراحل زیر است:
۴. استانداردسازی (Standardization) یا نرمال‌سازی (Normalization)
۵. رمزگذاری ویژگی‌های دسته‌ای (Categorical Encoding)
۶. مدل‌سازی (Modeling)
۷. ارزیابی مدل (Evaluation)

نتایج نهایی نشان می‌دهد که با استفاده از روش‌های مناسب پیش‌پردازش داده و انتخاب مدل بهینه، می‌توان مدلی موثر برای پیش‌بینی بیماری قلبی ایجاد کرد که از دقت و اعتمادپذیری بالایی برخوردار باشد.

متد کریسپ(CRISP-DM)

متد کریسپ (CRISP-DM) یک چارچوب منظم و جامع برای اجرای پروژه‌های داده‌کاوی است که به طور گسترده‌ای در صنایع مختلف، از جمله پزشکی، مورد استفاده قرار می‌گیرد. این متد، با فراهم کردن یک ساختار مشخص برای انجام پروژه‌ها، به تحلیلگران داده کمک می‌کند تا به نتایج دقیق‌تر و قابل اطمینان‌تری دست یابند.

پیش‌بینی سکته قلبی، یک مسئله حیاتی در حوزه پزشکی است. با استفاده از متد CRISP-DM، می‌توانیم به صورت سیستماتیک و دقیق، داده‌های جمع‌آوری شده مربوط به بیماران را ، پردازش و تحلیل کنیم تا الگوها و روابطی را که ممکن است به پیش‌بینی وقوع سکته قلبی کمک کنند، شناسایی نماییم. این امر به پزشکان اجازه می‌دهد تا اقدامات پیشگیرانه لازم را انجام داده و از بروز این عارضه خطرناک جلوگیری کنند.

مراحل متد CRISP-DM در پیش‌بینی سکته قلبی:

۱. درک کسب‌وکار:

- تعریف مسئله: هدف اصلی در این مرحله، تعریف دقیق مسئله پیش‌بینی سکته قلبی است. این شامل شناسایی عوامل خطر، تعیین گروه هدف و مشخص کردن معیارهای موفقیت مدل است.
- درک نیازهای کسب‌وکار: نیازهای پزشکان، بیماران و سازمان‌های بهداشتی باید در نظر گرفته شود تا مدل ایجاد شده کاربردی و مفید باشد.

۲. درک داده:

- جمع‌آوری داده: داده‌های مورد نیاز از منابع مختلف مانند پرونده‌های پزشکی، آزمایشگاه‌ها و بانک‌های اطلاعاتی جمع‌آوری می‌شوند.
- ارزیابی کیفیت داده: داده‌ها از نظر کامل بودن، دقت، سازگاری و مرتبط بودن ارزیابی می‌شوند.
- شناسایی متغیرها: متغیرهای مهمی که بر وقوع سکته قلبی تأثیر می‌گذارند (مانند سن، جنسیت، فشار خون، کلسترول، سابقه خانوادگی) شناسایی می‌شوند.

۳. آماده‌سازی داده:

- پاکسازی داده: داده‌های ناقص، تکراری و ناسازگار حذف یا اصلاح می‌شوند.
- تبدیل داده: داده‌ها به فرمتی مناسب برای تحلیل تبدیل می‌شوند (مثلاً تبدیل متغیرهای کیفی به کمی).
- انتخاب ویژگی‌ها: ویژگی‌های مهمی که بیشترین تأثیر را بر پیش‌بینی دارند، انتخاب می‌شوند.

۴. مدل‌سازی:

- انتخاب الگوریتم: الگوریتم‌های مناسب برای پیش‌بینی (مانند درخت تصمیم، شبکه‌های عصبی، ماشین‌های بردار پشتیبان) انتخاب می‌شوند.
- ساخت مدل: مدل بر اساس داده‌های آموزشی ساخته می‌شود.
- تعیین پارامترها: پارامترهای مدل بهینه می‌شوند تا بهترین عملکرد را داشته باشد.

۵. ارزیابی:

- تعیین معیارهای ارزیابی: معیارهایی مانند دقت، حساسیت، ویژگی و AUC برای ارزیابی عملکرد مدل انتخاب می‌شوند.
- ارزیابی مدل: مدل بر روی داده‌های آزمایشی ارزیابی می‌شود تا عملکرد آن در دنیای واقعی تخمین زده شود.

۶. استقرار:

- توسعه سیستم: مدل نهایی در یک سیستم کاربردی برای استفاده در محیط بالینی پیاده‌سازی می‌شود.
- نظارت و نگهداری: عملکرد مدل به طور مداوم نظارت می‌شود و در صورت نیاز به روزرسانی می‌شود.

در این بخش سعی شد یک دید کلی از مراحل کار در یک پروژه واقعی ارائه شود. عمده فعالیت صورت گرفته در این پژوهش با مرحله ارزیابی کیفیت داده شروع و در نهایت به ارزیابی مدل ختم می شود و عملیاتی در زمینه جمع آوری داده ها و همچنین استقرار مدل صورت نپذیرفته است.

در ادامه مراحل متد Crisp-DM در پروژه پیش بینی بیماری قلبی به تفصیل ارائه می‌گردد.

۱. درک مجموعه داده

بررسی اجمالی مجموعه داده Heart data

مجموعه داده مورد بررسی در این پروژه شامل ویژگی‌های مختلف مرتبط با سلامت قلب افراد است، از جمله سن، جنسیت، سطح کلسترول و فشار خون. تحلیل ابتدایی نشان داد که برخی ویژگی‌ها مانند شیب تست‌های استرس قلبی (slope) یا تعداد عروق کرونری (ca) اطلاعات ارزشمندی ارائه می‌دهند، اما مقادیر گمشده در آنها ممکن است عملکرد مدل را محدود کند. اضافه کردن فیلدهایی مانند سابقه خانوادگی یا شاخص توده بدنی می‌توانست پیش‌بینی را بهبود دهد.

مجموعه داده Heart data یک فایل CSV، شامل ۵۹۷ ردیف و ۱۴ ستون است که اطلاعات مختلفی در مورد بیماران و ویژگی‌های مرتبط با سلامت قلب آنها ارائه می‌دهد. برخی از نکات کلیدی در مورد مجموعه داده عبارت‌اند از:

ستون‌ها و نوع داده‌ها:

- ویژگی‌هایی مانند سن، جنسیت، نوع درد قفسه سینه، فشار خون، سطح کلسترول و غیره.
- ستون هدف (احتمال سکته قلبی) به نظر می‌رسد در ستون c باشد. توزیع متغیر هدف (c):
  - ٪۵۹ افراد سالم (۰)
  - ٪۴۱ افراد مبتلا به سکته قلبی (۱)

مقادیر گمشده:

- ستون‌هایی مثل slope، ca و thal مقادیر گمشده قابل‌توجهی دارند، که ممکن است بر عملکرد مدل تاثیرگذار باشد.
- برخی دیگر از ستون‌ها مانند فشار خون و سطح کلسترول نیز دارای مقادیر گمشده جزئی هستند.

بهبود ممکن در ویژگی‌ها:

اطلاعات بیشتری مثل سابقه خانوادگی بیماری‌های قلبی، عادت‌های غذایی، شاخص تـوده بـدنی (BMI) یـا وضـعیت فعالیـت بـدنی منظم می‌توانست برای پیش‌بینی بهتر مفید باشد.

```
====================================
Step: Analyzing Initial Data
====================================

Data Analysis - Before Cleaning
Rows: 597
Columns: 14
Missing Values: 787
Duplicate Rows: 1
Categorical Features: ['sex', 'chest pain', 'blood sugar', 'electrocardiographic ', 'exercise induced', 'slope', 'ca', 'thal', 'c']
Numerical Features: ['heart rate', 'Age (age in year)', 'blood pressure', 'depression ', 'cholestoral ']
```

فیلد های مجموعه داده Heart data

این فیلدها به‌صورت مستقیم یا غیرمستقیم تأثیرگذار بر پیش‌بینی بیماری‌های قلبی هستند و با ترکیب آنها می‌توان الگوریتمی موثر برای تحلیل و پیش‌بینی خطر سکته قلبی توسعه داد. اضافه کردن ویژگی‌هایی مانند شاخص تـوده بـدنی (BMI)، سـابقه خـانوادگی بیماری قلبی، سیگار کشیدن، الگوی خواب یا عادات غذایی می‌توانست به مدل کمک کند تا پیش‌بینی دقیق‌تری ارائه دهد.

۱. Age (سن)

افزایش سن یکی از عوامل خطر اصلی در بیماری‌های قلبی است، زیرا با گذر زمان شریان‌ها سفت‌تر شده و احتمـال گرفتگـی عـروق بیشتر می‌شود. این فیلد به‌صورت سال اندازه‌گیری شده و نشان‌دهنده سن فرد در هنگام جمع‌آوری داده است.

۲. Sex (جنسیت)

جنسیت نقش مهمی در بروز بیماری‌های قلبی دارد. مردان در سنین پایین‌تر نسبت به زنان بیشتر در معرض خطر هستند، اما پـس از یائسگی، خطر برای زنان نیز افزایش می‌یابد. این فیلد به‌صورت داده باینری (۱ برای مرد و ۰ برای زن) ثبت شده است.

۳. Chest Pain (درد قفسه سینه)

نوع درد قفسه سینه می‌تواند نشان‌دهنده وجود یا عدم وجود بیماری‌های قلبی باشد. دردهای مرتبط با آنژین معمولاً به‌دلیل کـاهش خون‌رسانی به عضله قلب اتفاق می‌افتد. این متغیر به چهار دسته تقسیم شده و بر اساس معاینه بالینی تعیین می‌شود.

- مقدار ۱: درد قفسه سینه معمولی (Typical Angina).
- مقدار ۲: درد غیرمعمول (Atypical Angina).
- مقدار ۳: درد بدون ارتباط با قلب (Non-Anginal Pain).
- مقدار ۴: بدون درد قفسه سینه.

۴. Blood Pressure (فشار خون)

فشار خون بالا (هایپرتانسیون) یکی از مهم‌ترین عوامل خطر برای بیماری‌های قلبی است، زیرا می‌تواند به دیواره رگ‌ها آسیب برساند و فشار بیشتری بر قلب وارد کند. این فیلد فشار خون سیستولیک را به میلی‌متر جیوه (mmHg) نشان می‌دهد و با دستگاه فشارسنج اندازه‌گیری می‌شود.

۵. Cholesterol (کلسترول)

سطح کلسترول بالا می‌تواند منجر به تجمع پلاک در عروق و افزایش خطر سکته قلبی شود. این فیلد سطح کلسترول کل خون را به میلی‌گرم بر دسی‌لیتر (mg/dL) نشان می‌دهد و از طریق آزمایش خون اندازه‌گیری می‌شود.

۶. Blood Sugar (قند خون ناشتا)

قند خون بالا، به‌ویژه در افراد دیابتی، خطر بیماری‌های قلبی را افزایش می‌دهد. مقدار این فیلد نشان‌دهنده وجود قند خون بالای ۱۲۰ mg/dL (۱: بله، ۰: خیر) است و از طریق آزمایش قند خون ناشتا تعیین می‌شود.

- مقدار ۱: بالاتر از ۱۲۰ (هایپرگلیسمی).
- مقدار ۰: کمتر از یا مساوی ۱۲۰.

۷. Electrocardiographic (نتایج الکتروکاردیوگرام)

تغییرات غیرطبیعی در ECG می‌تواند نشانه‌ای از مشکلات قلبی مانند ایسکمی یا هیپرتروفی بطن باشد. این فیلد به سه دسته (عادی، تغییرات ST-T، و هیپرتروفی) تقسیم شده و از طریق تحلیل نتایج ECG اندازه‌گیری می‌شود.

- مقدار ۰: عادی.
- مقدار ۱: وجود تغییرات ST-T (مانند الگوهای غیرطبیعی موج T).
- مقدار ۲: نشان‌دهنده هیپرتروفی بطن چپ احتمالی یا دیگر مشکلات.

۸. Heart Rate (بیشترین ضربان قلب)

ضربان قلب بالا می‌تواند نشان‌دهنده استرس قلبی باشد و در شرایطی خاص مانند تست ورزش اطلاعات ارزشمندی ارائه می‌دهد. این فیلد به‌صورت ضربان در دقیقه (bpm) ثبت شده است.

۹. Exercise Induced (درد ناشی از ورزش)

بروز درد قفسه سینه هنگام ورزش معمولاً نشان‌دهنده کاهش خون‌رسانی به قلب در زمان افزایش نیاز اکسیژن است. این فیلد به‌صورت داده باینری (۱: بله، ۰: خیر) ثبت می‌شود و در معاینات تست ورزش بررسی می‌گردد.

- مقدار ۱: درد ناشی از ورزش.
- مقدار ۰: بدون درد مرتبط با ورزش.

۱۰. Depression (شدت افسردگی)

افسردگی شدید می‌تواند خطر ابتلا به بیماری‌های قلبی را به طور قابل توجهی افزایش دهد .مطالعات نشان داده‌اند که افراد مبتلا به افسردگی شدید، بیشتر در معرض خطر ابتلا به بیماری‌های قلبی مانند حمله قلبی و سکته مغزی هستند.

۱۱. Slope (شیب قطعه ST)

شیب قطعه ST هنگام ورزش می‌تواند وضعیت خون‌رسانی به قلب را نشان دهد. شیب صعودی معمولاً عادی، شیب مسطح یا نزولی اغلب غیرطبیعی و مرتبط با مشکلات قلبی است. این متغیر به‌صورت طبقه‌بندی عددی ثبت می‌شود.

- مقدار ۱: شیب نزولی.
- مقدار ۲: شیب مسطح.
- مقدار ۳: شیب صعودی.

۱۲. CA (تعداد عروق کرونری رنگ‌شده)

تعداد عروق اصلی که با رنگ فلوروسکوپی دیده شده‌اند، میزان گرفتگی عروق کرونری را نشان می‌دهد. افزایش تعداد عروق رنگ‌شده نشان‌دهنده (از ۰ تا ۴) شدت بیماری است. این داده از طریق آنژیوگرافی به دست می‌آید.

۱۳. Thal (تالاسمی)

این فیلد وضعیت خون‌رسانی قلبی را بر اساس تست‌های رادیوایزوتوپ نشان می‌دهد. تالاسمی عادی، نقص برگشت‌پذیر، یا نقص دائمی (ثابت) اطلاعاتی از جریان خون غیرطبیعی ارائه می‌دهد.

- مقدار ۳: عادی.
- مقدار ۶: نقص ثابت.
- مقدار ۷: نقص برگشت‌پذیر.

۱۴. C (متغیر هدف)

این فیلد نشان‌دهنده وجود یا عدم وجود سکته قلبی است (۰: سالم، ۱: مبتلا به سکته قلبی). این متغیر نتیجه نهایی تشخیص مبتنی بر داده‌های آزمایشگاهی و بالینی است.

نوع ویژگی ها

ویژگی‌های یک مجموعه داده معمولاً به دو نوع اصلی تقسیم می‌شوند:

- کمی (Numerical)
- کیفی (Categorical).

تشخیص و تفکیک صحیح ویژگی‌ها (کمی، کیفی، پیوسته، گسسته) تأثیر مستقیمی بر کیفیت پیش‌پردازش داده و دقت مدل‌های یادگیری ماشین دارد. درک صحیح نوع داده‌ها به انتخاب روش‌های مناسب برای پاکسازی، تحلیل و مدل‌سازی کمک می‌کند.

نام	نوع داده	حالات
Age (سن)	کمی (Continuous)	-
Sex (جنسیت)	کیفی (Categorical, Binary)	0:زن, ۱: مرد
Chest Pain (درد قفسه سینه)	کیفی (Categorical)	1: Typical Angina, 2: Atypical Angina, 3: Non-Anginal Pain, 4 : بدون درد
Blood Pressure (فشار خون)	کمی (Continuous)	-
Cholesterol (کلسترول)	کمی (Continuous)	-
Blood Sugar (قند خون ناشتا)	کیفی (Categorical, Binary)	0: ≤120 mg/dL, 1: >120 mg/dL
Electrocardiographic (نتایج الکتروکاردیوگرام)	کیفی (Categorical)	۰: عادی, ۱: وجود تغییرات ST-T ۲: هیپرتروفی بطن چپ
Heart Rate (بیشترین ضربان قلب)	کمی (Continuous)	-
Exercise Induced (درد ناشی از ورزش)	کیفی (Categorical, Binary)	۰: بدون درد, ۱: درد ناشی از ورزش
Depression (شدت افسردگی)	کمی (Continuous)	-
Slope (شیب قطعهST)	کیفی (Categorical)	۱: شیب نزولی, ۲: شیب مسطح, ۳: شیب صعودی
CA (تعداد عروق کرونری رنگ‌شده)	کیفی (Categorical)	۰-۴
Thal (تالاسمی)	کیفی (Categorical)	۳: عادی, ۶: نقص ثابت, ۷: نقص برگشت‌پذیر
C (متغیر هدف)	کیفی (Categorical, Binary)	۰: سالم, ۱: سکته قلبی

تحلیل آماری ویژگی ها

تحلیل آماری ویژگی‌ها یکی از مراحل حیاتی در فرآیند تحلیل داده است. این مرحله به ما کمک می‌کند تا داده‌های خام و بی‌ساختار را به اطلاعات مفید و قابل تفسیر تبدیل کنیم. با استفاده از روش‌های آماری، می‌توانیم الگوها، روابط و روندهای پنهان در داده‌ها را کشف کرده و به درک عمیق‌تری از پدیده مورد مطالعه دست پیدا کنیم.

با توجه به اهمیت تفیکیک ویژگی ها در فرآیند پاکسازی، آماده سازی و مدلسازی، در ابتدا ویژگی ها با توجه به مقادیرشان به دو دسته Numerical , Categorical تقسیم شدند. به طور معمول با استفاده از نوع داده‌ی هر ویژگی، مطابق کد زیر تفکیک صورت میپذیرد.

```
"Numerical Features":        df.select_dtypes(include=["float64",
"int64"]).columns.tolist(),
"Categorical Features":         df.select_dtypes(include=["object",
"category"]).columns.tolist(),
```

از آنجا که در مجموعه داده مقادیر Int, Float درج شده و نام حالات برای هر ویژگی ثبت نشده بود، تابع detect_categorical_columns پیاده سازی شد. در این تابع براساس تنوع مقادیر هر ویژگی تشخیص داده میشود که آیا ویژگی Categorical است یا خیر.

```
def detect_categorical_columns(df, threshold=10):
  #Detecting categorical columns.
  categorical_cols = []
  for col in df.columns:
    unique_values = df[col].nunique()
    if unique_values <= threshold:
      categorical_cols.append(col)
  return categorical_cols
```

در ادامه تحلیل آماری برای هر ویژگی ارائه میشود و برای تحلیل هر ویژگی، در صورت عدد بودن دو نمودار هیستوگرام و باکس‌پلات و در صورت دسته بندی نمودار فراوانی و همچنین آمار توصیفی ارائه میشود. این ابزارها به ما کمک می‌کنند تا توزیع داده‌ها، وجود مقادیر پرت، تمرکز داده‌ها و سایر ویژگی‌های مهم این متغیر را بررسی کنیم.

۱. Age (سن)

تحلیل هیستوگرام:

- توزیع تقریباً نرمال: هیستوگرام نشان می‌دهد که توزیع سن تقریباً به شکل زنگوله‌ای (نرمال) است. این بدان معناست که اکثر افراد در محدوده سنی خاصی متمرکز شده‌اند و تعداد افراد با سنین بسیار پایین یا بسیار بالا کمتر است.
- میانگین و انحراف استاندارد: میانگین سن حدود ۵۱ سال و انحراف استاندارد آن حدود ۹ سال است. این نشان می‌دهد که اکثر افراد بین ۴۲ تا ۶۰ سال سن دارند.

- تقارن: توزیع سن تقریباً متقارن است، به این معنی که تعداد افرادی که سن کمتری از میانگین دارند تقریباً برابر با تعداد افرادی است که سن بیشتری از میانگین دارند.

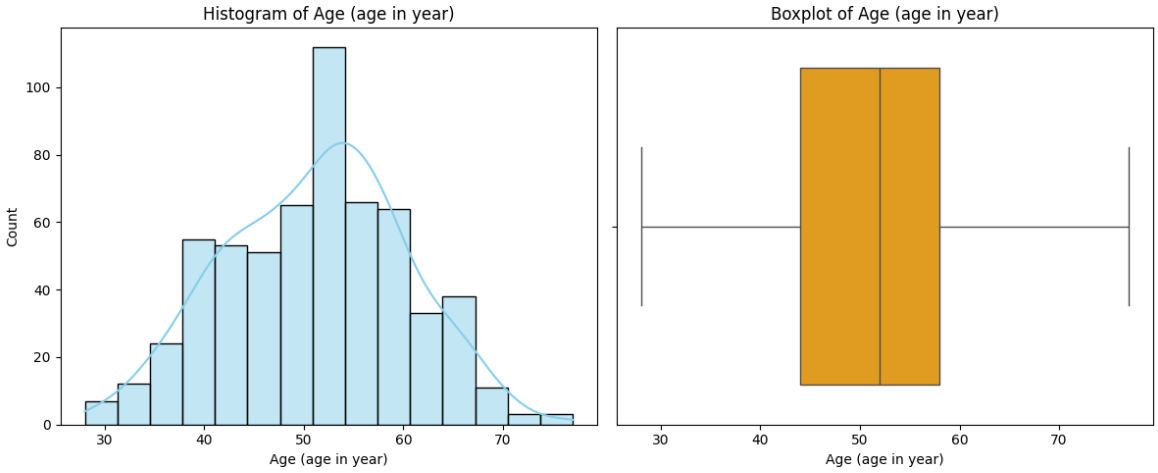
تحلیل باکس‌پلات:

- محدوده میان‌چرکی (IQR):IQR برابر با ۱۴ سال است که نشان می‌دهد ۵۰ درصد داده‌ها در محدوده ۱۴ ساله متمرکز شده‌اند.
- چارک اول و سوم: چارک اول (Q1) برابر با ۴۴ سال و چارک سوم (Q3) برابر با ۵۸ سال است. این نشان می‌دهد که ۲۵ درصد افراد کمتر از ۴۴ سال و ۲۵ درصد افراد بیشتر از ۵۸ سال سن دارند.
- رنج: محدوده تغییرات سن از ۲۸ تا ۷۷ سال است.
- مقادیر پرت: با توجه به باکس‌پلات، به نظر نمی‌رسد مقادیر پرت قابل توجهی در داده‌ها وجود داشته باشد.

بر اساس تحلیل‌های انجام شده، می‌توانیم به نتایج زیر دست پیدا کنیم:

- توزیع سنی: اکثر افراد در محدوده سنی ۴۲ تا ۶۰ سال قرار دارند و توزیع سن تقریباً نرمال است.
- پراکندگی داده‌ها: داده‌ها پراکندگی مناسبی دارند و مقادیر پرت قابل توجهی مشاهده نمی‌شود.
- میانگین سن: میانگین سن افراد حدود ۵۱ سال است.

```
========================================
Analyzing Field:      Age (age in year)
========================================
Numerical Field
count   597.000000
mean     51.182580
std       9.074366
min      28.000000
25%      44.000000
50%      52.000000
75%      58.000000
max      77.000000
Name:      Age (age in year), dtype:      float64
```



۲. Sex (جنسیت)

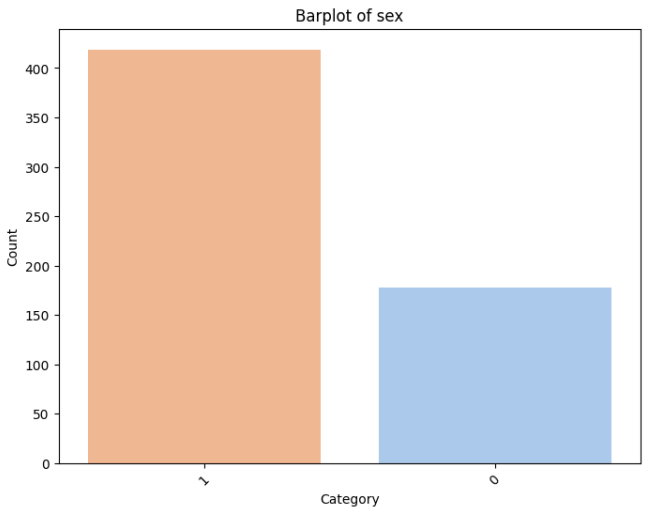
تحلیل نمودار فراوانی:

- دو دسته جنسیتی: نمودار نشان می‌دهد که ویژگی جنسیت به دو دسته اصلی تقسیم می‌شود که احتمالا به جنسیت‌های مذکر و مؤنث اشاره دارد.
- تفاوت تعداد: به وضوح مشاهده می‌شود که تعداد افراد در یکی از دسته‌ها (مذکر) به طور قابل توجهی بیشتر از دسته دیگر (مؤنث) است.
- توزیع نامتوازن: توزیع فراوانی بین دو دسته جنسیتی نامتوازن است. این نامتوازنی می‌تواند در تحلیل‌های بعدی و مدل‌سازی‌ها تأثیرگذار باشد و نیاز به اعمال برخی روش‌ها برای برطرف کردن آن (مانند نمونه‌برداری تصادفی یا روش‌های وزن‌دهی) وجود داشته باشد.

تحلیل آمار توصیفی:

- تعداد کل: تعداد کل مشاهدات در این ویژگی برابر با ۵۹۷ است.
- فراوانی هر دسته تعداد افراد در دسته اول ۴۱۹ نفر و در دسته دوم ۱۷۸ نفر است.
- درصد هر دسته: با محاسبه درصد هر دسته می‌توانیم به درک بهتری از نسبت هر یک از جنسیت‌ها در داده‌ها برسیم. برای مثال، درصد افراد در دسته اول حدود ۷۰٪ و در دسته دوم حدود ۳۰٪ است.

```
========================================
Analyzing Field:      sex
========================================
Categorical Field
sex
1   419
0   178
Name:      count, dtype:      int64
```



۳. Chest Pain (درد قفسه سینه)

تحلیل نمودار فراوانی:

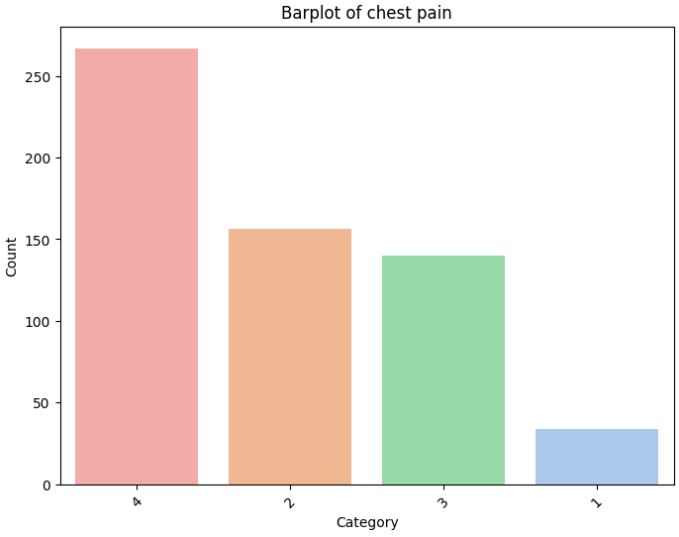
- چهار دسته درد: نمودار به وضوح چهار دسته مختلف از درد قفسه سینه را نشان می‌دهد که نشان‌دهنده شدت‌ها و انواع مختلف درد هستند.
- دسته غالب: دسته ۴ بیشترین فراوانی را دارد، به این معنی که بخش قابل توجهی از افراد در مجموعه داده، این نوع درد را تجربه کرده‌اند.
- کاهش تدریجی فراوانی: با حرکت از دسته ۴ به دسته ۱، فراوانی هر دسته کاهش می‌یابد. این نشان می‌دهد که انواع کمتر شایع درد قفسه سینه، شیوع کمتری در مجموعه داده دارند.

تحلیل آمار توصیفی:

- کل مشاهدات: مجموعه داده شامل ۵۹۷ مشاهده است.
- توزیع دسته‌ها:
  - دسته ۴: ۲۶۷ مشاهده (تقریباً ۴۵٪)
  - دسته ۲: ۱۵۶ مشاهده (تقریباً ۲۶٪)
  - دسته ۳: ۱۴۰ مشاهده (تقریباً ۲۳٪)
  - دسته ۱: ۳۴ مشاهده (تقریباً ۶٪)
- مقادیر گمشده: اگر مقادیر گمشده‌ای در این ویژگی وجود دارد، باید به درستی با آن‌ها برخورد شود (مثلاً با پر کردن آن‌ها یا حذف مشاهدات مربوطه).
- عدم تعادل داده‌ها: تفاوت قابل توجه در فراوانی دسته ۴ نسبت به سایر دسته‌ها ممکن است نیاز به تکنیک‌هایی مانند افزایش نمونه یا کاهش نمونه برای متعادل کردن مجموعه داده در مدل‌سازی داشته باشد.

تحلیل ویژگی درد قفسه سینه نشان می‌دهد که این ویژگی نقش مهمی در درک و پیش‌بینی بیماری‌های قلبی دارد. با بررسی توزیع انواع مختلف درد و ارتباط آن با سایر متغیرها، می‌توان به بینش‌های ارزشمندی دست یافت که در تشخیص و درمان بیماری‌های قلبی مفید خواهد بود.

```
========================================
Analyzing Field:        chest pain
========================================
Categorical Field
chest pain
4    267
2    156
3    140
1     34
Name:        count, dtype:        int64
```



۴. Blood Pressure (فشار خون)

تحلیل آمار توصیفی

- تعداد مشاهدات 596 فرد در این مطالعه بررسی شده‌اند.
- میانگین: میانگین فشار خون حدود ۱۳۲.۱۳ میلی‌متر جیوه است که نشان می‌دهد به‌طور متوسط، افراد در این گروه دارای فشار خون کمی بالاتر از حد نرمال هستند.
- انحراف استاندارد: انحراف استاندارد ۱۷.۶ نشان می‌دهد که مقادیر فشار خون در این گروه پراکندگی قابل توجهی دارند و برخی افراد فشار خون بسیار بالاتر یا پایین‌تر از میانگین دارند.
- حداقل و حداکثر: کمترین فشار خون اندازه‌گیری شده ۹۲ و بیشترین آن ۲۰۰ میلی‌متر جیوه بوده است. این نشان می‌دهد که محدوده تغییرات فشار خون در این گروه گسترده است.
- چارک‌ها:
  - 25درصد افراد فشار خونی کمتر از ۱۲۰ میلی‌متر جیوه دارند.
  - 50درصد افراد (میانه) فشار خونی کمتر از ۱۳۰ میلی‌متر جیوه دارند.
  - 75درصد افراد فشار خونی کمتر از ۱۴۰ میلی‌متر جیوه دارند.

تحلیل نمودار هیستوگرام:

- توزیع تقریباً نرمال: هیستوگرام نشان می‌دهد که توزیع فشار خون تقریباً به شکل زنگوله‌ای (نرمال) است، به این معنی که بیشترین افراد فشار خونی در محدوده میانگین دارند و به سمت مقادیر کمتر یا بیشتر، تعداد افراد کاهش می‌یابد.
- پراکندگی: وجود پراکندگی در داده‌ها به وضوح مشخص است، زیرا برخی از افراد فشار خون بسیار پایین‌تر یا بالاتر از میانگین دارند.
- مقادیر پرت: ممکن است چندین مقدار پرت در سمت راست نمودار (فشار خون بالا) وجود داشته باشد که نیاز به بررسی بیشتر دارد.

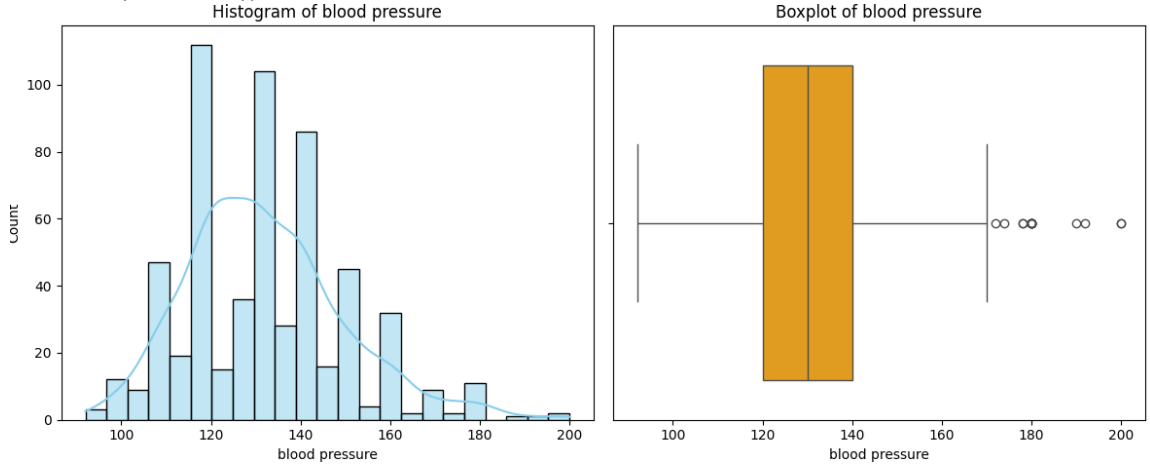
تحلیل نمودار باکس‌پلات:

- محدوده بین‌چرکی IQR :(IQR) نشان می‌دهد که ۵۰ درصد داده‌ها در چه محدوده‌ای قرار دارند. در این نمودار، IQRنسبتاً بزرگ است که نشان‌دهنده پراکندگی زیاد داده‌ها است.
- مقادیر پرت: در این نمودار، چندین مقدار پرت در سمت راست وجود دارد که نشان‌دهنده فشار خون بسیار بالا در برخی افراد است.
- تقارن: باکس‌پلات نشان می‌دهد که توزیع داده‌ها کمی به سمت راست متمایل است، به این معنی که تعداد بیشتری از افراد فشار خون بالاتر از میانگین دارند.

بر اساس تحلیل آمار توصیفی و نمودارها، می‌توان نتیجه گرفت که:

- فشار خون در این گروه از افراد پراکندگی قابل توجهی دارد و تعداد قابل توجهی از افراد فشار خون بالاتر از حد نرمال دارند.
- وجود مقادیر پرت نشان می‌دهد که ممکن است برخی از افراد مبتلا به بیماری‌های مرتبط با فشار خون بالا باشند. در پروژه واقعی باید بررسی شود که آیا مقادیر پرت به دلیل خطا در اندازه‌گیری یا عوامل دیگری مانند بیماری‌های همراه ایجاد شده‌اند.
- با بررسی همبستگی بین فشار خون و سایر متغیرهای مرتبط مانند سن، جنسیت، وزن، کلسترول و ... می توان به نتایج معناداری رسید.

```
========================================
Analyzing Field:       blood pressure
========================================
Numerical Field
count   596.000000
mean    132.129195
std      17.603812
min      92.000000
25%     120.000000
50%     130.000000
75%     140.000000
 max    200.000000
Name:       blood pressure, dtype:       float64
```



۵. Cholesterol (کلسترول)

تحلیل آمار توصیفی

- تعداد مشاهدات: 574 نفر در این مطالعه بررسی شده‌اند.
- میانگین: میانگین سطح کلسترول حدود ۲۴۸.۶۶ میلی‌گرم بر دسی‌لیتر است که نشان می‌دهد به‌طور متوسط، افراد دارای سطح کلسترول نسبتاً بالایی هستند.
- انحراف استاندارد: انحراف استاندارد ۵۹.۷۸ نشان می‌دهد، مقادیر کلسترول در این مجموعه داده پراکندگی قابل توجهی داشته و برخی افراد سطح کلسترول بسیار بالاتر یا پایین‌تر از میانگین دارند.
- حداقل و حداکثر: کمترین سطح کلسترول اندازه‌گیری شده ۸۵ و بیشترین آن ۶۰۳ میلی‌گرم بر دسی‌لیتر است. این نشان می‌دهد که محدوده تغییرات سطح کلسترول در این گروه گسترده است.
- چارک‌ها:
  - 25درصد افراد سطح کلسترول کمتر از ۲۱۱ میلی‌گرم بر دسی‌لیتر دارند.
  - 50درصد افراد (میانه) سطح کلسترول کمتر از ۲۴۲.۵ میلی‌گرم بر دسی‌لیتر دارند.
  - 75درصد افراد سطح کلسترول کمتر از ۲۷۸.۷۵ میلی‌گرم بر دسی‌لیتر دارند.

تحلیل نمودار هیستوگرام:

- توزیع تقریباً نرمال: هیستوگرام نشان می‌دهد که توزیع سطح کلسترول تقریباً به شکل زنگوله‌ای (نرمال) است، به این معنی که بیشترین افراد سطح کلسترول در محدوده میانگین دارند و به سمت مقادیر کمتر یا بیشتر، تعداد افراد کاهش می‌یابد.
- پراکندگی: وجود پراکندگی در داده‌ها به وضوح مشخص است، زیرا برخی از افراد سطح کلسترول بسیار پایین‌تر یا بالاتر از میانگین دارند.
- مقادیر پرت: ممکن است چندین مقدار پرت در سمت راست نمودار (سطح کلسترول بالا) وجود داشته باشد که نیاز به بررسی بیشتر دارد.

باکس‌پلات:

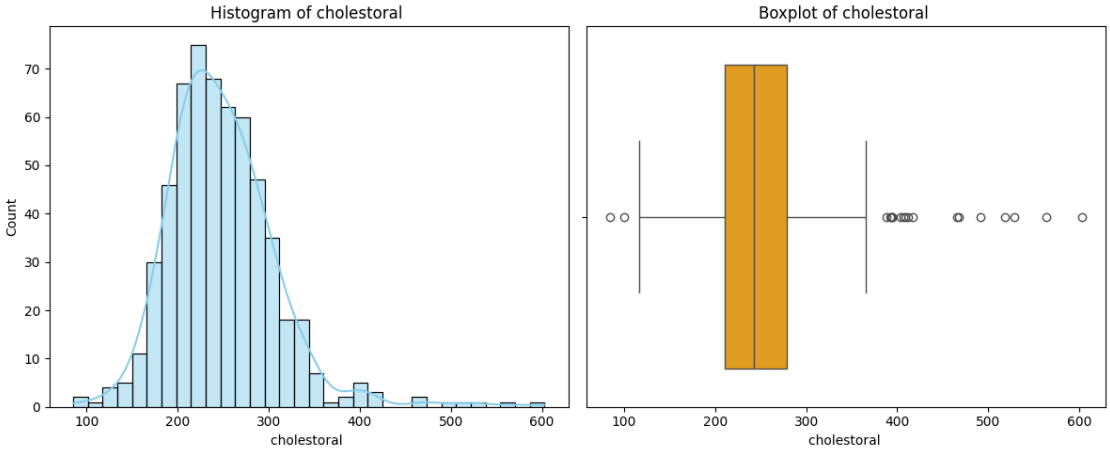
- محدوده بین‌چرکی IQR :(IQR) نشان می‌دهد که ۵۰ درصد داده‌ها در چه محدوده‌ای قرار دارند. در این نمودار، IQRنسبتاً بزرگ است که نشان‌دهنده پراکندگی زیاد داده‌ها است.
- مقادیر پرت: در این نمودار، چندین مقدار پرت در سمت راست وجود دارد که نشان‌دهنده سطح کلسترول بسیار بالا در برخی افراد است.
- تقارن: باکس‌پلات نشان می‌دهد که توزیع داده‌ها کمی به سمت راست متمایل است، به این معنی که تعداد بیشتری از افراد سطح کلسترول بالاتر از میانگین دارند.

بر اساس تحلیل آمار توصیفی و نمودارها، می‌توان نتیجه گرفت که:

- سطح کلسترول در این گروه از افراد پراکندگی قابل توجهی دارد.
- تعداد قابل توجهی از افراد سطح کلسترول بالاتر از حد نرمال دارند.

- وجود مقادیر پرت نشان می‌دهد که ممکن است برخی از افراد مبتلا به بیماری‌های مرتبط با کلسترول بالا باشند.
- با بررسی همبستگی بین کلسترول و سایر متغیرهای مانند سن، جنسیت، وزن، فشار خون و ... میتوان به نتایج معنا داری رسید.

```
========================================
Analyzing Field:      cholestoral
========================================
Numerical Field
count   574.000000
mean    248.655052
std      59.784805
min      85.000000
25%     211.000000
50%     242.500000
75%     278.750000
max     603.000000
Name:      cholestoral , dtype:      float64
```



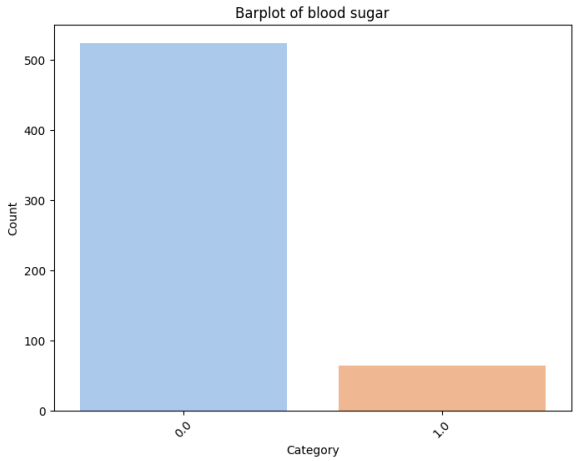
۶. Blood Sugar (قند خون ناشتا)

- تعداد مشاهدات: 589 نفر در این مطالعه بررسی شده‌اند.
- دسته‌ها: قند خون به دو دسته تقسیم شده است:
  - دسته ۰.۰: ۵۲۴ نفر (حدود ۸۹٪)
  - دسته ۱.۰: ۶۵ نفر (حدود ۱۱٪)

تحلیل نمودار نمودار میله‌ای:

- توزیع نامتقارن: نمودار میله‌ای نشان می‌دهد که توزیع قند خون به شدت نامتقارن است. اکثریت قریب به اتفاق افراد (حدود ۸۹٪) در دسته ۰ قرار دارند که احتمالاً نشان‌دهنده سطح قند خون نرمال است. تنها تعداد کمی از افراد (حدود ۱۱٪) در دسته ۱ قرار دارند که احتمالاً نشان‌دهنده سطح قند خون بالا یا دیابت است.
- تفاوت بین فراوانی دو دسته بسیار زیاد است و نشان‌دهنده تفاوت معنی‌داری در سطح قند خون افراد است.
- با بررسی همبستگی بین قندخون و سایر متغیرهای مانند سن، جنسیت، وزن، فشار خون و ... میتوان به نتایج معنا داری رسید.

==========
Analyzing Field: blood sugar
==========
Categorical Field
blood sugar
0.0 524
1.0 65
Name: count, dtype: int64



۷. Electrocardiographic (نتایج الکتروکاردیوگرام)

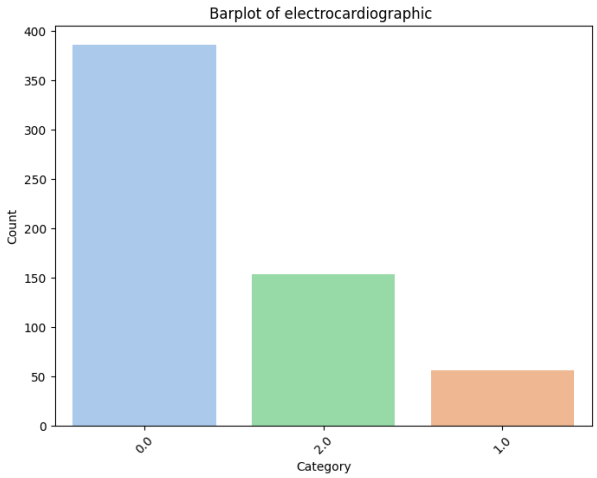
تحلیل آمار توصیفی

- تعداد مشاهدات: 596 نفر در این مطالعه بررسی شده‌اند.
- دسته‌ها: نتایج ECG به سه دسته تقسیم شده است:
  - دسته ۰.۰: ۳۸۶ نفر (حدود ۶۵٪)
  - دسته ۲.۰: ۱۵۴ نفر (حدود ۲۶٪)
  - دسته ۱.۰: ۵۶ نفر (حدود ۹٪)

تحلیل نمودار میله‌ای:

- توزیع نامتقارن: نمودار میله‌ای نشان می‌دهد که توزیع نتایج ECG به شدت نامتقارن است. اکثریت قریب به اتفاق افراد (حدود ۶۵٪) در دسته ۰ قرار دارند.
- تفاوت معنی‌دار: تفاوت بین فراوانی سه دسته بسیار زیاد است و نشان‌دهنده تفاوت معنی‌داری در نتایج ECG افراد است.

==========
Analyzing Field: electrocardiographic
==========
Categorical Field
electrocardiographic
0.0 386
2.0 154
1.0 56
Name: count, dtype: int64



۸. Heart Rate (ضربان قلب)

تحلیل آمار توصیفی

- تعداد مشاهدات: 596 نفر در این مطالعه بررسی شده‌اند.
- میانگین: میانگین ضربان قلب حدود ۱۴۴.۴۶ ضربه در دقیقه است که نشان می‌دهد به‌طور متوسط، افراد در این گروه دارای ضربان قلب نسبتاً بالایی هستند.
- انحراف استاندارد: انحراف استاندارد ۲۳.۷۹ نشان می‌دهد که مقادیر ضربان قلب در این گروه پراکندگی قابل توجهی دارند و برخی افراد ضربان قلب بسیار بالاتر یا پایین‌تر از میانگین دارند.
- حداقل و حداکثر: کمترین ضربان قلب اندازه‌گیری شده ۷۱ و بیشترین آن ۲۰۲ ضربه در دقیقه بوده است. این نشان می‌دهد که محدوده تغییرات ضربان قلب در این گروه گسترده است.
- چارک‌ها:
  - 25درصد افراد ضربان قلب کمتر از ۱۲۸ ضربه در دقیقه دارند.
  - 50درصد افراد (میانه) ضربان قلب کمتر از ۱۴۶ ضربه در دقیقه دارند.
  - 75درصد افراد ضربان قلب کمتر از ۱۶۲ ضربه در دقیقه دارند.

تحلیل نمودار هیستوگرام:

- توزیع تقریباً نرمال: هیستوگرام نشان می‌دهد که توزیع ضربان قلب تقریباً به شکل زنگوله‌ای (نرمال) است، به این معنی که بیشترین افراد ضربان قلب در محدوده میانگین دارند و به سمت مقادیر کمتر یا بیشتر، تعداد افراد کاهش می‌یابد.
- پراکندگی: وجود پراکندگی در داده‌ها به وضوح مشخص است، زیرا برخی از افراد ضربان قلب بسیار پایین‌تر یا بالاتر از میانگین دارند.
- مقادیر پرت: ممکن است چندین مقدار پرت در سمت راست نمودار (ضربان قلب بالا) وجود داشته باشد که نیاز به بررسی بیشتر دارد.

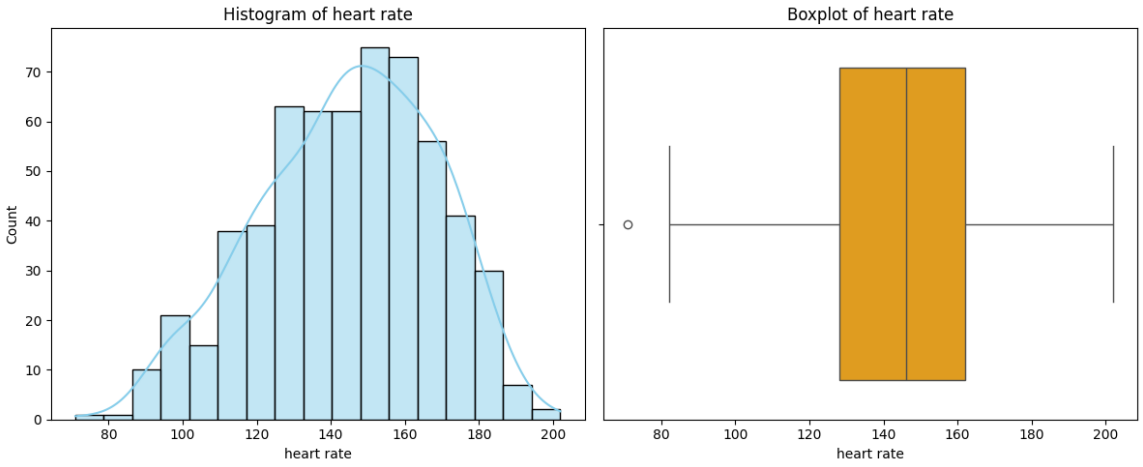
باکس‌پلات:

- محدوده بین‌چرکی IQR :(IQR) نشان می‌دهد که ۵۰ درصد داده‌ها در چه محدوده‌ای قرار دارند. در این نمودار، IQRنسبتاً بزرگ است که نشان‌دهنده پراکندگی زیاد داده‌ها است.
- مقادیر پرت: در این نمودار، چندین مقدار پرت در سمت راست وجود دارد که نشان‌دهنده ضربان قلب بسیار بالا در برخی افراد است.
- تقارن: باکس‌پلات نشان می‌دهد که توزیع داده‌ها کمی به سمت راست متمایل است، به این معنی که تعداد بیشتری از افراد ضربان قلب بالاتر از میانگین دارند.

بر اساس تحلیل آمار توصیفی و نمودارها، می‌توان نتیجه گرفت که:

- ضربان قلب در این گروه از افراد پراکندگی قابل توجهی دارد.
- تعداد قابل توجهی از افراد ضربان قلب بالاتر از حد نرمال دارند.
- وجود مقادیر پرت نشان می‌دهد که ممکن است برخی از افراد مبتلا به بیماری‌های قلبی یا شرایطی باشند که باعث افزایش ضربان قلب می‌شود.

```
========================================
Analyzing Field:       heart rate
========================================
Numerical Field
count    596.000000
mean     144.456376
std       23.794282
min       71.000000
25%      128.000000
50%      146.000000
75%      162.000000
max      202.000000
Name:       heart rate, dtype:       float64
```



۹. Exercise Induced (درد ناشی از ورزش)

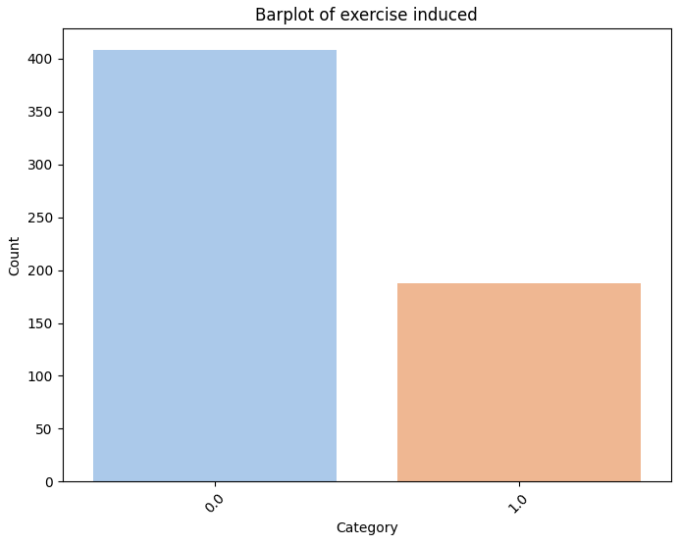
تحلیل آمار توصیفی

- تعداد مشاهدات: 596 نفر در این مطالعه بررسی شده‌اند.
- دسته‌ها: وضعیت ورزش القایی به دو دسته تقسیم شده است:
  - دسته ۰.۰: ۴۰۸ نفر (حدود ۶۸٪)
  - دسته ۱.۰: ۱۸۸ نفر (حدود ۳۲٪)

تحلیل نمودار میله‌ای:

- توزیع نامتقارن: نمودار میله‌ای نشان می‌دهد که توزیع وضعیت این ویژگی به شدت نامتقارن است. اکثریت قریب به اتفاق افراد (حدود ۶۸٪) در دسته ۰.۰ قرار دارند.
- تفاوت بین فراوانی دو دسته بسیار زیاد است و نشان‌دهنده تفاوت معنی‌داری در وضعیت افراد است.

```
========================================
Analyzing Field:      exercise induced
========================================
Categorical Field
exercise induced
0.0   408
1.0   188
Name:      count, dtype:      int64
```



۱۰. Depression (شدت افسرگی)

تحلیل آمار توصیفی

- تعداد مشاهدات: 597نفر در این مطالعه بررسی شده‌اند.
- میانگین: میانگین شدت افسردگی حدود ۰.۸۱ است که نشان می‌دهد به‌طور متوسط، افراد در این گروه دارای سطح نسبتاً پایینی از افسردگی هستند.

- انحراف استاندارد: انحراف استاندارد ۱.۰۶ نشان می‌دهد که مقادیر شدت افسردگی در این گروه پراکندگی قابل توجهی دارند و برخی افراد سطح افسردگی بسیار بالاتر یا پایین‌تر از میانگین دارند.
- حداقل و حداکثر: کمترین شدت افسردگی اندازه‌گیری شده ۰ (یعنی عدم وجود افسردگی) و بیشترین آن ۶.۲ بوده است. این نشان می‌دهد که محدوده تغییرات شدت افسردگی در این گروه گسترده است.
- چارک‌ها:
  - 25درصد افراد شدت افسردگی کمتر از ۰ (یعنی عدم وجود افسردگی) دارند.
  - 50درصد افراد (میانه) شدت افسردگی کمتر از ۰.۲ دارند.
  - 75درصد افراد شدت افسردگی کمتر از ۱.۵ دارند.

تحلیل نمودارها

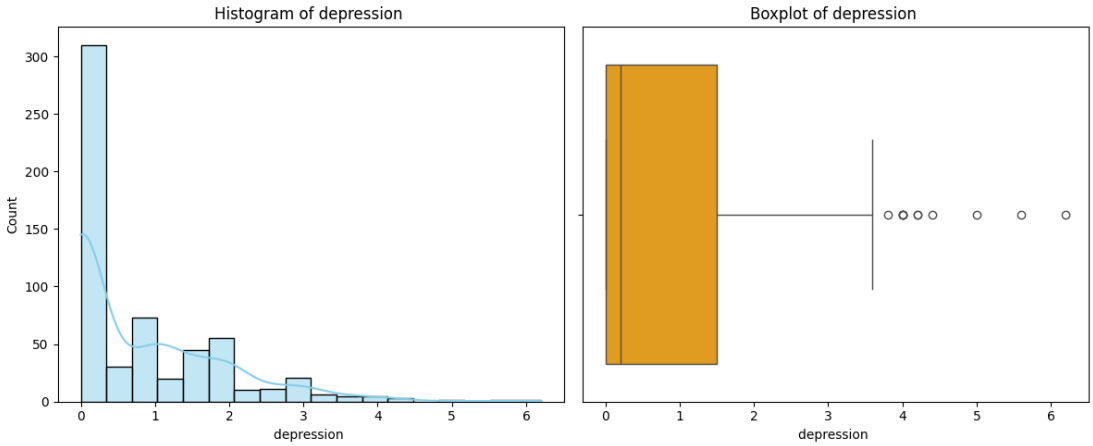
هیستوگرام:

- توزیع نامتقارن به سمت راست: هیستوگرام نشان می‌دهد که توزیع شدت افسردگی به شدت به سمت راست نامتقارن است. این بدان معنی است که اکثریت افراد سطح افسردگی پایینی دارند و تعداد کمی از افراد سطح افسردگی بسیار بالا دارند.
- پراکندگی: وجود پراکندگی در داده‌ها به وضوح مشخص است، زیرا برخی از افراد سطح افسردگی بسیار بالاتر از میانگین دارند.
- مقادیر پرت: ممکن است چندین مقدار پرت در سمت راست نمودار (سطح افسردگی بالا) وجود داشته باشد که نیاز به بررسی بیشتر دارد.

باکس‌پلات:

- محدوده بین‌چرکی IQR :(IQR) نشان می‌دهد که ۵۰ درصد داده‌ها در چه محدوده‌ای قرار دارند. در این نمودار، IQRنسبتاً بزرگ است که نشان‌دهنده پراکندگی زیاد داده‌ها است.
- مقادیر پرت: در این نمودار، چندین مقدار پرت در سمت راست وجود دارد که نشان‌دهنده سطح افسردگی بسیار بالا در برخی افراد است.
- تقارن: باکس‌پلات به وضوح نشان می‌دهد که توزیع داده‌ها به شدت به سمت راست متمایل است.

```
========================================
Analyzing Field:      depression
========================================
Numerical Field
count   597.000000
mean      0.816248
std       1.067938
min       0.000000
25%       0.000000
50%       0.200000
75%       1.500000
max       6.200000
Name:      depression , dtype:      float64
```



۱۱. Slope (شیب قطعه ST)

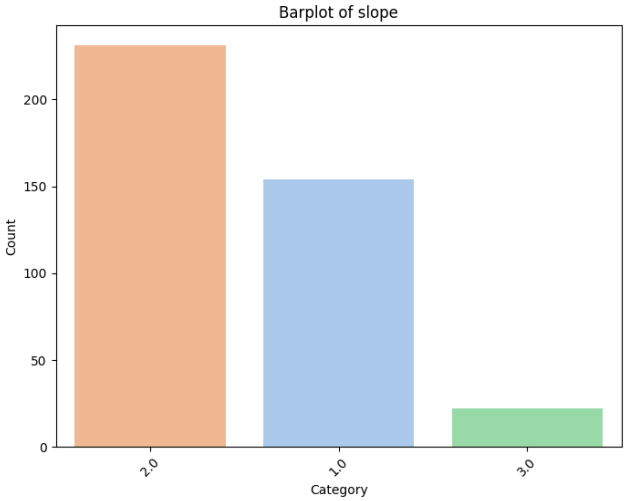
تحلیل آمار توصیفی

- تعداد مشاهدات: 407 مورد در این مطالعه بررسی شده‌اند.
- دسته‌ها: مقادیر شیب به سه دسته تقسیم شده است:
  - دسته ۲.۰: ۲۳۱ مورد (حدود ۵۶.۷٪)
  - دسته ۱.۰: ۱۵۴ مورد (حدود ۳۷.۸٪)
  - دسته ۳.۰: ۲۲ مورد (حدود ۵.۴٪)

تحلیل نمودار میله‌ای:

- توزیع نامتقارن: نمودار میله‌ای نشان می‌دهد که توزیع مقادیر شیب به شدت نامتقارن است. اکثریت قریب به اتفاق موارد (حدود ۵۶.۷٪) در دسته ۲.۰ قرار دارند.
- تفاوت بین فراوانی سه دسته بسیار زیاد است و نشان‌دهنده تفاوت معنی‌داری در مقادیر شیب موارد است.

```
========================================
Analyzing Field:      slope
========================================
Categorical Field
slope
2.0   231
1.0   154
3.0    22
Name:      count, dtype:      int64
```



CA (تعداد عروق کرونری رنگ‌شده)

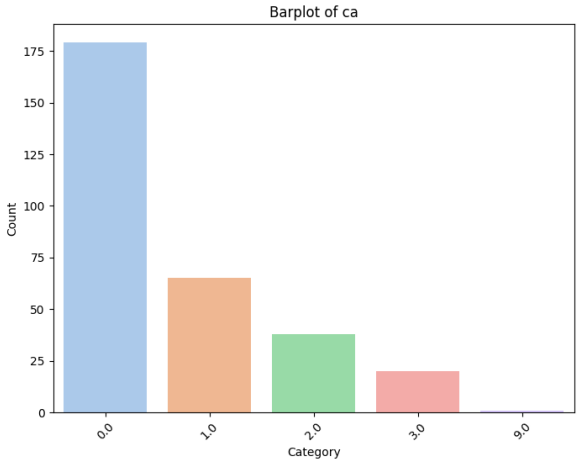
تحلیل آمار توصیفی

- تعداد مشاهدات: 303 مورد در این مطالعه بررسی شده‌اند.
- دسته‌ها: مقادیر ca به پنج دسته تقسیم شده است:
  - دسته ۰.۰: ۱۷۹ مورد (حدود ۵۹٪)
  - دسته ۱.۰: ۶۵ مورد (حدود ۲۱.۴٪)
  - دسته ۲.۰: ۳۸ مورد (حدود ۱۲.۵٪)
  - دسته ۳.۰: ۲۰ مورد (حدود ۶.۶٪)
  - دسته ۹.۰: ۱ مورد (حدود ۰.۳٪)

تحلیل نمودار میله‌ای:

- توزیع نامتقارن: نمودار میله‌ای نشان می‌دهد که توزیع مقادیر ca به شدت نامتقارن است. اکثریت قریب به اتفاق موارد (حدود ۵۹٪) در دسته ۰.۰ قرار دارند.
- تفاوت بین فراوانی دسته‌ها بسیار زیاد است و نشان‌دهنده تفاوت معنی‌داری در مقادیر ca موارد است.

```
========================================
Analyzing Field:      ca
========================================
Categorical Field
ca
0.0   179
1.0    65
2.0    38
3.0    20
9.0     1
Name:      count, dtype:      int64
```



۱۲. Thal (تالاسمی)

تحلیل آمار توصیفی

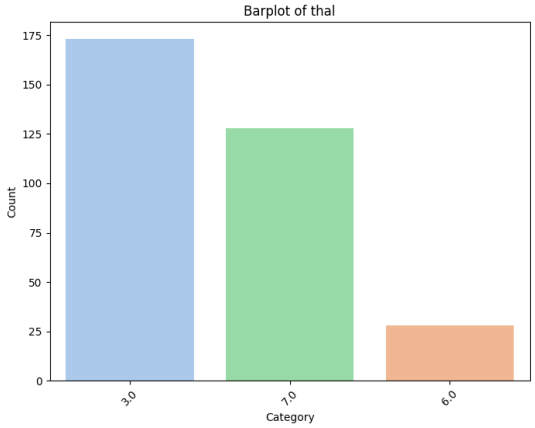
- تعداد مشاهدات: 329 مورد در این مطالعه بررسی شده‌اند.
- دسته‌ها: مقادیر thal به سه دسته تقسیم شده است:
  - دسته ۳.۰: ۱۷۳ مورد (حدود ۵۲.۶٪)
  - دسته ۷.۰: ۱۲۸ مورد (حدود ۳۸.۹٪)
  - دسته ۶.۰: ۲۸ مورد (حدود ۸.۵٪)

تحلیل نمودار میله‌ای:

توزیع نامتقارن: نمودار میله‌ای نشان می‌دهد که توزیع مقادیر thal به شدت نامتقارن است. اکثریت قریب به اتفاق موارد (حدود ۵۲.۶٪) در دسته ۳.۰ قرار دارند.

- تفاوت بین فراوانی دسته‌ها بسیار زیاد است و نشان‌دهنده تفاوت معنی‌داری در مقادیر thal موارد است.

```
========================================
Analyzing Field:      thal
========================================
Categorical Field
thal
3.0   173
7.0   128
6.0    28
Name:      count, dtype:      int64
```



۱۳. C (متغیر هدف)

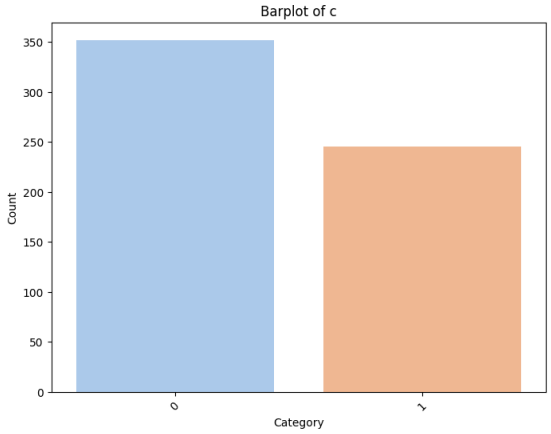
تحلیل آمار توصیفی

- تعداد کل مشاهدات: 597 مورد در این مطالعه بررسی شده‌اند.
- دسته‌ها: متغیر c به دو دسته تقسیم شده است:
  - دسته ۰: 352 مورد (حدود ۵۹ درصد)
  - دسته ۱: 245 مورد (حدود ۴۱ درصد)

تحلیل نمودار میله‌ای:

- توزیع نامتقارن: نمودار میله‌ای نشان می‌دهد که توزیع مقادیر متغیر c به شدت نامتقارن است. اکثریت قریب به اتفاق موارد (حدود ۵۹ درصد) در دسته ۰ قرار دارند.
- تفاوت بین فراوانی دو دسته بسیار زیاد است و نشان‌دهنده تفاوت معنی‌داری در مقادیر متغیر c موارد است.

```
========================================
Analyzing Field:        c
========================================
Categorical Field
c
0    352
1    245
Name:        count, dtype:        int64
```

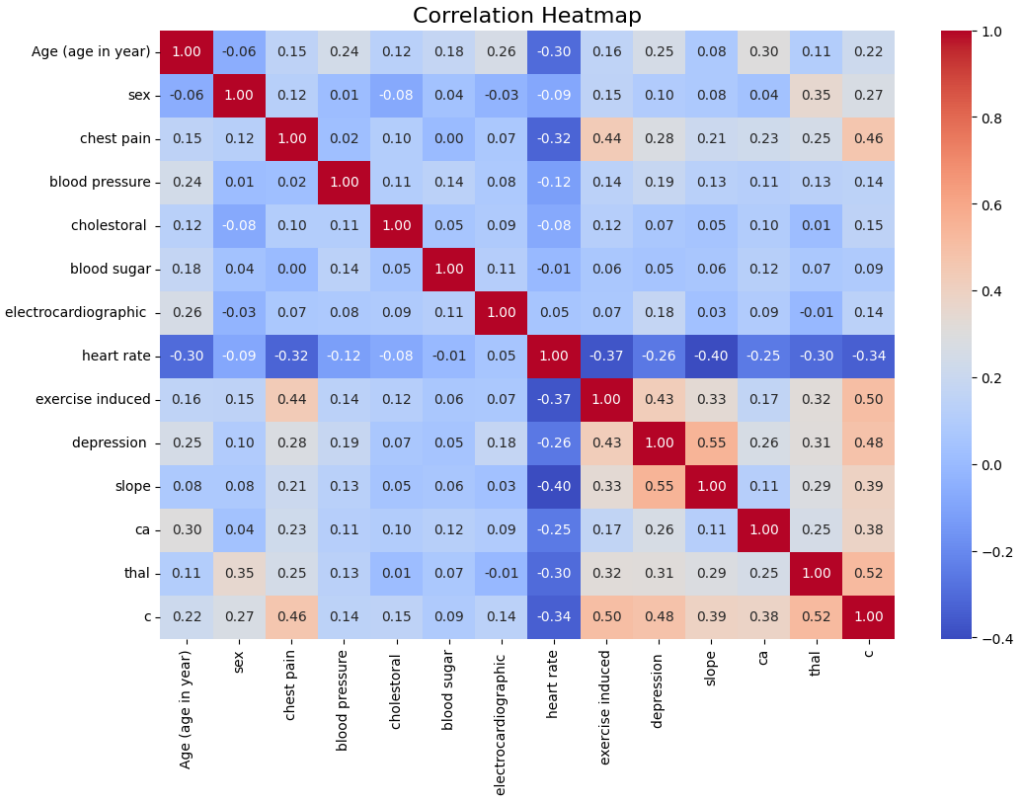


تحلیل همبستگی‌های بین ویژگی‌ها

ماتریس همبستگی، قدرت و جهت روابط خطی بین متغیرهای مختلف مجموعه داده را نشان می‌دهد. مقادیر همبستگی در بازه [-1, 1] قرار دارند:

- مقادیر نزدیک به ۱: همبستگی مثبت قوی (افزایش یک ویژگی باعث افزایش دیگری می‌شود).
- مقادیر نزدیک به -۱: همبستگی منفی قوی (افزایش یک ویژگی باعث کاهش دیگری می‌شود).
- مقادیر نزدیک به ۰: عدم وجود رابطه خطی قوی.

ماتریس همبستگی مجموعه داده Heart Data تصویر زیر می باشد:



تحلیل روابط کلیدی:

۱. ویژگی chest pain و c:

- مقدار همبستگی نشان‌دهنده یک رابطه مثبت نسبتاً قوی بین درد قفسه سینه و متغیر c است. این رابطه نشان می‌دهد که نوع درد قفسه سینه می‌تواند تأثیرگذار بر پیش‌بینی c و بیماری قلبی باشد.

۲. ویژگی exercise induced و thal:

- مقدار همبستگی نشان‌دهنده یک رابطه مثبت متوسط است. این می‌تواند نشان دهد که نوع تالاسمی ممکن است با اثرات ورزش روی بیمار مرتبط باشد.

۳. ویژگی slope و depression:

- مقدار همبستگی حاکی از یک رابطه مثبت نسبتاً قوی است. این ارتباط می‌تواند نشان دهد که شیب قطعه ST با میزان افسردگی (شاید به عنوان نماینده مشکلات قلبی) در ارتباط است.

۴. ویژگی heart rate و exercise induced:

- مقدار همبستگی نشان‌دهنده یک رابطه منفی است. این ممکن است نشان دهد که افرادی که در اثر ورزش دچار مشکلات می‌شوند، احتمالاً ضربان قلب پایین‌تری دارند.

۵. ویژگی ca و Age:

- مقدار همبستگی حاکی از یک رابطه مثبت است. این نشان می‌دهد که با افزایش سن، تعداد انسداد عروق قلبی ممکن است افزایش یابد.

۶. ویژگی‌های با همبستگی پایین یا نزدیک به صفر:

متغیرهایی مانند blood sugar و electrocardiographic تقریباً هیچ رابطه خطی مشخصی با دیگر ویژگی‌ها ندارند. این متغیرها ممکن است تأثیر مستقیمی بر سایر ویژگی‌ها نداشته باشند.

۲. مدیریت داده‌های گمشده

در فرایند جمع‌آوری، پردازش و تحلیل داده‌ها، اغلب با مسئله‌ی وجود داده‌های گمشده یا ناقص مواجه می‌شویم. این داده‌ها می‌توانند به دلایل مختلفی مانند خطاهای انسانی، مشکلات فنی در جمع‌آوری داده‌ها، یا عدم پاسخگویی افراد به پرسش‌ها از بین رفته باشند. وجود داده‌های گمشده می‌تواند بر دقت و اعتبار نتایج تحلیل‌ها تأثیرگذار باشد و در صورت عدم مدیریت صحیح، به نتایج نادرست و گمراه‌کننده منجر شود.

انتخاب روش مناسب برای مدیریت داده‌های گمشده به عوامل مختلفی بستگی دارد، از جمله:

- میزان داده‌های گمشده: اگر تعداد داده‌های گمشده کم باشد، ممکن است حذف مشاهدات یا جایگزینی با مقدار ثابت مناسب باشد. اما اگر تعداد داده‌های گمشده زیاد باشد، استفاده از روش‌های پیشرفته‌تر مانند روش‌های یادگیری ماشین توصیه می‌شود.
- نوع داده :روش‌های مدیریت داده‌های گمشده برای داده‌های کمی و کیفی متفاوت است.

دراین پروژه مدیریت داده های گمشده براساس نوع داده صورت گرفت به این ترتیب که برای ویژگی های عددی مقادیر گمشده با میانه و برای ویژگی های دسته ای با مد جایگزین شدند.

جدول و نمودار زیر نتایج بررسی تعداد مقادیر گمشده در مجموعه داده برای هر ویژگی است:

```
========================================
Step: Exploring and Handling Missing Values
========================================
Step: Detecting Missing Values
Missing Values per Column:
blood pressure              1
cholestoral                23
blood sugar                 8
electrocardiographic        1
heart rate                  1
exercise induced            1
slope                     190
ca                        294
thal                      268
dtype: int64
```

همچنین میتوانید روش رسیدگی به مقادیر گمشده هر ویژگی را مشاهده نمائید.

```
Step: Handling Missing Values
Filling missing values in blood pressure with median.
Filling missing values in cholestoral with median.
Filling missing values in blood sugar with mode.
Filling missing values in electrocardiographic  with mode.
Filling missing values in heart rate with median.
Filling missing values in exercise induced with mode.
Filling missing values in slope with mode.
Filling missing values in ca with mode.
Filling missing values in thal with mode.
Missing values handled.
```

۳. مدیریت داده های تکراری

داده‌های تکراری یکی از مشکلات رایج در مجموعه داده‌ها هستند که می‌توانند نتایج تحلیل‌ها را تحت تأثیر قرار دهند. روش‌های مدیریت داده‌های تکراری به نوع داده و هدف تحلیل بستگی دارد. در این پروژه یک سطر داده تکراری وجود داشت که حذف شد.

```
========================================
Step: Detecting and Removing Duplicates
========================================

Step: Detecting Duplicate Rows
Number of duplicate rows: 1

Step: Removing Duplicate Rows
Removed 1 duplicate rows.
Rows remaining: 596
```

۴. مدیریت داده های نویزی

داده‌های نویزی به داده‌هایی گفته می‌شود که حاوی خطا، بی‌دقتی یا اطلاعات اضافی هستند و از دقت و صحت داده‌های اصلی می‌کاهند. این نویزها می‌توانند در فرآیند جمع‌آوری، ذخیره‌سازی یا پردازش داده‌ها ایجاد شوند. مدیریت موثر داده‌های نویزی، گامی حیاتی در پیش‌پردازش داده‌ها و بهبود دقت مدل‌های یادگیری ماشین است.

روش‌های مدیریت داده‌های نویزی به نوع داده و ماهیت نویز بستگی دارد.

داده‌های کمی:(Numerical)

- تشخیص داده‌های پرت :(Outliers)
  - روش‌های آماری :استفاده از روش‌هایی مانند، نمودار جعبه‌ای و روش‌های مبتنی بر چارک‌ها برای شناسایی داده‌های پرت.
  - روش‌های مبتنی بر چگالی.
- هموارسازی :(Smoothing)
  - میانگین‌گیری محلی :جایگزینی مقدار هر داده با میانگین همسایگان آن.
  - فیلترهای دیجیتال :استفاده از فیلترهای مختلف برای کاهش نویز.
- تبدیلات :
  - تبدیل لگاریتمی :برای داده‌های با توزیع نامتقارن.
  - استانداردسازی :برای مقیاس‌بندی داده‌ها.

داده‌های کیفی:(Categorical)

- حذف سوابق نادرست :حذف سوابقی که به وضوح نادرست یا غیرقابل اعتماد هستند.
- تبدیل به داده‌های عددی :تبدیل داده‌های کیفی به داده‌های عددی برای استفاده از روش‌های عددی.
- استفاده از الگوریتم‌های خوشه‌بندی :برای شناسایی و حذف داده‌های نویزی که از سایر داده‌ها فاصله زیادی دارند.

در این پروژه داده های دسته ای براساس فراوانی مقادیرشان از نظر و جود داده های نویزی مورد بررسی قرار گرفتند، همین طور که در خروجی مشاهده می نمائید و براساس اطلاعاتی که از ویژگی های مجموعه داده وجود دارد، به جز یک مقدار در ویژگـی Ca کـه برابر ۹ است، مقادیر خارج از محدوده ویژگی ها مشاهده نشد.

```
======================================
Step: Detecting and Removing Noise in Categorical Features
======================================

Step: Detecting Noise in Features

Unique values in sex:
sex
1    419
0    177
Name: count, dtype: int64

Unique values in chest pain:
chest pain
4    267
2    155
3    140
1     34
Name: count, dtype: int64

Unique values in blood sugar:
blood sugar
0.0    531
1.0     65
Name: count, dtype: int64

Unique values in electrocardiographic :
electrocardiographic
0.0    386
2.0    154
1.0     56
Name: count, dtype: int64

Unique values in exercise induced:
exercise induced
0.0    408
1.0    188
Name: count, dtype: int64

Unique values in slope:
slope
2.0    420
1.0    154
3.0     22
Name: count, dtype: int64

Unique values in ca:
ca
0.0    472
1.0     65
2.0     38
```

```
3.0     20
9.0      1
Name: count, dtype: int64

Unique values in thal:
thal
3.0    440
7.0    128
6.0     28
Name: count, dtype: int64

Unique values in c:
c
0    351
1    245
Name: count, dtype: int64
Noise detection completed.
```

۵. مدیریت داده‌های پرت (Outliers)

داده‌های پرت به داده‌هایی گفته می‌شود که به طور قابل توجهی از سایر داده‌ها متفاوت هستند و ممکن است بر نتایج تحلیل آماری تأثیر منفی بگذارند. این داده‌ها می‌توانند ناشی از خطاهای اندازه‌گیری، خطاهای ورود داده‌ها، یا رخدادهای نادر و غیرمعمول باشند. مدیریت صحیح داده‌های پرت برای دستیابی به نتایج دقیق و قابل اعتماد در تحلیل داده‌ها بسیار مهم است.

روش های تشخیص داده های پرت

روش IQR (Interquartile Range)

IQR یا دامنه میان‌چارکه‌ای، فاصله بین چارک اول و سوم یک مجموعه داده است. این شاخص نشان‌دهنده پراکندگی داده‌ها است. با استفاده از IQR می‌توانیم داده‌های پرت را به صورت زیر شناسایی کنیم:

- محاسبه حد پایین Q1 - 1.5 * IQR
- محاسبه حد بالا Q3 + 1.5 * IQR
- هر داده‌ای که خارج از این محدوده باشد، به عنوان یک داده پرت بالقوه در نظر گرفته می‌شود.
- به توزیع داده‌ها حساسیت کمتری دارد و برای داده‌های با توزیع غیر نرمال مناسب‌تر است.
- کمتر تحت تأثیر داده‌های پرت شدید قرار می‌گیرد.

روش Z-Score

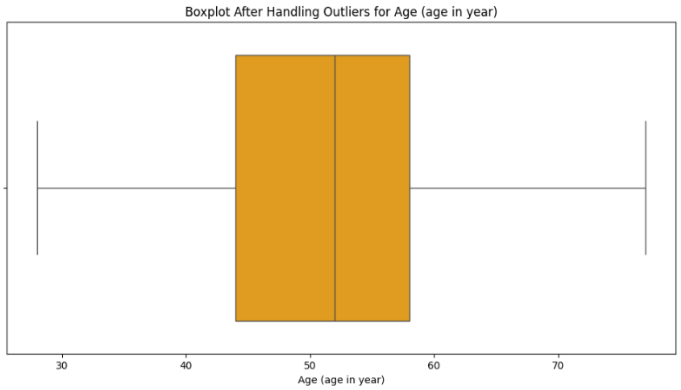
Z-Score، فاصله یک داده تا میانگین را بر حسب انحراف استاندارد بیان می‌کند. به عبارت دیگر، Z-Score نشان می‌دهد که یک داده چند انحراف استاندارد از میانگین فاصله دارد. به طور معمول، داده‌هایی که Z-Score آن‌ها از ۳ یا کمتر از -۳ باشد، به عنوان داده‌های پرت در نظر گرفته می‌شوند.

- برای داده‌هایی با توزیع تقریباً نرمال مناسب‌تر است.
- به داده‌های پرت حساس‌تر است.

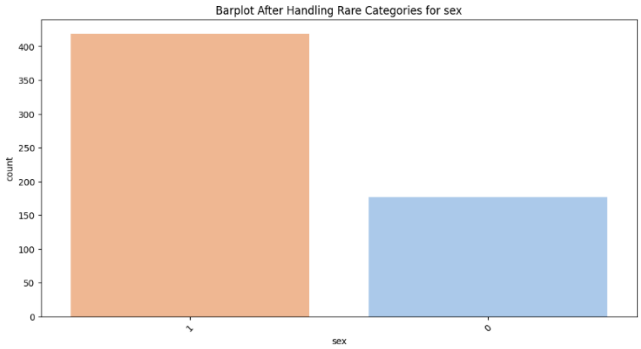
در این پروژه، برای تشخیص داده‌های پرت از دو روش Z-Score, IQR استفاده شده است به این ترتسیب کـه ابتـدا بـا اسـتفاده از روش آزمون Shapiro-Wilk توزیع داده ها تشخیص داده میشـود و اگـر توزیـع داده هـای نرمـال باشـد روش Z-Score و در صورت غیرنرمال بودن توزیع داده ها از روش IQR استفاده میشود.

برای رسیدگی به داده های پرت شناسایی شده برای داده های عددی از روش Clipping استفاده شد و با مقادیر آسـتانه بـرای هـر ویژگی جایگزین شدند و برای داده های دسته ای در صورتی که مقادیر یک یا چند حالت در یک ویژگی کمتـر از ۵ درصـد داده هـا باشد در دسته Others قرار گرفته است. می توانید خروجی عملیات انجام شده را در تصویر مشاهده نمائید.

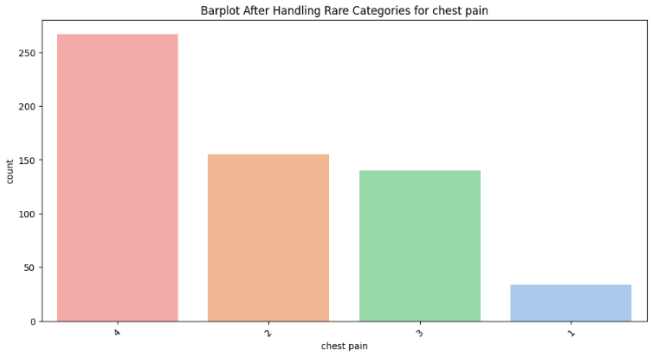
```
========================================
Step: Detecting and Removing Outliers
========================================

Step: Detecting and Handling Outliers

Analyzing Column: Age (age in year)
Type: Numerical

Analyzing Distribution for: Age (age in year)
Shapiro-Wilk Test: Statistic=0.9942, p-value=0.0227
The distribution of Age (age in year) appears to be Non-Normal.
Using IQR to Detect Outliers...
Number of Outliers Detected: 0
```



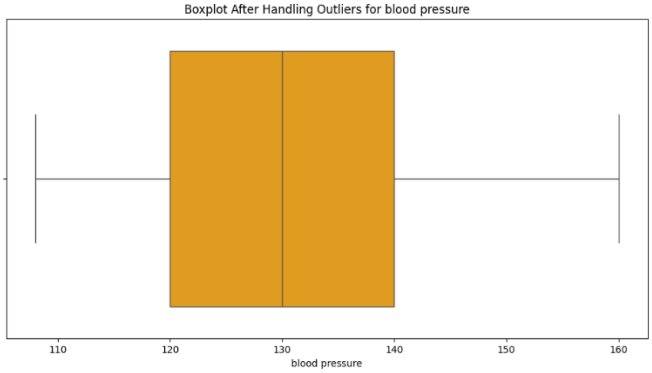

```
Analyzing Column: sex
Type: Categorical
Value Counts:
sex
1    419
0    177
Name: count, dtype: int64
Rare Categories: []
```



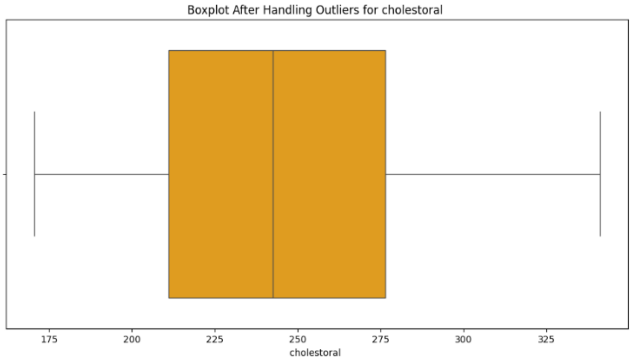

```
Analyzing Column: chest pain
Type: Categorical
Value Counts:
chest pain
4    267
2    155
3    140
1     34
Name: count, dtype: int64
Rare Categories: []
```



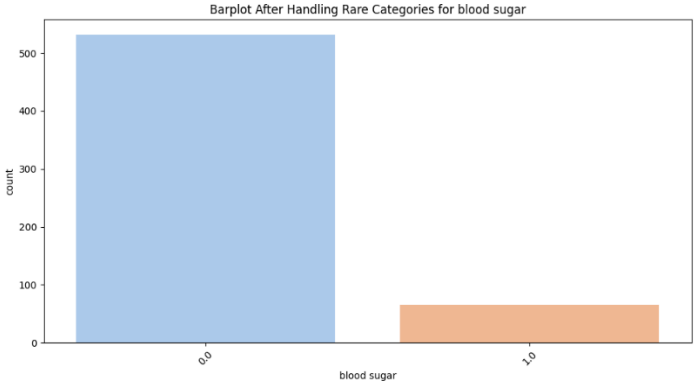

```
Analyzing Column: blood pressure
Type: Numerical

Analyzing Distribution for: blood pressure
Shapiro-Wilk Test: Statistic=0.9599, p-value=0.0000
The distribution of blood pressure appears to be Non-Normal.
Using IQR to Detect Outliers...
Number of Outliers Detected: 17
Handling Outliers by Clipping...
```



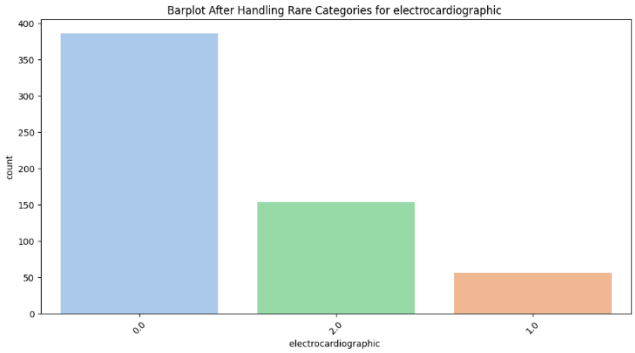
```
Analyzing Column: cholestoral
Type: Numerical

Analyzing Distribution for: cholestoral
Shapiro-Wilk Test: Statistic=0.9199, p-value=0.0000
The distribution of cholestoral  appears to be Non-Normal.
Using IQR to Detect Outliers...
Number of Outliers Detected: 19
Handling Outliers by Clipping...
```



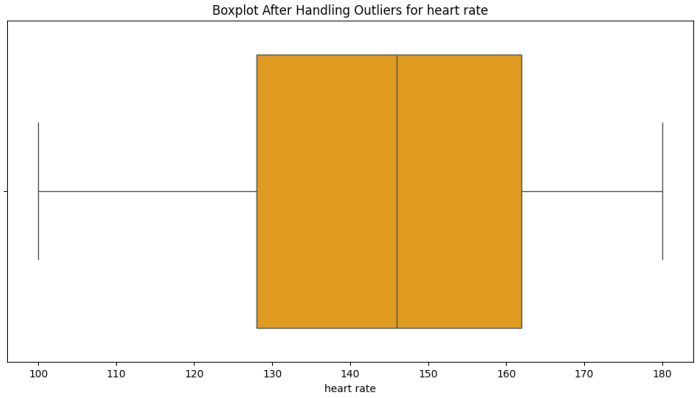

```
Analyzing Column: blood sugar
Type: Categorical
Value Counts:
blood sugar
0.0    531
1.0     65
Name: count, dtype: int64
Rare Categories: []
```



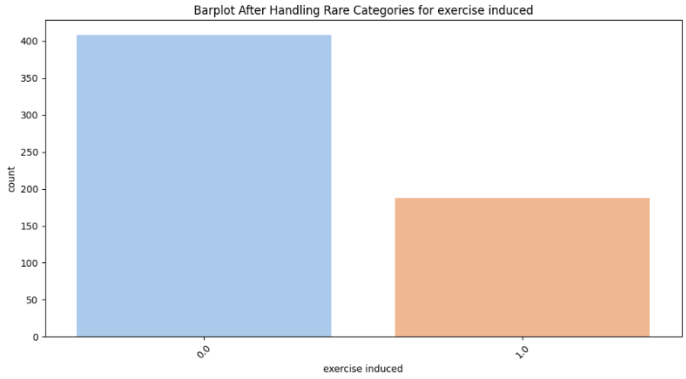

```
Analyzing Column: electrocardiographic
Type: Categorical
Value Counts:
electrocardiographic
0.0    386
2.0    154
1.0     56
Name: count, dtype: int64
Rare Categories: []
```



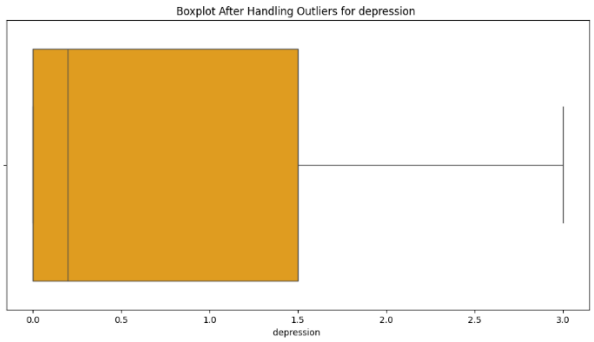

```
Analyzing Column: heart rate
Type: Numerical

Analyzing Distribution for: heart rate
Shapiro-Wilk Test: Statistic=0.9862, p-value=0.0000
The distribution of heart rate appears to be Non-Normal.
Using IQR to Detect Outliers...
Number of Outliers Detected: 1
Handling Outliers by Clipping...
```



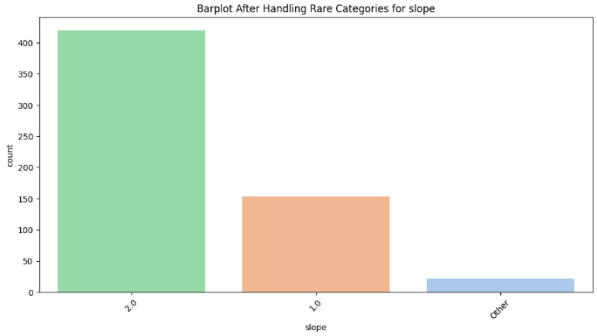

```
Analyzing Column: exercise induced
Type: Categorical
Value Counts:
exercise induced
0.0    408
1.0    188
Name: count, dtype: int64
Rare Categories: []
```



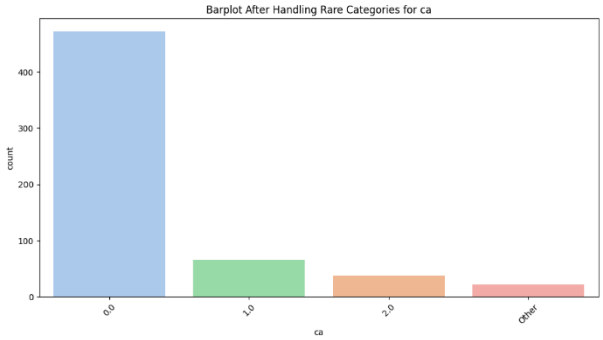
```
Analyzing Column: depression
Type: Numerical

Analyzing Distribution for: depression
Shapiro-Wilk Test: Statistic=0.7803, p-value=0.0000
The distribution of depression  appears to be Non-Normal.
Using IQR to Detect Outliers...
Number of Outliers Detected: 11
Handling Outliers by Clipping...
```



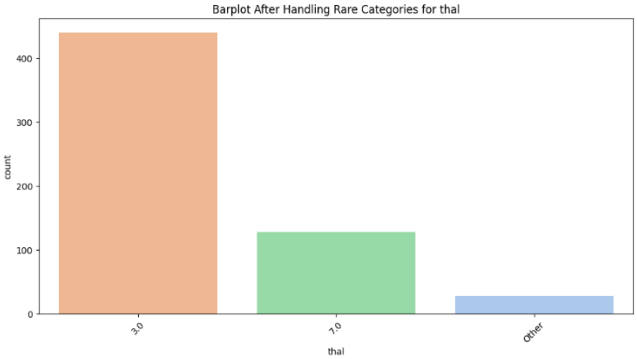

```
Analyzing Column: slope
Type: Categorical
Value Counts:
slope
2.0    420
1.0    154
3.0     22
Name: count, dtype: int64
Rare Categories: [3.0]
Handling Rare Categories by Combining into 'Other'
```



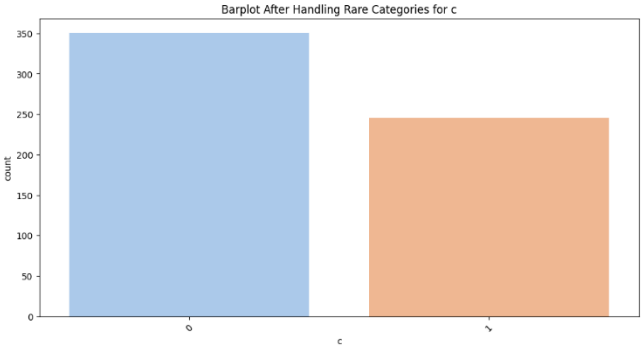

```
Analyzing Column: ca
Type: Categorical
Value Counts:
ca
0.0    472
1.0     65
2.0     38
3.0     20
9.0      1
Name: count, dtype: int64
Rare Categories: [3.0, 9.0]
Handling Rare Categories by Combining into 'Other'
```




```
Analyzing Column: thal
Type: Categorical
Value Counts:
thal
3.0    440
7.0    128
6.0     28
Name: count, dtype: int64
Rare Categories: [6.0]
Handling Rare Categories by Combining into 'Other'
```




```
Analyzing Column: c
Type: Categorical
Value Counts:
c
0    351
1    245
Name: count, dtype: int64
Rare Categories: []
```



٦. خلاصه نتایج حاصل از پاکسازی مجموعه داده

نتایج بدست آمده از پاکسازی و ذخیره سازی مجموعه داده را می توانید مشاهده نمائید. چنانچه مشخص است بعد از عملیات پاکسازی مقادیر داده های گمشده و تکراری صفر است.

```
========================================
Step: Analyzing Data After Cleaning
========================================
Data Analysis - After Cleaning
Rows: 596
Columns: 14
Missing Values: 0
Duplicate Rows: 0
Categorical Features: ['sex', 'chest pain', 'blood sugar',
'electrocardiographic ', 'exercise induced', 'slope', 'ca', 'thal', 'c']
Numerical Features: ['cholestoral ', 'blood pressure', 'heart rate',
'depression ', 'Age (age in year)']

Comparison of Data Statistics Before and After Cleaning:
           Metric  Before Cleaning  After Cleaning
0            Rows              597             596
1         Columns               14              14
2  Missing Values              787               0
3  Duplicate Rows                1               0
========================================
Step: Save new dataset After Cleaning
========================================
 Cleaned dataset saved.
```

٧. آماده‌سازی داده‌ها

آماده‌سازی داده‌ها (Data Preparation)یکی از مراحل حیاتی در هر پروژه تحلیل داده است. این مرحله شامل مجموعه عملیاتی است که داده‌هایی پاکسازی شده به فرمت مناسب برای مدلسازی تبدیل میشوند. این مرحله شامل تقسیم داده های آموزش آزمایش و بررسی توازن بین کلاس ها و تبدیل داده است. سازگار و قابل تحلیل تبدیل می‌کند. هدف اصلی از آماده‌سازی داده‌ها، ایجاد یک پایگاه داده با کیفیت بالا است که بتوان از آن برای مدل‌سازی، تجزیه و تحلیل و تصمیم‌گیری استفاده کرد.

مراحل کلیدی در آماده‌سازی داده‌ها

۸. تقسیم داده‌ها

تقسیم داده‌ها به مجموعه‌های آموزش، اعتبارسنجی و آزمون یکی از مراحل مهم در آماده‌سازی داده‌ها است. این کار به منظور ارزیابی عملکرد مدل‌های یادگیری ماشین انجام می‌شود.

- مجموعه آموزش :برای آموزش مدل استفاده می‌شود.
- مجموعه اعتبارسنجی :برای تنظیم پارامترهای مدل استفاده می‌شود.
- مجموعه آزمون :برای ارزیابی نهایی عملکرد مدل استفاده می‌شود.

در این پروژه مجموعه داده به دو بخش داده آموزش و آزمون با نسبت ۸۰ به ۲۰ تقسیم شد.

۹. بررسی توازن داده‌ها

توازن داده‌ها به معنای توزیع یکسان نمونه‌ها در هر کلاس است. در مسائل طبقه‌بندی، اگر داده‌ها نامتوازن باشند (یعنی تعداد نمونه‌ها در کلاس‌های مختلف بسیار متفاوت باشد)، ممکن است مدل یادگیری ماشین به سمت کلاس اکثریت گرایش پیدا کند. مهمترین نکته در مدیریت داده های نامتوازن انجام عملیات تنها بر روی **داده های آموزش** است. برای حل این مشکل می‌توان از روش‌هایی مانند:

- Undersampling: کاهش تعداد نمونه‌های کلاس اکثریت
- Oversampling: افزایش تعداد نمونه‌های کلاس اقلیت
- SMOTE (Synthetic Minority Over-sampling Technique): ایجاد نمونه‌های مصنوعی برای کلاس اقلیت.

در این پروژه در تابع detect_handle_imbalance بررسی شد که اگر تعداد مقادیر یک کلاس کمتر از ۰.۲ است، عملیات SMOTE جهت توازن داده برای کلاس ها انجام شود.

۱۰. تبدیل داده‌ها

در مرحله آماده‌سازی داده، تبدیل داده یکی از مهم‌ترین مراحل است. این مرحله به ما کمک می‌کند تا داده‌ها را به فرمتی مناسب برای مدل‌سازی و تحلیل تبدیل کنیم. نوع تبدیل داده‌ها به نوع ویژگی (کمی یا کیفی) بستگی دارد.

انواع تبدیل برای ویژگی‌های کمی:

نرمال‌سازی (Normalization) :

نرمال‌سازی مقیاس‌بندی داده‌ها به طوری که همه ویژگی‌ها در محدوده مشخصی (مثلاً ۰ تا ۱) قرار بگیرند. این کار به خصوص در الگوریتم‌هایی که به مقیاس داده‌ها حساس هستند (مانند الگوریتم‌های مبتنی بر فاصله) مفید است. روش‌های رایج نرمال‌سازی عبارتند از :

- Min-Max Normalization
- Z-score Normalization

استانداردسازی (Standardization)

تبدیل داده‌ها به طوری که میانگین آن‌ها صفر و انحراف استاندارد آن‌ها یک شود. این روش نیز برای الگوریتم‌هایی که به مقیاس داده‌ها حساس هستند مفید است.

- **مقیاس‌بندی لگاریتمی :**برای داده‌هایی که توزیع لگاریتمی دارند، استفاده از مقیاس‌بندی لگاریتمی می‌تواند مفید باشد.
- **باکت‌بندی :(Binning)** تقسیم مقادیر پیوسته به گروه‌های گسسته.

انواع تبدیل برای ویژگی‌های کیفی

- **One-Hot Encoding**: برای هر مقدار ممکن از ویژگی یک ستون جدید ایجاد می‌کنیم و اگر آن مقدار برای یک نمونه خاص وجود داشته باشد، مقدار آن ستون را ۱ و در غیر این صورت ۰ قرار می‌دهیم.
- **Label Encoding**: به هر مقدار یک عدد منحصر به فرد اختصاص می‌دهیم. این روش برای ویژگی‌هایی که ترتیب خاصی دارند (مانند رتبه) مناسب است.
- **کدگذاری با استفاده از اعداد صحیح :**به هر مقدار یک عدد صحیح اختصاص می‌دهیم.
- **Embedding**: برای ویژگی‌های کیفی با تعداد مقادیر زیاد، می‌توان از روش Embedding استفاده کرد که در آن هر مقدار به یک بردار چگال نگاشت می‌شود.

در این پروژه از روش Min-Max Normalization برای تبدیل داده های عددی و از روش Label Encoding برای تبدیل داده های دسته ای استفاده شد. دقت کنید حاصل عملیات تبدیل بر روی داده های آموزش بر روی داده های آزمون فیت میشود. در نهایت مجموعه داده جدید با هدف قابلیت استفاده مجدد و صرفه جویی در زمات ذخیره شد.

پس از انجام عملیات پاکسازی و آماده سازی داده، نوبت به مدلسازی میرسد، در ادامه الگوریتم های مدلسازی استفاده شده و پارامتر های آنها به تفصیل شرح داده میشود.

۱۱.مدلسازی

مدلسازی در یادگیری ماشین و علم داده، فرآیندی است که در آن از داده‌های موجود برای ایجاد یک مدل ریاضی استفاده می‌شود. این مدل می‌تواند برای پیش‌بینی، طبقه‌بندی، خوشه‌بندی و سایر وظایف تحلیل داده‌ها به کار رود. هدف اصلی مدلسازی، یافتن الگوها و روابط پنهان در داده‌ها و استفاده از این الگوها برای تصمیم‌گیری‌های آگاهانه است. در این پروژه از الگوریتم‌های بالا ابتدا با مقادیر پارامترهای پیش فرض برای مدلسازی استفاده شد و پس از بررسی نتایج با هدف بهینه سازی پارامترها از الگوریتم Grid Search استفاده و بهترین پارامترها انتخاب شد. برای ارزیابی عملکرد مدل‌ها و جلوگیری از Overfitting از K-fold Cross Validation استفاده شد. تمامی موارد تست شده و کد الگوریتم های پیاده شده به پیوست تقدیم میگردد.

الگوریتم های استفاده شده در مدلسازی

رگرسیون لجستیک(Logistic Regression)

رگرسیون لجستیک علی‌رغم نامش، بیشتر برای مسائل **طبقه‌بندی** استفاده می‌شود. این مدل یک تابع سیگموئید را برای تبدیل خروجی به یک احتمال بین ۰ تا ۱ به کار می‌برد. به عبارت دیگر، رگرسیون لجستیک احتمال وقوع یک رویداد را بر اساس ویژگی‌های ورودی تخمین می‌زند. کاربرد عمده آن در طبقه بندی است.

پارامترها :

- C: کنترل قدرت Regularization، مقادیر بالاتر C باعث ایجاد مدل پیچیده‌تری می‌شود که ممکن است به داده‌های آموزش بیش از حد برازش شود (overfitting). مقادیر پایین‌تر C باعث ایجاد مدل ساده‌تری می‌شود که ممکن است به داده‌های آموزش کم‌برازش شود(underfitting) .
- Solver: الگوریتم بهینه‌سازی است که برای یافتن بهترین ضرایب مدل استفاده می‌شود. انتخاب solver مناسب می‌تواند بر سرعت و دقت مدل تأثیر بگذارد.

جنگل تصادفی(Random Forest)

جنگل تصادفی یک الگوریتم انسامبلی است که از مجموعه‌ای از درختان تصمیم تشکیل شده است. هر درخت تصمیم به تنهایی یک مدل ساده است، اما ترکیب آن‌ها می‌تواند مدل پیچیده‌تری را ایجاد کند که توانایی تعمیم‌پذیری بهتری دارد. کاربرد عمده آن در طبقه بندی، رگرسیون و انتخاب ویژگی است.

پارامترها :

- n_estimators: تعداد درختان در جنگل. هر چه تعداد درختان بیشتر باشد، مدل پیچیده‌تر می‌شود و ممکن است دقت آن افزایش یابد، اما زمان آموزش نیز بیشتر می‌شود.
- max_depth : حداکثر عمق هر درخت. درختان عمیق‌تر می‌توانند الگوهای پیچیده‌تری را مدل‌سازی کنند، اما ممکن است به داده‌های آموزش بیش از حد برازش شوند.

XGBoost

XGBoost (eXtreme Gradient Boosting) یک الگوریتم تقویت گرادیان است که بر پایه درخت‌های تصمیم ساخته شده است. این الگوریتم بسیار سریع و دقیق است و در بسیاری از مسابقات یادگیری ماشین برنده شده است. کاربرد عمده آن در طبقه بندی، رگرسیون است.

پارامترها :

- n_estimators:تعداد درختان. مشابه جنگل تصادفی.
- learning_rate: نرخ یادگیری. نرخ یادگیری پایین باعث می‌شود مدل به آرامی یاد بگیرد و احتمال overfitting کاهش می‌یابد.
- max_depth: حداکثر عمق هر درخت. مشابه جنگل تصادفی.

K-Nearest Neighbors (KNN)

KNNیکی از ساده‌ترین الگوریتم‌های یادگیری ماشین است. در این الگوریتم، برای طبقه‌بندی یک نمونه جدید، به k نزدیک‌ترین نمونه در داده‌های آموزشی نگاه می‌کنیم و بر اساس رای اکثریت، کلاس نمونه جدید را تعیین می‌کنیم. . کاربرد عمده آن در طبقه بندی، رگرسیون است.

پارامترها :

- n_neighbors: تعداد همسایگان نزدیک. مقدار k بر پیچیدگی مدل تأثیر می‌گذارد. مقادیر کوچک k باعث ایجاد مدل پیچیده‌تری می‌شود و مقادیر بزرگ k باعث ایجاد مدل ساده‌تری می‌شود.
- Weights: نحوه وزن‌دهی همسایه‌ها.

درخت تصمیم(Decision Tree)

درخت تصمیم یک مدل مبتنی بر قوانین است که برای طبقه‌بندی و رگرسیون استفاده می‌شود. درخت تصمیم با ایجاد یک سری سوالات درباره ویژگی‌های داده‌ها، به یک تصمیم نهایی می‌رسد. کاربرد عمده آن در طبقه بندی، رگرسیون است.

پارامترها :

- max_depth: حداکثر عمق درخت. مشابه جنگل تصادفی.
- Criterion: معیار تقسیم‌بندی گره‌ها. دو معیار اصلی gini و entropy هستند.

شبکه عصبی مصنوعی(Neural Network)

شبکه‌های عصبی الهام گرفته از مغز انسان هستند و از تعداد زیادی نورون مصنوعی تشکیل شده‌اند که در لایه‌های مختلف سازماندهی شده‌اند. شبکه‌های عصبی قادر به یادگیری الگوهای بسیار پیچیده در داده‌ها هستند. کاربرد در طیف گسترده‌ای از مسائل از جمله طبقه‌بندی، رگرسیون، خوشه‌بندی و تولید متن

پارامترها :

hidden_layer_sizes: تعداد نورون‌ها در هر لایه پنهان. تعداد لایه‌ها و تعداد نورون‌ها بر پیچیدگی مدل تأثیر می‌گذارند.

Activation: تابع فعال‌سازی تعیین می‌کند که خروجی یک نورون چگونه محاسبه شود.

learning_rate_init: نرخ یادگیری اولیه.

max_iter: حداکثر تعداد تکرارها. تعداد تکرار بیشتر به مدل اجازه می‌دهد تا بهینه‌تر شود، اما ممکن است زمان آموزش را افزایش دهد.

Solver: الگوریتم بهینه‌سازی.

ماشین بردار پشتیبان(Support Vector Machine)

SVMیک الگوریتم قدرتمند برای مسائل طبقه‌بندی است. SVM سعی می‌کند بهترین خط جداکننده (یا هایپرپلن) را بین داده‌های دو کلاس پیدا کند. . کاربرد عمده آن در طبقه بندی، رگرسیون است.

پارامترها :

- C: پارامتر تنظیم کننده نرمال‌سازی. مشابه رگرسیون لجستیک.
- Kernel: تابع هسته تعیین می‌کند که چگونه داده‌ها در فضای ویژگی نگاشت شوند.

انتخاب مدل مناسب به عوامل مختلفی از جمله:

- نوع مسئله :طبقه‌بندی، رگرسیون، خوشه‌بندی و...
- حجم داده‌ها :برای داده‌های بزرگ، مدل‌هایی مانند XGBoost و جنگل تصادفی مناسب‌تر هستند.
- ویژگی‌های داده‌ها :نوع داده‌ها (کمی، کیفی)، تعداد ویژگی‌ها و توزیع داده‌ها.
- زمان محاسبات :برخی مدل‌ها مانند شبکه‌های عصبی ممکن است زمان آموزش بیشتری نیاز داشته باشند.
- تعمیم‌پذیری :مدل باید بتواند بر روی داده‌های جدید نیز عملکرد خوبی داشته باشد.

کاربردهای اصلی	تعمیم‌پذیری	سرعت آموزش	پیچیدگی	مدل
طبقه‌بندی دو کلاسه	خوب برای داده‌های خطی	سریع	ساده	رگرسیون لجستیک
طبقه‌بندی، رگرسیون	خوب	متوسط	متوسط	جنگل تصادفی
طبقه‌بندی، رگرسیون	بسیار خوب	کندتر از جنگل تصادفی	پیچیده	XGBoost
طیف گسترده‌ای از مسائل	بسیار خوب برای داده‌های پیچیده	کند	بسیار پیچیده	شبکه عصبی
طبقه‌بندی، رگرسیون	متوسط	سریع برای داده‌های کوچک	ساده	KNN
طبقه‌بندی، رگرسیون	ممکن است به داده‌های آموزش بیش از حد برازش شود	سریع	متوسط	درخت تصمیم
طبقه‌بندی	خوب برای داده‌های با ابعاد بالا	کندتر از برخی الگوریتم‌های دیگر	متوسط	SVM

ارزیابی عملکرد مدل‌ها

در یادگیری ماشین، ارزیابی عملکرد مدل‌ها از اهمیت بسیار بالایی برخوردار است. پس از آموزش یک مدل، نیاز داریم تا کیفیت و دقت آن را بسنجیم. اینجاست که معیارهای ارزیابی وارد عمل می‌شوند. این معیارها به ما کمک می‌کنند تا عملکرد مدل را به صورت کمی اندازه‌گیری کرده و مدل‌های مختلف را با هم مقایسه کنیم. انتخاب معیار ارزیابی مناسب به نوع مسئله (طبقه‌بندی، رگرسیون، خوشه‌بندی و ...) و هدف از مدل‌سازی بستگی دارد.

از مهمترین مزایای استفاده از معیارهای ارزیابی می توان به موارد زیر اشاره کرد:

- انتخاب بهترین مدل :از بین چندین مدل آموزش دیده، مدلی را که بهترین عملکرد را دارد انتخاب کنیم.

- تنظیم پارامترها :پارامترهای مدل را به گونه‌ای تنظیم کنیم که بهترین نتیجه را حاصل کنیم.
- درک نقاط قوت و ضعف مدل :با استفاده از معیارهای مختلف، می‌توانیم نقاط قوت و ضعف مدل را شناسایی کرده و برای بهبود آن اقدام کنیم.

از مهمترین معیار های ارزیابی که دراین پروژه نیز برای ارزیابی عملکرد مدل استفاده شده اند، می توان به موارد زیر اشاره کرد:

ماتریس درهم‌ریختگی (Confusion Matrix)

ماتریس درهم‌ریختگی یک جدول است که عملکرد یک مدل طبقه‌بندی را به صورت خلاصه نشان می‌دهد. این ماتریس تعداد نمونه‌های درست و نادرست طبقه‌بندی شده را برای هر کلاس نشان می‌دهد. از روی این ماتریس می‌توانیم معیارهای دیگری مانند Precision ، Recall و F1-Score را محاسبه کنیم.

- TP (True Positive): نمونه‌هایی که به درستی مثبت پیش‌بینی شده‌اند .
- FP (False Positive): نمونه‌هایی که به اشتباه مثبت پیش‌بینی شده‌اند .
- FN (False Negative): نمونه‌هایی که به اشتباه منفی پیش‌بینی شده‌اند .
- TN (True Negative): نمونه‌هایی که به درستی منفی پیش‌بینی شده‌اند.

	پیش‌بینی شده: مثبت	پیش‌بینی شده: منفی
واقعی: مثبت	TP (True Positive)	FP (False Positive)
واقعی: منفی	FN (False Negative)	TN (True Negative)

دقت(Precision)

دقت نشان می‌دهد از بین نمونه‌هایی که مدل به عنوان مثبت پیش‌بینی کرده است، چه تعداد واقعاً مثبت بوده‌اند. به عبارت دیگر، چقدر می‌توانیم به پیش‌بینی‌های مثبت مدل اعتماد کنیم. کاربرد اصلی آن زمانی است که هزینه اشتباه طبقه‌بندی مثبت بالا باشد.

```
Precision = TP / (TP + FP)
```

فراخوانی(Recall)

فراخوانی نشان می‌دهد از بین تمام نمونه‌های مثبت واقعی، چه تعداد توسط مدل به درستی شناسایی شده‌اند. به عبارت دیگر، چقدر مدل توانسته است نمونه‌های مثبت را پیدا کند. . کاربرد اصلی آن زمانی است که هزینه از دست دادن نمونه‌های مثبت بالا باشد

```
Recall = TP / (TP + FN)
```

F1-Score

F1-Score میانگین هارمونیک دقت و فراخوانی است و تعادلی بین این دو معیار ایجاد می‌کند. کاربرد اصلی آن زمانی است که هم دقت و هم فراخوانی مهم باشند و بخواهیم تعادلی بین آن‌ها برقرار کنیم.

```
F1-Score = 2 * (Precision * Recall) / (Precision + Recall)
```

AUC-ROC

AUC (Area Under the Curve)مساحت زیر منحنی ROC است. منحنی ROC نموداری است کـه نـرخ مثبت‌هـای صـحیح (True Positive Rate)را بر حسب نرخ مثبت‌های کاذب (False Positive Rate) رسم می‌کند. AUC نشان می‌دهد که یـک مدل تا چه اندازه می‌تواند نمونه‌های مثبت را از نمونه‌های منفی تشخیص دهد. هرچه AUC به ۱ نزدیک‌تر باشد، عملکرد مدل بهتـر است. کاربرد اصلی آن زمانی است که بخواهیم عملکرد کلی مدل را در آستانه‌های مختلف ارزیابی کنیم.

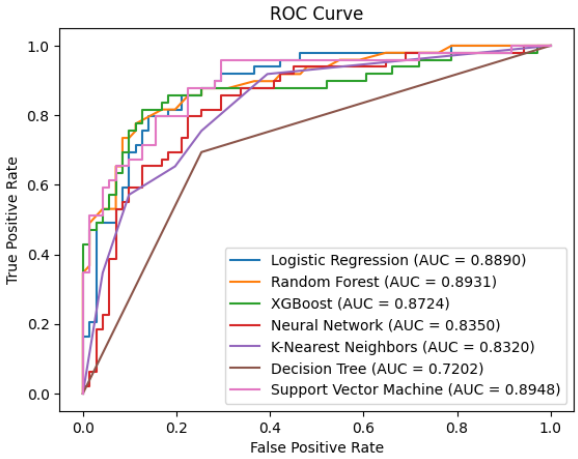
در ادامه شاهد خروجی مرحله مدلسازی با پارامتر های پیشفرض هستید:

```
Training Logistic Regression model...
Evaluating Logistic Regression...
Accuracy: 0.8167
Precision: 0.8140
Recall: 0.7143
F1-Score: 0.7609
AUC-ROC: 0.8890

Confusion Matrix:
[[63  8]
 [14 35]]
Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.89      0.85        71
           1       0.81      0.71      0.76        49

    accuracy                           0.82       120
   macro avg       0.82      0.80      0.81       120
weighted avg       0.82      0.82      0.81       120

Logistic Regression Test Accuracy: 0.8166666666666667

Training Random Forest model...
Evaluating Random Forest...
Accuracy: 0.8417
Precision: 0.8571
Recall: 0.7347
F1-Score: 0.7912
AUC-ROC: 0.8931

Confusion Matrix:
[[65  6]
 [13 36]]
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.92      0.87        71
           1       0.86      0.73      0.79        49

    accuracy                           0.84       120
   macro avg       0.85      0.83      0.83       120
weighted avg       0.84      0.84      0.84       120
```

```
Random Forest Test Accuracy: 0.8416666666666667

Training XGBoost model...
Evaluating XGBoost...
Accuracy: 0.8333
Precision: 0.8085
Recall: 0.7755
F1-Score: 0.7917
AUC-ROC: 0.8724

Confusion Matrix:
[[62  9]
 [11 38]]
Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.87      0.86        71
           1       0.81      0.78      0.79        49

    accuracy                           0.83       120
   macro avg       0.83      0.82      0.83       120
weighted avg       0.83      0.83      0.83       120

XGBoost Test Accuracy: 0.8333333333333334

Training Neural Network model...
Evaluating Neural Network...
Accuracy: 0.7583
Precision: 0.7632
Recall: 0.5918
F1-Score: 0.6667
AUC-ROC: 0.8350

Confusion Matrix:
[[62  9]
 [20 29]]
Classification Report:
              precision    recall  f1-score   support

           0       0.76      0.87      0.81        71
           1       0.76      0.59      0.67        49

    accuracy                           0.76       120
   macro avg       0.76      0.73      0.74       120
weighted avg       0.76      0.76      0.75       120

Neural Network Test Accuracy: 0.7583333333333333

Training K-Nearest Neighbors model...
Evaluating K-Nearest Neighbors...
Accuracy: 0.7417
Precision: 0.6957
Recall: 0.6531
F1-Score: 0.6737
AUC-ROC: 0.8320

Confusion Matrix:
```

```
[[57 14]
 [17 32]]
Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.80      0.79        71
           1       0.70      0.65      0.67        49

    accuracy                           0.74       120
   macro avg       0.73      0.73      0.73       120
weighted avg       0.74      0.74      0.74       120

K-Nearest Neighbors Test Accuracy: 0.7416666666666667

Training Decision Tree model...
Evaluating Decision Tree...
Accuracy: 0.7250
Precision: 0.6538
Recall: 0.6939
F1-Score: 0.6733
AUC-ROC: 0.7202

Confusion Matrix:
[[53 18]
 [15 34]]
Classification Report:
              precision    recall  f1-score   support

           0       0.78      0.75      0.76        71
           1       0.65      0.69      0.67        49

    accuracy                           0.72       120
   macro avg       0.72      0.72      0.72       120
weighted avg       0.73      0.72      0.73       120

Decision Tree Test Accuracy: 0.725

Training Support Vector Machine model...
Evaluating Support Vector Machine...
Accuracy: 0.8000
Precision: 0.7907
Recall: 0.6939
F1-Score: 0.7391
AUC-ROC: 0.8948

Confusion Matrix:
[[62  9]
 [15 34]]
Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.87      0.84        71
           1       0.79      0.69      0.74        49

    accuracy                           0.80       120
   macro avg       0.80      0.78      0.79       120
weighted avg       0.80      0.80      0.80       120
```

```
Support Vector Machine Test Accuracy: 0.8
```




```
Summary of Results:
                              Accuracy  Precision    Recall  F1-Score  \
Logistic Regression           0.816667   0.813953  0.714286  0.760870
Random Forest                 0.841667   0.857143  0.734694  0.791209
XGBoost                       0.833333   0.808511  0.775510  0.791667
Neural Network                0.758333   0.763158  0.591837  0.666667
K-Nearest Neighbors (KNN)     0.741667   0.695652  0.653061  0.673684
Decision Tree                 0.725000   0.653846  0.693878  0.673267
Support Vector Machine (SVM)  0.800000   0.790698  0.693878  0.739130

                               AUC-ROC
Logistic Regression           0.889049
Random Forest                 0.893073
XGBoost                       0.872377
Neural Network                0.835010
K-Nearest Neighbors (KNN)     0.831992
Decision Tree                 0.720178
Support Vector Machine (SVM)  0.894797
```

در مسائل پزشکی، هزینه اشتباه طبقه‌بندی می‌تواند بسیار بالا باشد. بنابراین، علاوه بر Accuracy کلی، باید به Precision و Recall برای هر کلاس نیز توجه شود. در مسائل پزشکی، اغلب تعادل بین Precision و Recall اهمیت دارد. برای مثال، در تشخیص سرطان، شناسایی همه موارد مثبت (فراخوانی بالا) بسیار مهم است، حتی اگر به قیمت افزایش تعداد مثبت کاذب (کاهش دقت) باشد.

K-fold Cross Validation

K-fold Cross Validation یک روش قدرتمند در یادگیری ماشین است که برای ارزیابی عملکرد مدل و جلوگیری از بیش‌برازش (Overfitting)استفاده می‌شود. این روش به ویژه زمانی مفید است که حجم داده‌ها محدود باشد. در این روش، داده‌های آموزشی به K قسمت مساوی تقسیم می‌شوند. در هر تکرار، یکی از این K قسمت به عنوان مجموعه تست و بقیه به عنوان مجموعه آموزش

استفاده می‌شود. به عبارت دیگر، مدل K بار آموزش می‌بیند و ارزیابی می‌شود. در نهایت، نتایج حاصل از هر تکرار میانگین‌گیری شده و به عنوان برآورد نهایی عملکرد مدل در نظر گرفته می‌شود. مزایای اصلی این روش شامل موارد زیر میباشد:

- جلوگیری از بیش‌برازش: با تقسیم داده‌ها به چندین قسمت و آموزش مدل بر روی قسمت‌های مختلف، از وابستگی بیش از حد مدل به داده‌های آموزشی جلوگیری می‌شود.
- برآورد دقیق‌تر عملکرد: با میانگین‌گیری نتایج حاصل از تکرارهای مختلف، برآوردی دقیق‌تر از عملکرد مدل بر روی داده‌های دیده نشده بدست می‌آید.
- انتخاب بهترین مدل: می‌توان از این روش برای مقایسه عملکرد مدل‌های مختلف و انتخاب بهترین مدل استفاده کرد.

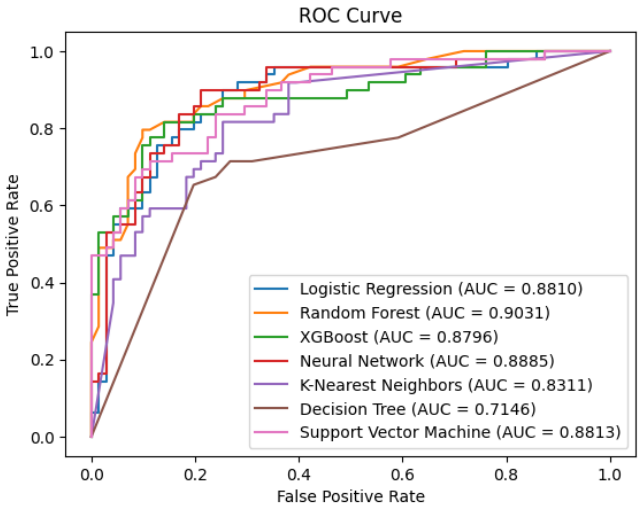
در ادامه شاهد خروجی مرحله مدلسازی با پارامتر های پیشفرض و عملیات K-fold Cross Validation هستید:

```
Training Logistic Regression model...
Evaluating Logistic Regression...
Accuracy: 0.8508
Precision: 0.8571
Recall: 0.7653
F1-Score: 0.8086
AUC-ROC: 0.8842

Confusion Matrix:
[[255  25]
 [ 46 150]]
Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.91      0.88       280
           1       0.86      0.77      0.81       196

    accuracy                           0.85       476
   macro avg       0.85      0.84      0.84       476
weighted avg       0.85      0.85      0.85       476

Logistic Regression Train Accuracy: 0.8508403361344538
Logistic Regression Test Accuracy: 0.8

Training Random Forest model...
Evaluating Random Forest...
Accuracy: 0.8403
Precision: 0.8226
Recall: 0.7806
F1-Score: 0.8010
AUC-ROC: 0.8909

Confusion Matrix:
[[247  33]
 [ 43 153]]
Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.88      0.87       280
           1       0.82      0.78      0.80       196
```

```
    accuracy                           0.84       476
   macro avg       0.84      0.83      0.83       476
weighted avg       0.84      0.84      0.84       476

Random Forest Train Accuracy: 0.8403361344537815
Random Forest Test Accuracy: 0.8416666666666667

Training XGBoost model...
Evaluating XGBoost...
Accuracy: 0.8235
Precision: 0.8077
Recall: 0.7500
F1-Score: 0.7778
AUC-ROC: 0.8865

Confusion Matrix:
[[245  35]
 [ 49 147]]
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.88      0.85       280
           1       0.81      0.75      0.78       196

    accuracy                           0.82       476
   macro avg       0.82      0.81      0.82       476
weighted avg       0.82      0.82      0.82       476

XGBoost Train Accuracy: 0.8235294117647058
XGBoost Test Accuracy: 0.8166666666666667

Training Neural Network model...
Evaluating Neural Network...
Accuracy: 0.8529
Precision: 0.8580
Recall: 0.7704
F1-Score: 0.8118
AUC-ROC: 0.8932

Confusion Matrix:
[[255  25]
 [ 45 151]]
Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.91      0.88       280
           1       0.86      0.77      0.81       196

    accuracy                           0.85       476
   macro avg       0.85      0.84      0.85       476
weighted avg       0.85      0.85      0.85       476

Neural Network Train Accuracy: 0.8529411764705882
Neural Network Test Accuracy: 0.8

Training K-Nearest Neighbors model...
Evaluating K-Nearest Neighbors...
```

```
Accuracy: 0.8235
Precision: 0.8146
Recall: 0.7398
F1-Score: 0.7754
AUC-ROC: 0.8854

Confusion Matrix:
[[247  33]
 [ 51 145]]
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.88      0.85       280
           1       0.81      0.74      0.78       196

    accuracy                           0.82       476
   macro avg       0.82      0.81      0.82       476
weighted avg       0.82      0.82      0.82       476

K-Nearest Neighbors Train Accuracy: 0.8235294117647058
K-Nearest Neighbors Test Accuracy: 0.7583333333333333

Training Decision Tree model...
Evaluating Decision Tree...
Accuracy: 0.7605
Precision: 0.7113
Recall: 0.7041
F1-Score: 0.7077
AUC-ROC: 0.7442

Confusion Matrix:
[[224  56]
 [ 58 138]]
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.80      0.80       280
           1       0.71      0.70      0.71       196

    accuracy                           0.76       476
   macro avg       0.75      0.75      0.75       476
weighted avg       0.76      0.76      0.76       476

Decision Tree Train Accuracy: 0.7605042016806722
Decision Tree Test Accuracy: 0.725

Training Support Vector Machine model...
Evaluating Support Vector Machine...
Accuracy: 0.8298
Precision: 0.8142
Recall: 0.7602
F1-Score: 0.7863
AUC-ROC: 0.9043

Confusion Matrix:
[[246  34]
 [ 47 149]]
```

```
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.88      0.86       280
           1       0.81      0.76      0.79       196

    accuracy                           0.83       476
   macro avg       0.83      0.82      0.82       476
weighted avg       0.83      0.83      0.83       476

Support Vector Machine Train Accuracy: 0.8298319327731093
Support Vector Machine Test Accuracy: 0.8083333333333333
```




```
Summary of Results:
                                Accuracy  Precision    Recall  F1-Score  \
Logistic Regression             0.850840   0.857143  0.765306  0.808625
Random Forest                   0.840336   0.822581  0.780612  0.801047
XGBoost                         0.823529   0.807692  0.750000  0.777778
Neural Network                  0.852941   0.857955  0.770408  0.811828
K-Nearest Neighbors (KNN)       0.823529   0.814607  0.739796  0.775401
Decision Tree                   0.760504   0.711340  0.704082  0.707692
Support Vector Machine (SVM)    0.829832   0.814208  0.760204  0.786280

                                 AUC-ROC
Logistic Regression             0.884220
Random Forest                   0.890944
XGBoost                         0.886534
Neural Network                  0.893203
K-Nearest Neighbors (KNN)       0.885405
Decision Tree                   0.744178
Support Vector Machine (SVM)    0.904264
```

۱۲.بهینه سازی پارامترها

بهینه‌سازی پارامترها یکی از مهم‌ترین مراحل در ساخت و توسعه مدل‌های یادگیری ماشین است. پارامترها، تنظیماتی هستند که قبل از شروع فرایند آموزش به مدل داده می‌شوند و بر عملکرد نهایی مدل تأثیر مستقیم دارند. به عنوان مثال، در یک شبکه عصبی، تعداد لایه‌ها، تعداد نورون‌ها در هر لایه، نرخ یادگیری و تابع فعال‌سازی همگی پارامترهایی هستند که باید به دقت انتخاب شوند. مهمترین مزایای بهینه سازی پارامترها عبارتند از:

- افزایش دقت مدل: انتخاب مناسب پارامترها باعث می‌شود مدل بتواند الگوهای پیچیده‌تری را در داده‌ها شناسایی کند و در نتیجه دقت پیش‌بینی‌ها افزایش یابد.
- کاهش خطا: پارامترهای نامناسب می‌توانند منجر به بیش‌برازش (Overfitting) یا کم‌برازش (Underfitting) شوند که هر دو باعث کاهش دقت مدل می‌شوند.
- بهبود سرعت آموزش: برخی از پارامترها می‌توانند بر سرعت همگرایی الگوریتم آموزش تأثیر بگذارند.
- کاهش پیچیدگی مدل: نتخاب مناسب پارامترها می‌تواند به ساده‌سازی مدل و کاهش زمان محاسبات کمک کند.

روش‌های بهینه‌سازی پارامترها

- Grid Search : در این روش، تمام ترکیبات ممکن از مقادیر پارامترها آزمایش می‌شوند و بهترین ترکیب بر اساس یک معیار ارزیابی انتخاب می‌شود.
- Random Search: در این روش، ترکیبات پارامترها به صورت تصادفی انتخاب می‌شوند و معمولاً کارآمدتر از Grid Search است.

در زیر جدولی از پارامترهای اصلی هر مدل و مقادیر پیشنهادی برای آن‌ها ارائه شده است.

مدل	پارامترها	مقادیر پیشنهادی	توضیحات
رگرسیون لجستیک	C	0.1, 1, 10	کنترل قدرت منظم‌سازی
	solver	liblinear, lbfgs	الگوریتم بهینه‌سازی
جنگل تصادفی	n_estimators	50, 100, 200	تعداد درختان
	max_depth	None, 10, 20	حداکثر عمق هر درخت
XGBoost	n_estimators	50, 100, 200	تعداد درختان
	learning_rate	0.01, 0.1, 0.2	نرخ یادگیری
	max_depth	3, 5, 2007	حداکثر عمق هر درخت
شبکه عصبی	hidden_layer_sizes	(50,), (100,), (100, 50)	تعداد نورون‌ها در هر لایه پنهان
	activation	relu, tanh	تابع فعال‌سازی
	learning_rate_init	0.001, 0.01	نرخ یادگیری اولیه
	max_iter	500, 1000	حداکثر تعداد تکرارها
	solver	adam, lbfgs, sgd	الگوریتم بهینه‌سازی
K-Nearest Neighbors	n_neighbors	3, 5, 2007	تعداد همسایگان نزدیک
	weights	uniform, distance	نحوه وزن‌دهی همسایه‌ها
درخت تصمیم	max_depth	None, 10, 20	حداکثر عمق درخت
	criterion	gini, entropy	معیار تقسیم‌بندی گره‌ها
ماشین بردار پشتیبان	C	0.1, 1, 10	پارامتر تنظیم کننده نرمال‌سازی

	kernel	linear, rbf	تابع هسته

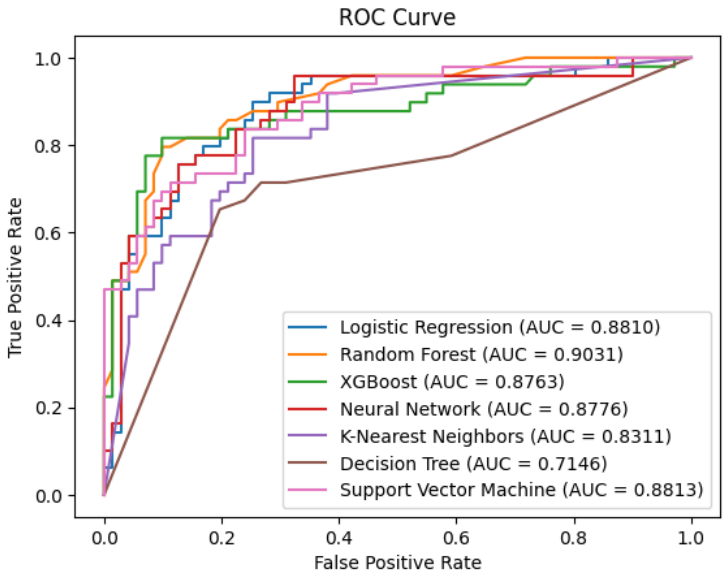
در ادامه شاهد خروجی مرحله مدلسازی برای پارامتر های بهینه شده هستید:

```
Training Logistic Regression model...
Best Parameters for Logistic Regression: {'C': 0.1, 'solver': 'liblinear'}
Evaluating OptimizedLogistic Regression...
Accuracy: 0.8000
Precision: 0.8049
Recall: 0.6735
F1-Score: 0.7333
AUC-ROC: 0.8810

Confusion Matrix:
[[63  8]
 [16 33]]
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.89      0.84        71
           1       0.80      0.67      0.73        49

    accuracy                           0.80       120
   macro avg       0.80      0.78      0.79       120
weighted avg       0.80      0.80      0.80       120

Logistic Regression Test Accuracy: 0.8

Training Random Forest model...
Best Parameters for Random Forest: {'max_depth': None, 'n_estimators': 50}
Evaluating OptimizedRandom Forest...
Accuracy: 0.8417
Precision: 0.8571
Recall: 0.7347
F1-Score: 0.7912
AUC-ROC: 0.9031

Confusion Matrix:
[[65  6]
 [13 36]]
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.92      0.87        71
           1       0.86      0.73      0.79        49

    accuracy                           0.84       120
   macro avg       0.85      0.83      0.83       120
weighted avg       0.84      0.84      0.84       120

Random Forest Test Accuracy: 0.8416666666666667

Training XGBoost model...
Best Parameters for XGBoost: {'learning_rate': 0.1, 'max_depth': 5,
'n_estimators': 50}
```

```
Evaluating OptimizedXGBoost...
Accuracy: 0.8667
Precision: 0.8837
Recall: 0.7755
F1-Score: 0.8261
AUC-ROC: 0.8763

Confusion Matrix:
[[66  5]
 [11 38]]
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.93      0.89        71
           1       0.88      0.78      0.83        49

    accuracy                           0.87       120
   macro avg       0.87      0.85      0.86       120
weighted avg       0.87      0.87      0.87       120

XGBoost Test Accuracy: 0.8666666666666667

Training Neural Network model...
Best Parameters for Neural Network: {'activation': 'relu',
'hidden_layer_sizes': (100,), 'learning_rate_init': 0.001, 'max_iter': 500,
'solver': 'sgd'}
Evaluating OptimizedNeural Network...
Accuracy: 0.7917
Precision: 0.8000
Recall: 0.6531
F1-Score: 0.7191
AUC-ROC: 0.8776

Confusion Matrix:
[[63  8]
 [17 32]]
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.89      0.83        71
           1       0.80      0.65      0.72        49

    accuracy                           0.79       120
   macro avg       0.79      0.77      0.78       120
weighted avg       0.79      0.79      0.79       120

Neural Network Test Accuracy: 0.7916666666666666

Training K-Nearest Neighbors model...
Best Parameters for K-Nearest Neighbors: {'n_neighbors': 5, 'weights':
'distance'}
Evaluating OptimizedK-Nearest Neighbors...
Accuracy: 0.7583
Precision: 0.7083
Recall: 0.6939
F1-Score: 0.7010
AUC-ROC: 0.8311
```

```
Confusion Matrix:
[[57 14]
 [15 34]]
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.80      0.80        71
           1       0.71      0.69      0.70        49

    accuracy                           0.76       120
   macro avg       0.75      0.75      0.75       120
weighted avg       0.76      0.76      0.76       120

K-Nearest Neighbors Test Accuracy: 0.7583333333333333

Training Decision Tree model...
Best Parameters for Decision Tree: {'criterion': 'gini', 'max_depth': 10}
Evaluating OptimizedDecision Tree...
Accuracy: 0.7250
Precision: 0.6600
Recall: 0.6735
F1-Score: 0.6667
AUC-ROC: 0.7146

Confusion Matrix:
[[54 17]
 [16 33]]
Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.76      0.77        71
           1       0.66      0.67      0.67        49

    accuracy                           0.72       120
   macro avg       0.72      0.72      0.72       120
weighted avg       0.73      0.72      0.73       120

Decision Tree Test Accuracy: 0.725

Training Support Vector Machine model...
Best Parameters for Support Vector Machine: {'C': 0.1, 'kernel': 'rbf'}
Evaluating Optimized Support Vector Machine...
Accuracy: 0.8083
Precision: 0.8095
Recall: 0.6939
F1-Score: 0.7473
AUC-ROC: 0.8813

Confusion Matrix:
[[63  8]
 [15 34]]
Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.89      0.85        71
           1       0.81      0.69      0.75        49
```

```
              precision    recall  f1-score   support

    accuracy                         0.81       120
   macro avg       0.81      0.79      0.80       120
weighted avg       0.81      0.81      0.81       120

Support Vector Machine Test Accuracy: 0.8083333333333333
```



Summary of Results:

	Accuracy	Precision	Recall	F1-Score	\
Logistic Regression	0.800000	0.804878	0.673469	0.733333	
Random Forest	0.841667	0.857143	0.734694	0.791209	
XGBoost	0.866667	0.883721	0.775510	0.826087	
Neural Network	0.791667	0.800000	0.653061	0.719101	
K-Nearest Neighbors (KNN)	0.758333	0.708333	0.693878	0.701031	
Decision Tree	0.725000	0.660000	0.673469	0.666667	
Support Vector Machine (SVM)	0.808333	0.809524	0.693878	0.747253	

	AUC-ROC
Logistic Regression	0.881000
Random Forest	0.903133
XGBoost	0.876258
Neural Network	0.877551
K-Nearest Neighbors (KNN)	0.831130
Decision Tree	0.714573
Support Vector Machine (SVM)	0.881288

۱۳.تحلیل نتایج و جمع بندی

در مسائل پزشکی، هزینه اشتباه طبقه‌بندی می‌تواند بسیار بالا باشد. بنابراین، علاوه بر Accuracy ، باید به Precision و Recall و تعادل بین آنها برای هر کلاس نیز توجه شود. در انتخاب مدل مناسب برای پیش‌بینی بیماری قلبی، باید به عوامل مختلفی مانند دقت مورد نیاز، هزینه اشتباه، تفسیرپذیری و پیچیدگی مدل توجه کرد.

با توجه به نتایج حاصل از اجرای مدل های یادگیری ماشین بر روی مجموعه داده Heart Data برای پیش‌بینی بیماری قلبی می توان به این جمع بندی رسید که مدل XGBoost بهترین عملکرد کلی را داشته است. این مدل بالاترین دقت، فراخوانی، F1-Score و AUC-ROC را در بین مدل‌های دیگر دارد. این نشان می‌دهد که XGBoost توانسته است بیماران قلبی را با دقت بسیار خوبی شناسایی کند و تعادل مناسبی بین دقت و فراخوانی برقرار کند. مدل Random Forest نیز عملکرد بسیار خوبی داشته و پس از XGBoost در رتبه دوم قرار دارد. مدل ماشین بردار پشتیبان نیز نتایج قابل قبولی را نشان می‌دهد و در برخی معیارها با XGBoostبرابری می‌کند. Neural Networkو Logistic Regression و KNN این مدل‌ها عملکرد نسبتا خوبی دارند، اما نسبت به XGBoost و Random Forest عملکرد ضعیف‌تری از خود نشان داده اند. Decision Tree این مدل ساده‌ترین مدل در این مجموعه است و عملکرد آن نسبت به سایر مدل‌ها ضعیف‌تر است.

بنابراین، مدل‌های پیچیده‌تر مانند XGBoost و Random Forest در مقایسه با مدل‌های ساده‌تر مانند درخت تصمیم عملکرد بهتری در پیش‌بینی بیماری قلبی دارند. همچنین، مدل XGBoost به عنوان بهترین مدل برای این مجموعه داده مشخص شده است.