دانشگاه صنعتی
کرمانشاه

دانشکده مهندسی

گروه IT

درس :

داده کاوی پیشرفته

موضوع:

گزارش کار پروژه ی

رگرسیون خطی

دانشجو :

امیرحسین شهریاری قیماسی

استاد:

دکتر سجاد احمدیان

سال تحصیلی :

1403-1

1. مقدمه:

هدف این گزارش، تحلیل داده‌ها و ارزیابی عملکرد مدل‌های مختلف رگرسیون برای پیش‌بینی شاخص عملکرد دانش‌آموزان (Performance Index) است. داده‌های استفاده‌شده شامل ویژگی‌هایی نظیر ساعات مطالعه، نمرات قبلی، میزان خواب، تمرین نمونه سوالات، و شرکت در فعالیت‌های فوق‌برنامه است. این ویژگی‌ها به‌عنوان متغیرهای مستقل، تأثیر خود را بر شاخص عملکرد (متغیر وابسته) نشان می‌دهند.

در این پروژه، از مدل‌های مختلف رگرسیون شامل رگرسیون خطی، Lasso، Bayesian Ridge، Support Vector Regression (SVR)، و Decision Tree Regressor استفاده شده است. هدف اصلی این است که با تحلیل آماری و بررسی دقت هر مدل از طریق معیارهایی نظیر RMSE و R^2 و همچنین ارزیابی توزیع باقی‌مانده‌ها (Residuals)، بهترین مدل برای پیش‌بینی شاخص عملکرد تعیین شود.

در ادامه، هر یک از مدل‌ها با روش‌های استاندارد آماری نظیر Shapiro-Wilk و D'Agostino برای بررسی نرمال بودن باقی‌مانده‌ها و همچنین با معیارهای کمی برای ارزیابی دقت، تحلیل و مقایسه شده‌اند.

2. رگرسیون:

رگرسیون در آمار و یادگیری ماشین، به یک روش تحلیل داده گفته می‌شود که برای مدل‌سازی و بررسی رابطه بین یک متغیر وابسته (هدف) و یک یا چند متغیر مستقل (پیش‌بینی‌کننده) استفاده می‌شود.

رگرسیون تلاش می‌کند رابطه ریاضی بین متغیرها را بیابد و با استفاده از آن، پیش‌بینی‌هایی انجام دهد. در واقع، هدف این است که با استفاده از مقادیر متغیرهای مستقل، مقدار متغیر وابسته را تخمین بزنند.

انواع رگرسیون:

1. رگرسیون خطی (Linear Regression):

- در این روش، رابطه بین متغیرها با یک خط مستقیم مدل‌سازی می‌شود.

- معادله آن معمولاً به صورت $y = mx + c$ است که y متغیر وابسته، x متغیر مستقل، و m شیب خط و c عرض از مبدأ است.

2. رگرسیون چندگانه (Multiple Regression):

- مشابه رگرسیون خطی است، اما شامل چندین متغیر مستقل است.

- معادله آن به صورت $y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$ است.

3. رگرسیون لجستیک (Logistic Regression):

- برای پیش‌بینی متغیرهای وابسته گسسته (مانند دسته‌بندی باینری) استفاده می‌شود.

- خروجی معمولاً به صورت احتمال ارائه می‌شود.

4. رگرسیون غیرخطی (Non-Linear Regression):

- زمانی استفاده می‌شود که رابطه بین متغیرها غیرخطی باشد و به یک منحنی نیاز باشد.

5. رگرسیون چندکی (Quantile Regression):

- به جای میانگین، توزیع چندکی متغیر وابسته را مدل‌سازی می‌کند.

6. رگرسیون ریج (Ridge Regression) و لاسو (Lasso Regression):

- روش‌های منظم‌سازی برای جلوگیری از بیش‌برازش (Overfitting) هستند که در داده‌های پیچیده استفاده می‌شوند.

در این گزارش به رگرسیون خطی و غیر خطی می پردازیم.

3. رگرسیون خطی:

رگرسیون خطی یکی از ساده‌ترین و پرکاربردترین روش‌های رگرسیون است. هدف این روش، مدل‌سازی رابطه بین یک متغیر وابسته (yy) و یک یا چند متغیر مستقل (x1,x2,...,xnx1,x2,...,xn) به کمک یک خط مستقیم است.

معادله:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n + \epsilon$$

b_0 عرض از مبدأ

b_i ضرایب شیب (ضرایب رگرسیون)

ϵ خطا یا نویز مدل

فرضیات:

رابطه بین متغیرها خطی است.

نویزها مستقل و دارای توزیع نرمال هستند.

همسان‌واریانس: واریانس خطاها ثابت است.

کاربرد:

مدل‌سازی و پیش‌بینی مقادیر پیوسته.

تحلیل رابطه بین متغیرها.

3.1. روش Lasso:

LASSO (Least Absolute Shrinkage and Selection Operator) نوعی رگرسیون خطی با تکنیک منظم‌سازی (Regularization) است که ضرایب متغیرها را به کمک یک اصطلاح جریمه کوچک می‌کند. این روش به ویژه برای جلوگیری از بیش‌برازش (Overfitting) و انتخاب ویژگی مناسب است.

$$\min\{\sum_{i=1}^{n} (y_i - b0 - \sum_{j=1}^{p} b_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |b_j|\}$$

λ : پارامتر جریمه که شدت کوچک‌سازی ضرایب را کنترل می‌کند.

$|b_j|$ جریمه برای مقدار مطلق ضرایب.

ویژگی‌ها:

باعث می‌شود برخی از ضرایب به صفر تبدیل شوند (انتخاب ویژگی).

مناسب برای داده‌های با تعداد زیاد متغیر مستقل و ارتباطات پیچیده.

کاربرد:

تحلیل داده‌های دارای ویژگی‌های زیاد (مانند داده‌های ژنتیکی یا مالی).

انتخاب خودکار مهم‌ترین متغیرها.

3.2. روش Bayesian Ridge :

رگرسیون ریج بیزی یک روش منظم‌سازی بر پایه رویکرد بیزی است که مشابه رگرسیون ریج (Ridge Regression) عمل می‌کند، اما از توزیع احتمالی برای تخمین ضرایب استفاده می‌کند. در این روش، به جای ضرایب ثابت، به ضرایب یک توزیع احتمال اختصاص داده می‌شود.

$$\min\{\sum_{i=1}^{n} (y_i - b0 - \sum_{j=1}^{p} b_j x_{ij})^2 + \lambda \sum_{j=1}^{p} b_j^2\}$$

اما در روش بیزی، λ λ و b_j به‌صورت توزیع احتمال مدل‌سازی می‌شوند.

ویژگی‌ها:

توزیع‌های احتمال اولیه (Priors) و پسین (Posteriors) در تعیین ضرایب نقش دارند.

انعطاف‌پذیری بیشتری در مدل‌سازی داده‌های دارای نویز یا پیچیدگی زیاد فراهم می‌کند.

احتمال وقوع مقادیر خاصی از ضرایب را محاسبه می‌کند.

کاربرد:

داده‌هایی که دارای عدم‌قطعیت هستند یا نویز بالایی دارند.

پیش‌بینی‌هایی که نیازمند اندازه‌گیری عدم‌قطعیت هستند.

تفاوت‌های اصلی:

ویژگی‌ها	رگرسیون خطی	لاسو	ریج بیزی
منظم‌سازی	ندارد	$L1$-جریمه	$L2$-جریمه بیزی
حذف ویژگی‌ها	خیر	بله (ضرایب صفر می‌شود)	خیر
مدل‌سازی احتمالی	خیر	خیر	بله
مناسب برای داده‌های پیچیده	محدود	بله	بله

4. رگرسیون غیرخطی:

رگرسیون غیرخطی زمانی استفاده می‌شود که رابطه بین متغیرهای مستقل و وابسته غیرخطی است و نمی‌توان آن را با یک خط مستقیم مدل‌سازی کرد. در این روش، مدل ریاضی می‌تواند اشکال مختلفی مانند نمایی، لگاریتمی، چندجمله‌ای یا ترکیبی از این موارد داشته باشد.

$$y = f(x) + \epsilon$$

- $f(x)$:یک تابع غیرخطی که باید متناسب با داده‌ها انتخاب یا تخمین زده شود.

- ϵ : خطای مدل..

ویژگی‌ها:

- انعطاف‌پذیری بیشتری در مدل‌سازی داده‌های پیچیده دارد.

- پیدا کردن بهترین تابع f(x) چالش‌برانگیز است.

کاربرد:

- داده‌های بیولوژیکی، اقتصادی یا فیزیکی که روابط غیرخطی بین متغیرها دارند.

- مدل‌سازی فرآیندهایی که شامل اثرات اشباع یا آستانه‌ای هستند.

4.1. روش SVR:

رگرسیون با روش ماشین بردار پشتیبان (Support Vector Regression - SVR) یا SVR نوعی روش یادگیری ماشین است که هدف آن یافتن یک تابع با خطای پیش‌بینی کوچک و حاشیه خطای بزرگ (در محدوده ϵ) است.

مفهوم اصلی:

SVR تلاش می‌کند تا نقاط داده‌ها را در یک محدوده مشخص ϵ نگه دارد و در عین حال یک تابع خطی یا غیرخطی مناسب پیدا کند.

فرمول:

SVR از تابع زیر برای بهینه‌سازی استفاده می‌کند:

$$\min \frac{1}{2} || w ||^2 \quad \text{subject to: } | y_i - (w \cdot xi + b) | \leq \epsilon$$

- $||w||^2$: تلاش برای مینیمم کردن پیچیدگی مدل.

- ϵ : حاشیه قابل قبول خطا.

ویژگی‌ها:

- می‌تواند با استفاده از کرنل‌ها روابط غیرخطی را مدل کند (مانند کرنل RBF یا چندجمله‌ای).

- مقاوم در برابر بیش‌برازش.

کاربرد:

- پیش‌بینی داده‌های پیوسته (مانند قیمت سهام یا دما).

- مناسب برای مجموعه داده‌های کوچک اما پیچیده.

4.2. روش Decision Tree Regressor:

رگرسیون با روش درخت تصمیم (Decision Tree Regressor) یا درخت تصمیم یک روش غیرپارامتری برای مدل‌سازی رابطه بین متغیرهای مستقل و وابسته است. این روش داده‌ها را به صورت بازگشتی به زیرمجموعه‌های کوچک‌تر تقسیم می‌کند و در هر گره تصمیم می‌گیرد که چگونه مقدار متغیر وابسته پیش‌بینی شود.

فرآیند:

1. داده‌ها بر اساس یک معیار مانند میانگین مربعات خطا (MSE) تقسیم می‌شوند.

2. در هر مرحله، بهترین ویژگی و مقدار آستانه برای تقسیم انتخاب می‌شوند.

3. تقسیمات ادامه می‌یابد تا زمانی که یک شرط توقف برقرار شود (مانند عمق درخت یا تعداد نمونه‌ها در یک گره).

ویژگی‌ها:

- بسیار انعطاف‌پذیر و ساده برای تفسیر.

- مستعد بیش‌برازش است، مگر اینکه محدودیت‌هایی مانند عمق درخت یا تعداد حداقلی نمونه‌ها اعمال شود.

کاربرد:

- پیش‌بینی داده‌هایی با ساختار سلسله‌مراتبی یا روابط پیچیده.

- داده‌های حساس به ویژگی‌های خاص (مانند داده‌های طبقه‌بندی‌شده).

مقایسه SVR و Decision Tree Regressor:

ویژگی‌ها	SVR	درخت تصمیم
نوع مدل	مبتنی بر کرنل	مبتنی بر درخت
خطی/غیرخطی	هم خطی و هم غیرخطی	غیرخطی
انعطاف‌پذیری	زیاد (با انتخاب کرنل مناسب)	زیاد
بیش‌برازش	کمتر (با تنظیم C و کرنل)	مستعد بیش‌برازش
سرعت پیش‌بینی	معمولاً کندتر	سریع‌تر
تفسیرپذیری	دشوار (به‌ویژه با کرنل‌ها)	ساده و قابل فهم

انتخاب مناسب:

- رگرسیون غیرخطی: برای مدل‌هایی با روابط پیچیده بین متغیرها.

- SVR: مناسب برای داده‌های با حجم کوچک و دارای پیچیدگی غیرخطی.

- Decision Tree Regressor: مناسب برای داده‌هایی که تصمیم‌گیری‌های سلسله‌مراتبی در آن‌ها اهمیت دارد.

5. روش انجام :

در این بخش از گزارش در قسمت اول روش رگرسیون خطی و نحوه ی انجام کار ودر بخش دوم روش رگرسیون غیرخطی را بررسی می کنیم. اما قبل از آن بخش بررسی های آماری و دیتاست را بررسی می کنیم.

دیتاست:

این دیتاست که از وبسایت Kaggle بدست آمده است مربوط به عملکرد تحصیلی دانش‌آموزان است و عواملی که می‌توانند بر این عملکرد تأثیر بگذارند را بررسی می‌کند. ستون‌های موجود نشان‌دهنده اطلاعاتی درباره عادت‌های مطالعه، فعالیت‌های فوق‌برنامه، میزان خواب و تمرین نمونه سوالات توسط دانش‌آموزان است. هدف اصلی این داده‌ها، بررسی و پیش‌بینی شاخص عملکرد (Performance Index) بر اساس این عوامل است.

ساختار اطلاعات:

هر ردیف نماینده اطلاعات یک دانش‌آموز است.

ویژگی‌ها (متغیرها):

- ساعت‌های مطالعه (Hours Studied): تعداد ساعت‌هایی که هر دانش‌آموز مطالعه کرده است.
- نمرات قبلی (Previous Scores): میانگین نمرات قبلی دانش‌آموز در امتحانات یا آزمون‌ها.
- فعالیت‌های فوق‌برنامه (Extracurricular Activities): آیا دانش‌آموز در فعالیت‌های فوق‌برنامه شرکت کرده است (باینری: Yes یا No).
- ساعت‌های خواب (Sleep Hours): تعداد ساعت‌های خواب روزانه.
- تعداد نمونه سوالات تمرین شده (Sample Question Papers Practiced): تعداد نمونه سوالاتی که دانش‌آموز برای تمرین استفاده کرده است.
- شاخص عملکرد (Performance Index): خروجی یا هدف این مجموعه داده که عملکرد دانش‌آموز را بر اساس معیارهای مشخص اندازه‌گیری می‌کند (اعداد پیوسته).

نمونه ی دیتا

	Hours Studied	Previous Scores	Extracurricular Activities	Sleep Hours	Sample Question Papers Practiced	Performance Index
8821	5	46	No	6	2	32
7739	7	77	No	5	6	67
5260	4	51	Yes	9	5	36
1741	7	65	Yes	7	0	53

8366	4	98	No	8	5	82
8812	9	93	Yes	6	7	93
653	3	95	Yes	5	1	74
2429	1	68	No	6	4	39
1472	2	85	Yes	9	1	61
418	7	81	No	6	0	70
4973	2	94	No	6	0	70
7769	6	57	Yes	9	8	45
7722	9	91	Yes	5	2	88
8435	2	52	No	6	0	28
1762	4	91	No	9	7	74

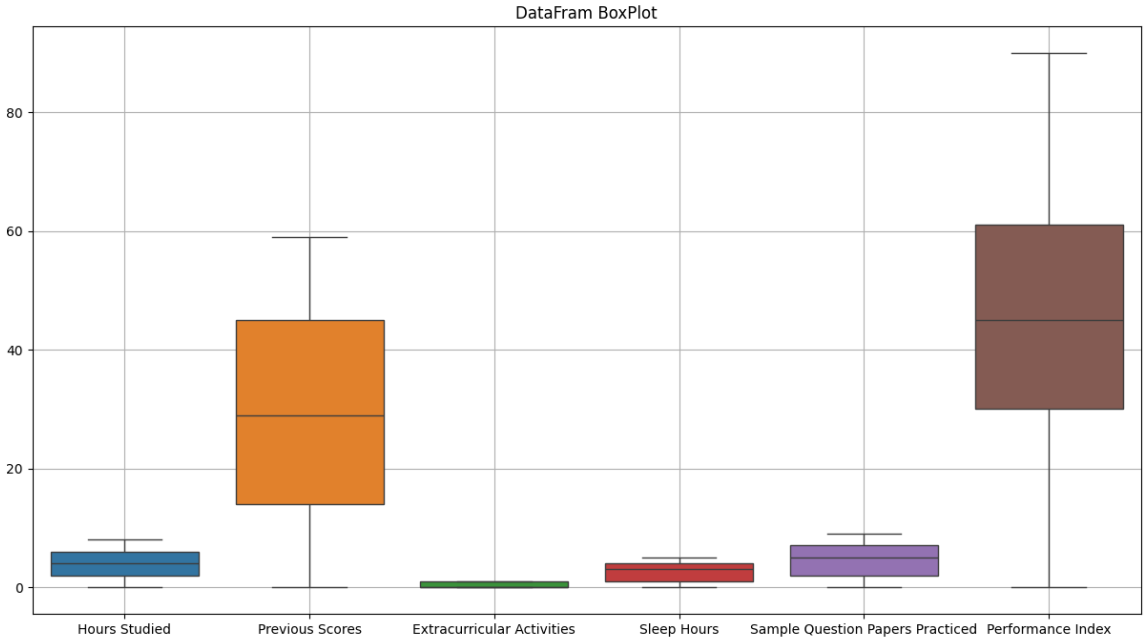
بررسی ها آماری

در این بخش چند تست آماری انجام شده را ارائه می دهیم

	Hours Studied	Previous Scores	Sleep Hours	Sample Question Papers Practiced	Performance Index
count	10000	10000	10000	10000	10000
mean	4/9929	69/4457	6/5306	4/5833	55/2248
std	2/589308796	17/34315225	1/695862977	2/867347778	19/2125578
min	1	40	4	0	10
25%	3	54	5	2	40
50%	5	69	7	5	55
75%	7	85	8	7	71
max	9	99	9	9	100

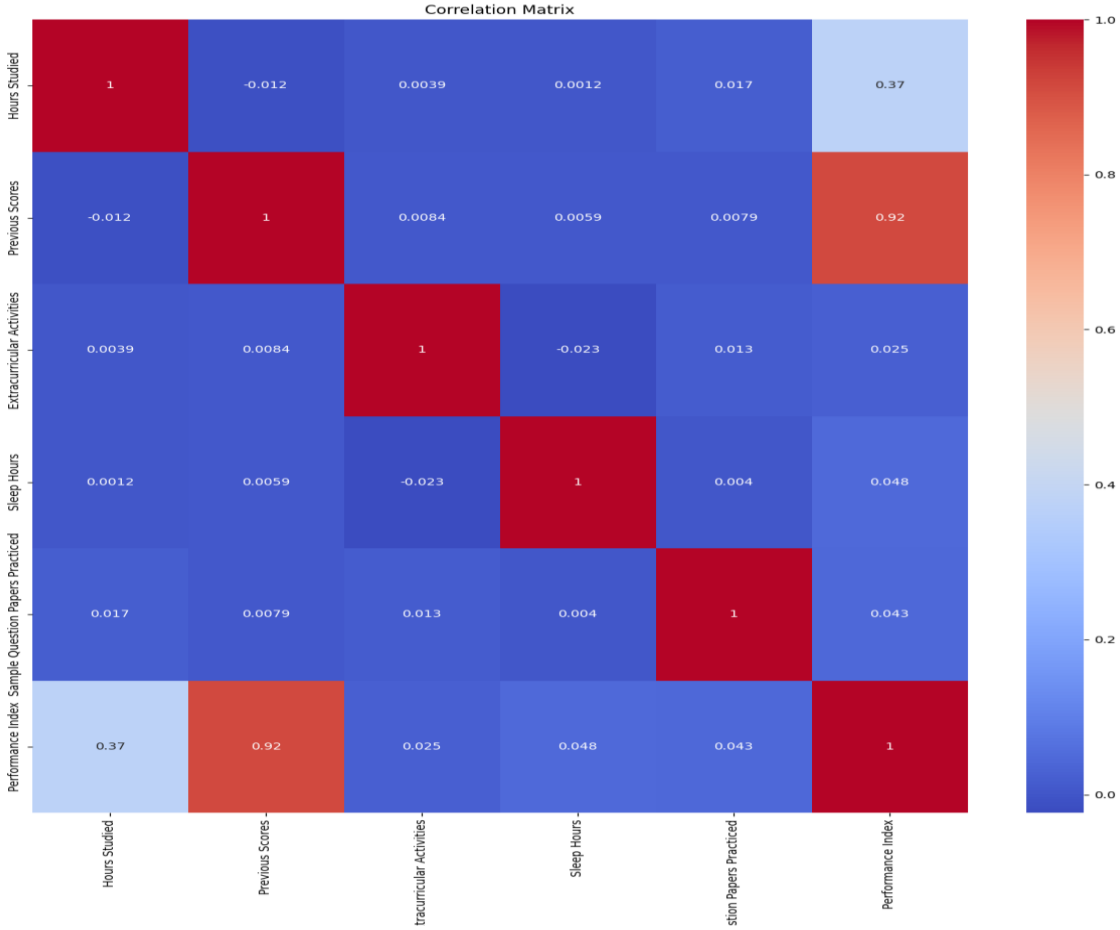
نمودار BoxPlot

یک نمایش گرافیکی از توزیع داده‌ها که اطلاعاتی درباره میانه، چارک‌ها، دامنه بین‌چارکی و نقاط پرت ارائه می‌دهد.

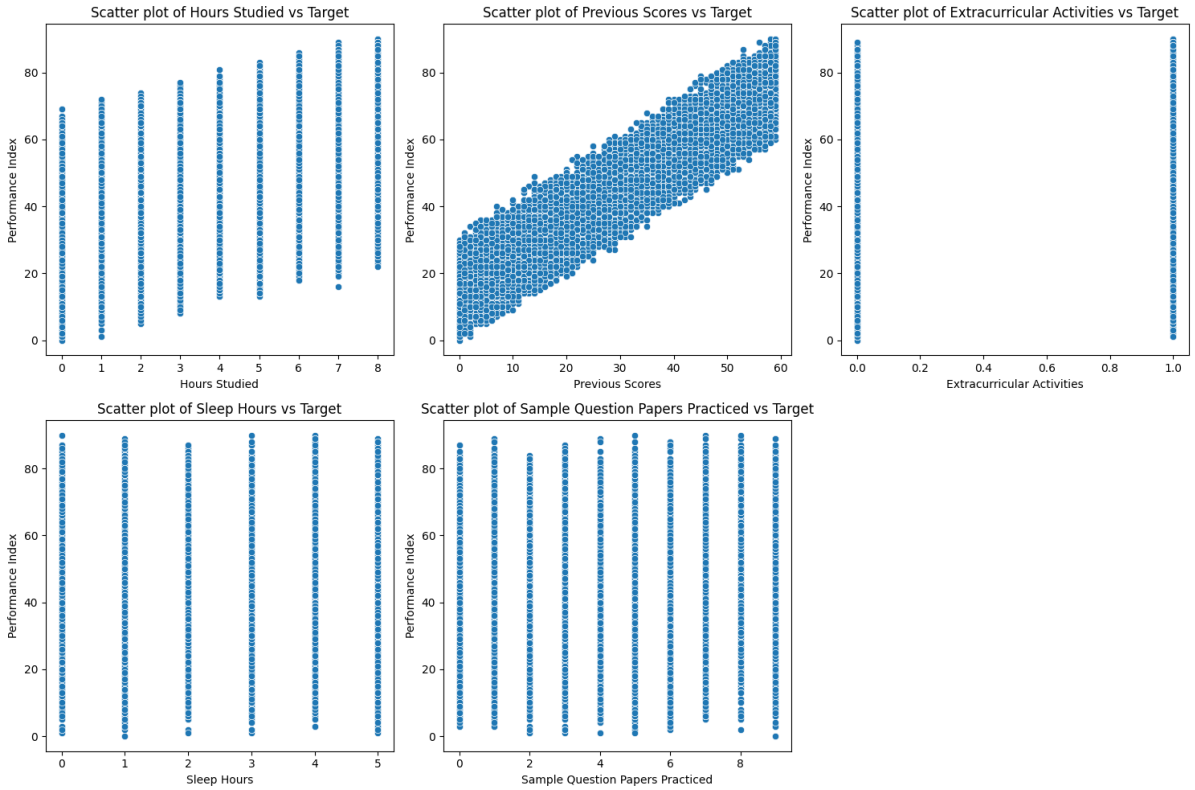


ماتریس همبستگی

یک جدول عددی که میزان تغییرات مشترک (همبستگی) بین جفت متغیرها را نشان می‌دهد.



نمودار همبستگی:



تست Variance Inflation Factor (VIF)

VIF یکی از معیارهای آماری برای تشخیص چندخطی بودن (Multicollinearity) بین متغیرهای مستقل در یک مدل رگرسیون است. چندخطی بودن زمانی رخ می‌دهد که متغیرهای مستقل به شدت با یکدیگر همبستگی داشته باشند، که می‌تواند باعث ناپایداری ضرایب رگرسیون و کاهش دقت مدل شود.

VIF به ما می‌گوید که چه مقدار از تغییرات یک متغیر مستقل خاص توسط سایر متغیرهای مستقل توضیح داده می‌شود.

Feature	VIF
Hours Studied	2/714584
Previous Scores	2/954507
Extracurricular Activities	1/819782
Sleep Hours	2/620819
Sample Question Papers Practiced	2/848774

مفهوم مقادیر VIF:

- $VIF = 1$: متغیر کاملاً مستقل است (بدون همبستگی با سایر متغیرها).

- $1 < VIF < 5$: میزان همبستگی قابل قبول است.

- $VIF > 5$: همبستگی بالاست و باید بررسی شود.

- $VIF > 10$: نشان‌دهنده وجود مشکل جدی چندخطی بودن است و ممکن است نیاز به حذف یا ترکیب متغیرها باشد.

چرا VIF مهم است؟

- ناپایداری ضرایب: چندخطی بودن باعث می‌شود ضرایب تخمینی بسیار حساس به تغییرات جزئی در داده‌ها باشند.

- تفسیر اشتباه: ضرایب ممکن است علامت اشتباه داشته باشند یا معنادار نشوند.

- دقت مدل: باعث کاهش توان پیش‌بینی مدل می‌شود.

5.1. رگرسیون خطی

تفاسیر به روش های زیر خواهد بود :

نمودار Residplot :

نموداری است که برای تجزیه و تحلیل خطاهای باقی‌مانده (Residuals) در یک مدل رگرسیون استفاده می‌شود. این نمودار نقاطی را نمایش می‌دهد که هر نقطه نشان‌دهنده اختلاف بین مقادیر پیش‌بینی‌شده و مقادیر واقعی است.

نمودار qqplot :

Quantile-Quantile Plot یک نمودار آماری است که برای مقایسه توزیع یک مجموعه داده با یک توزیع مرجع (معمولاً نرمال) استفاده می‌شود. این نمودار نقاطی را نشان می‌دهد که هر نقطه نمایانگر یک کوانتیل از داده واقعی در مقابل کوانتیل متناظر از توزیع مرجع است.

تست Shapiro-Wilk :

این تست بررسی می‌کند که آیا داده‌ها به طور قابل ملاحظه‌ای از یک توزیع نرمال منحرف شده‌اند یا خیر. این تست به‌ویژه برای داده‌های با اندازه کوچک تا متوسط مناسب است. مقدار p-value حاصل نشان می‌دهد

که آیا فرض نرمال بودن رد می‌شود یا پذیرفته می‌شود (معمولاً اگر $p<0.05$، فرض نرمال بودن رد می‌شود).

تست D'Agostino :

این تست نیز نرمال بودن داده‌ها را بررسی می‌کند، اما تمرکزش بر اساس محاسبه چولگی (Skewness) و کشیدگی (Kurtosis) داده‌هاست. این تست بیشتر برای داده‌های با اندازه متوسط تا بزرگ مناسب است و از آماره K^2 استفاده می‌کند.

RMSE (Root Mean Squared Error):

معیاری برای اندازه‌گیری میزان خطای مدل در پیش‌بینی است. این مقدار میانگین خطاهای مربعی را به دست می‌آورد و سپس ریشه دوم آن را می‌گیرد. مقدار کمتر RMSE نشان‌دهنده دقت بیشتر مدل است. RMSE همیشه در واحد متغیر هدف (y) بیان می‌شود و حساس به مقادیر پرت است.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

R^2 (Coefficient of Determination):

نشان‌دهنده درصد واریانس متغیر وابسته (yy) است که توسط مدل توضیح داده شده است. مقدار R2R2 بین 0 و 1 است؛ هرچه به 1 نزدیک‌تر باشد، نشان‌دهنده عملکرد بهتر مدل است. مقدار 0 به معنای این است که مدل هیچ توضیحی برای واریانس داده‌ها ندارد، و مقدار 1 به معنای توضیح کامل واریانس است.

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}}$$

5.1.1. روش lasso

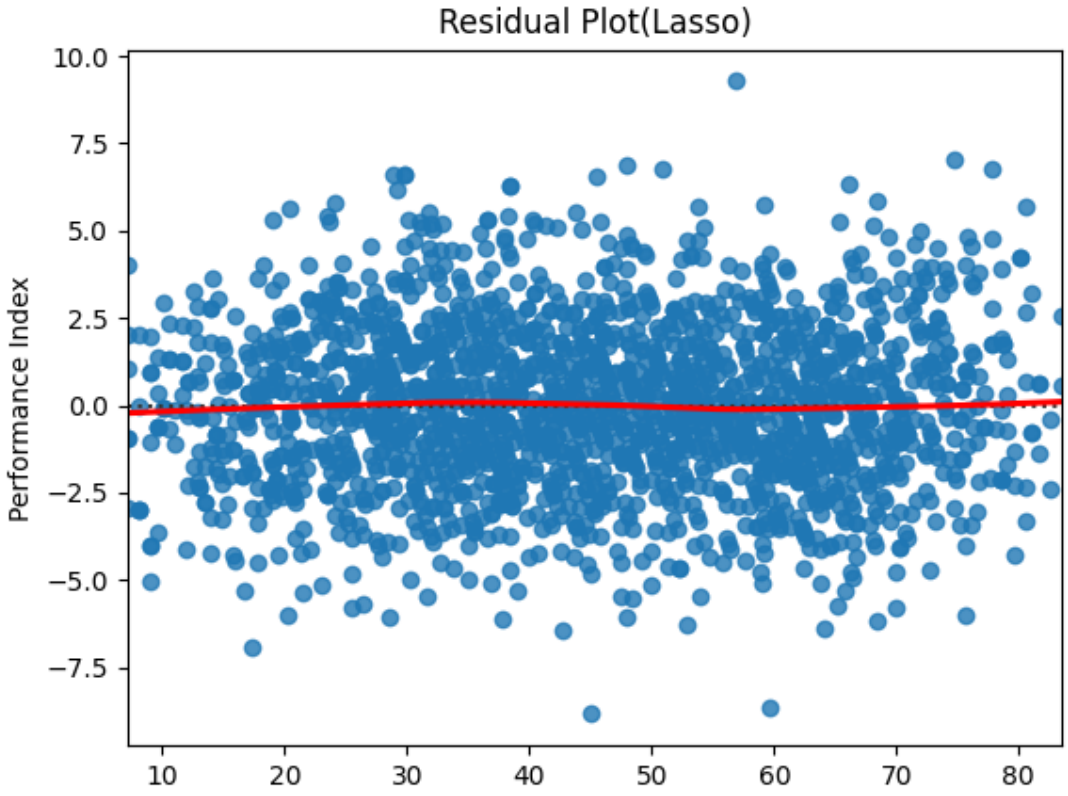
نمودار Residplot:

پراکندگی تصادفی نقاط: مدل مناسب است و فرض خطی بودن برقرار است.

الگوی واضح یا منحنی شکل در نقاط: مدل خطی مناسب نیست و ممکن است به یک مدل غیرخطی نیاز باشد.

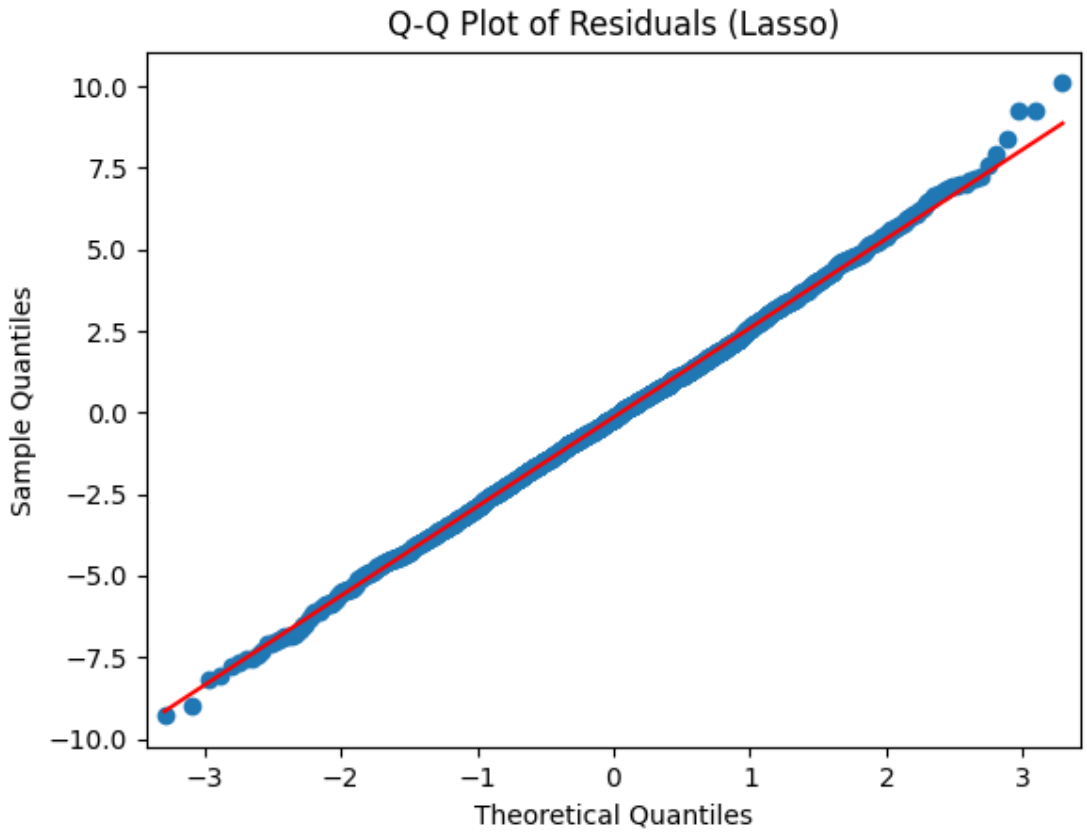
پراکندگی تغییرپذیر (Fan Shape): نشانه‌ای از واریانس نابرابر (Heteroscedasticity) در داده‌ها.

نقاط پرت: نشان‌دهنده داده‌های غیرعادی که ممکن است نیاز به بررسی بیشتر داشته باشند.



نمودار qqplot

این نمودار نشان می‌دهد که کوانتیل‌های باقی‌مانده‌ها (Residuals) با کوانتیل‌های توزیع نرمال تئوری مقایسه شده‌اند. بیشتر نقاط به صورت نزدیک به خط قرمز (خط مرجع) قرار دارند، که نشان می‌دهد باقی‌مانده‌ها تقریباً از توزیع نرمال پیروی می‌کنند. در انتهای نمودار (نواحی کوانتیل‌های بسیار کوچک و بزرگ)، چند نقطه انحراف اندکی از خط دارند، که می‌تواند نشان‌دهنده وجود مقادیر پرت باشد. به طور کلی، این نمودار تأیید می‌کند که فرض نرمال بودن باقی‌مانده‌ها به خوبی برقرار است و مدل از این لحاظ معتبر است.



تست Shapiro-Wilk

Shapiro-Wilk Test p-value: 0.53

مقدار p بالاتر از 0.05 است، به این معنی که نمی‌توان فرض نرمال بودن توزیع داده‌ها را رد کرد.

بنابراین، داده‌ها از نظر تست شاپیرو-ویلک می‌توانند به عنوان داده‌های نرمال در نظر گرفته شوند.

تست D'Agostino

D'Agostino Test p-value: 0.24

مشابه تست قبلی، مقدار p بزرگ‌تر از 0.05 است، که نشان می‌دهد این تست نیز فرض نرمال بودن داده‌ها را رد نمی‌کند.

بر اساس این نتیجه، داده‌ها با نرمال بودن سازگار هستند.

نتیجه‌گیری کلی:

هر دو تست نشان می‌دهند که داده‌ها از یک توزیع نرمال پیروی می‌کنند. بنابراین، می‌توان از فرض نرمال بودن برای این داده‌ها در تحلیل‌های آماری بعدی استفاده کرد.

مقدار RMSE :

Root Mean Squared Error: 2.74

این عدد در واحد متغیر وابسته (Target) است و نشان‌دهنده خطای متوسط مدل است. مقدار کوچک 2.74 نشان‌دهنده دقت بالای مدل در پیش‌بینی‌ها است.

مقدار R^2 :

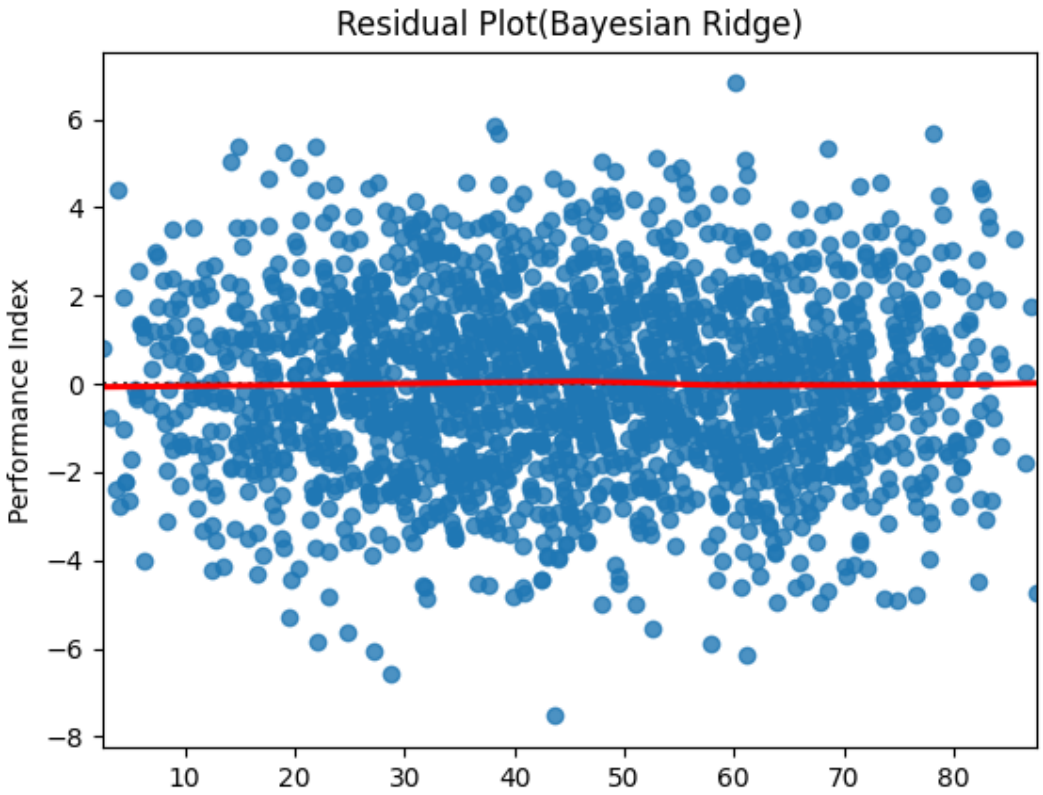
R^2 Score: 0.9797

این مقدار بسیار نزدیک به 1 است، که نشان‌دهنده این است که مدل عملکرد بسیار خوبی در توضیح داده‌ها دارد.

5.1.2. روش Bayesian Ridge

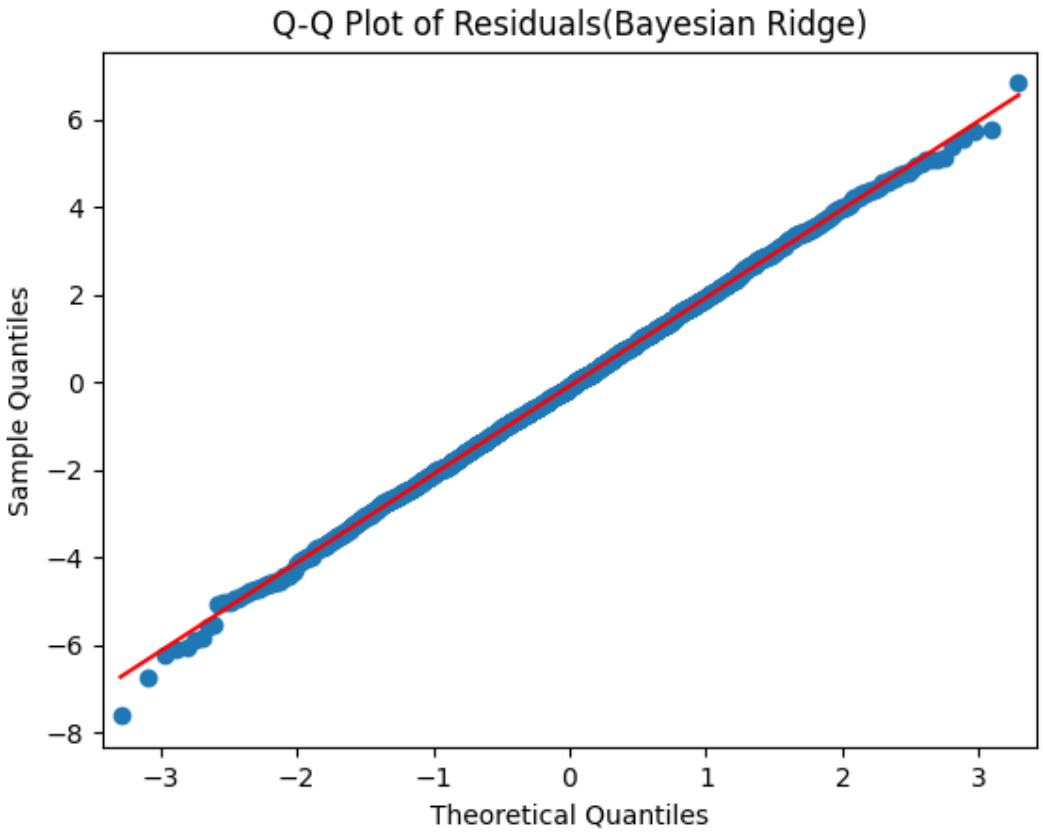
نمودار Residplot:

نمودار Residplot نشان‌دهنده پراکندگی تصادفی باقی‌مانده‌ها در اطراف خط افقی صفر است. این پراکندگی تصادفی نشان می‌دهد که فرض‌های همسان واریانس و خطی بودن مدل رعایت شده‌اند و مدل به خوبی داده‌ها را پیش‌بینی می‌کند.



نمودار qqplot :

نمودار QQPlot نشان می‌دهد که باقی‌مانده‌ها تقریباً به طور کامل بر روی خط مرجع قرار گرفته‌اند، که این امر بیانگر این است که توزیع باقی‌مانده‌ها با توزیع نرمال سازگار است. انحراف‌های جزئی در انتهای نمودار قابل مشاهده است، اما به قدری کم هستند که تأثیر قابل توجهی بر عملکرد مدل ندارند.



تست Shapiro-Wilk :

Shapiro-Wilk Test p-value: 0.98

مشابه تست شاپیرو-ویلک، این مقدار نیز بزرگ‌تر از 0.05 است، بنابراین فرض نرمال بودن باقی‌مانده‌ها تأیید می‌شود.

تست D'Agostino :

D'Agostino Test p-value: 0.96

مشابه تست شاپیرو-ویلک، این مقدار نیز بزرگ‌تر از 0.05 است، بنابراین فرض نرمال بودن باقی‌مانده‌ها تأیید می‌شود.

مقدار RMSE :

Root Mean Squared Error: 2.02

خطای متوسط پیش‌بینی این مدل بسیار کوچک است، که نشان‌دهنده دقت بالای مدل در پیش‌بینی مقادیر هدف است.

مقدار R^2 :

R^2 Score: 0.9890

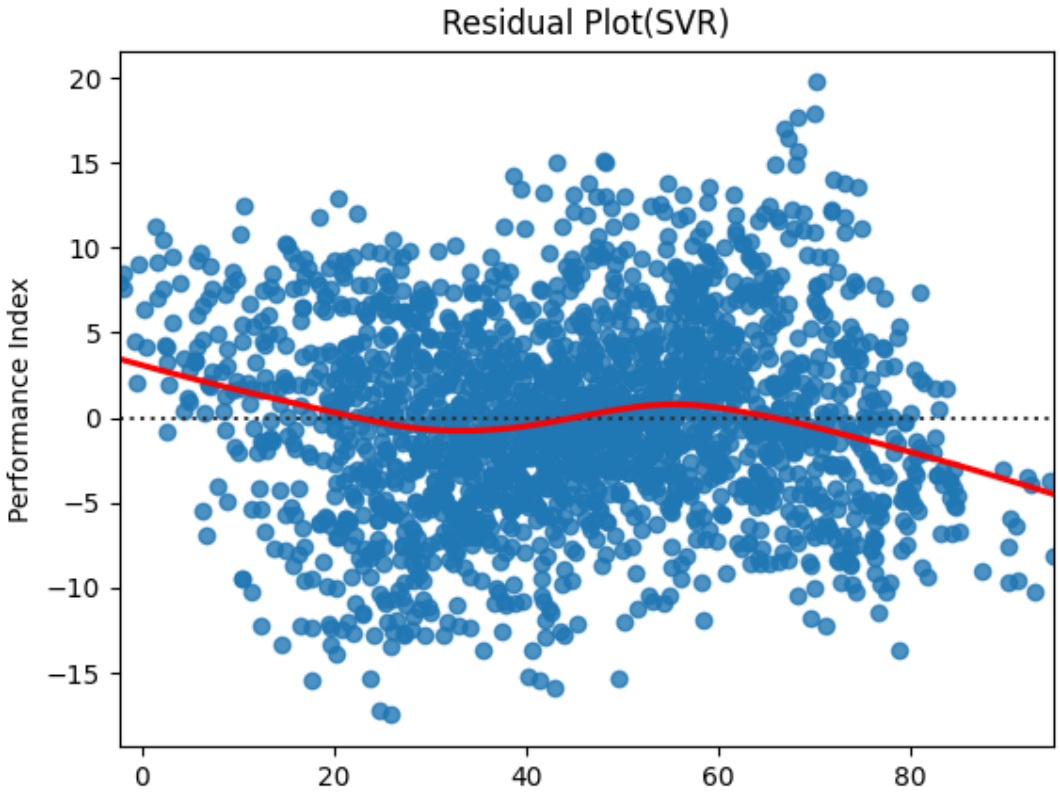
مقدار بسیار نزدیک به 1 است و نشان می‌دهد که 98.9٪ از واریانس متغیر هدف توسط مدل توضیح داده می‌شود، که عملکرد مدل را بسیار خوب نشان می‌دهد.

5.2. رگرسیون غیر خطی

5.2.1. روش SVR

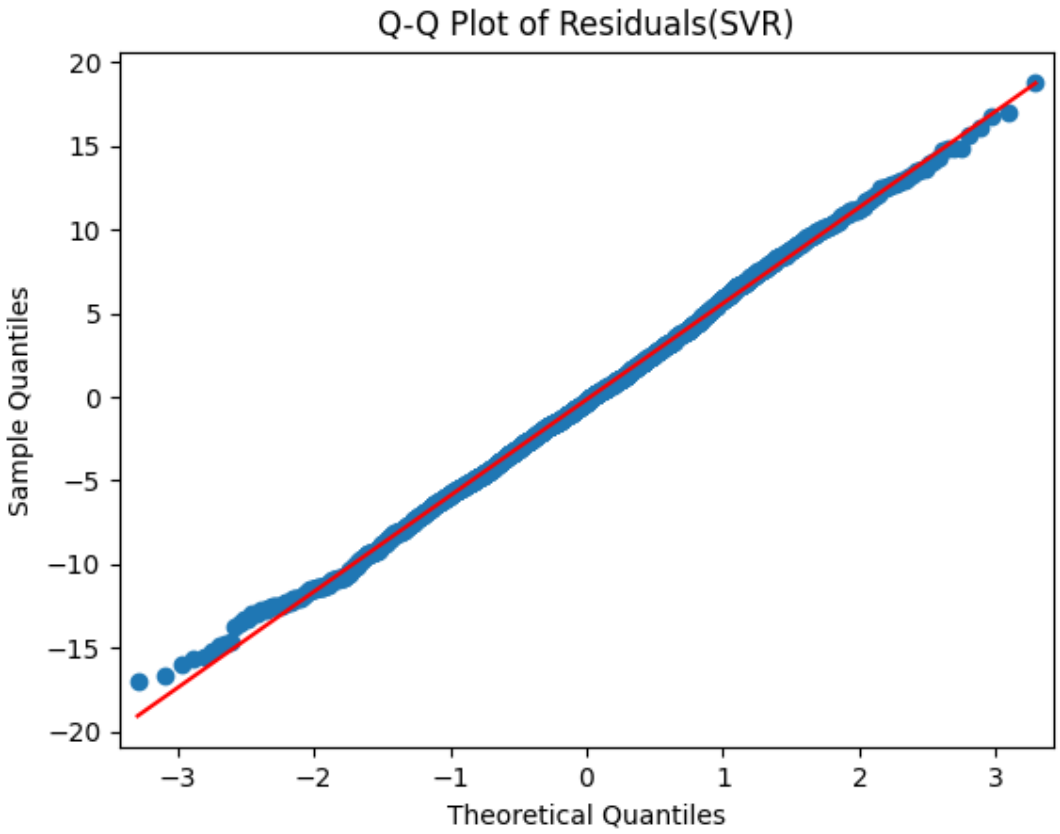
نمودار Residplot:

در این مدل، نمودار Residplot الگوهایی نشان می‌دهد، به‌خصوص در مقادیر انتهایی. این الگوها بیانگر این است که مدل در توضیح داده‌ها در برخی بازه‌ها دچار مشکل شده است و ممکن است مدل بهینه‌سازی بیشتری نیاز داشته باشد.



نمودار qqplot :

در این مدل، نقاط در بخش‌های ابتدایی و انتهایی نمودار از خط مرجع فاصله گرفته‌اند، که نشان‌دهنده انحراف از توزیع نرمال در کوانتیل‌های بسیار کوچک و بزرگ است. این امر ممکن است به دلیل خطای بیشتر مدل در پیش‌بینی مقادیر خارج از محدوده باشد.



تست Shapiro-Wilk :

Shapiro-Wilk Test p-value: 0.05

مقدار p دقیقاً برابر با 0.05 است، که نشان می‌دهد داده‌ها در مرز پذیرش فرض نرمال بودن قرار دارند.

تست D'Agostino :

D'Agostino Test p-value: 0.06

مقدار p کمی بزرگ‌تر از 0.05 است، که نشان می‌دهد نرمال بودن باقی‌مانده‌ها با تقریب قابل قبولی تأیید می‌شود.

مقدار RMSE :

Root Mean Squared Error: 5.75

مقدار RMSE در این مدل نسبتاً بزرگ است، که نشان‌دهنده خطای پیش‌بینی بیشتر مدل SVR است.

مقدار R^2 :

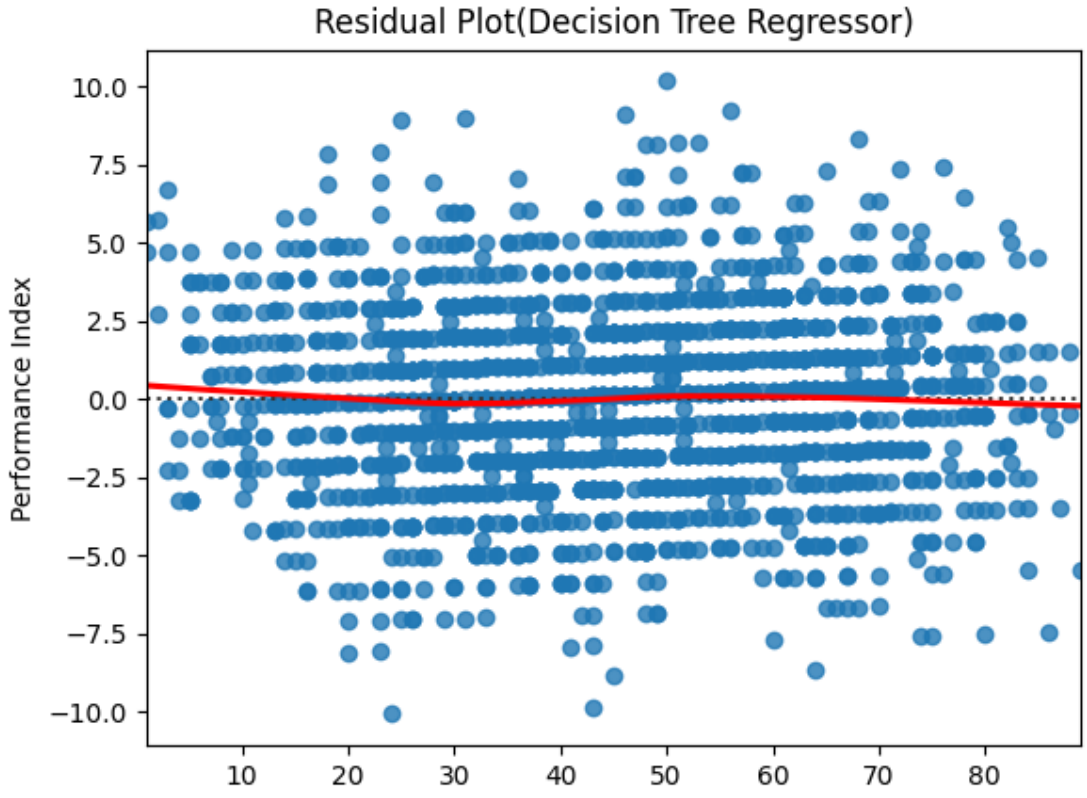
R^2 Score: 0.9109

این مقدار نشان می‌دهد که مدل 91.09٪ از واریانس متغیر هدف را توضیح می‌دهد، که عملکرد مناسبی است اما به خوبی Bayesian Ridge نیست.

5.2.2. روش Decision Tree Regressor

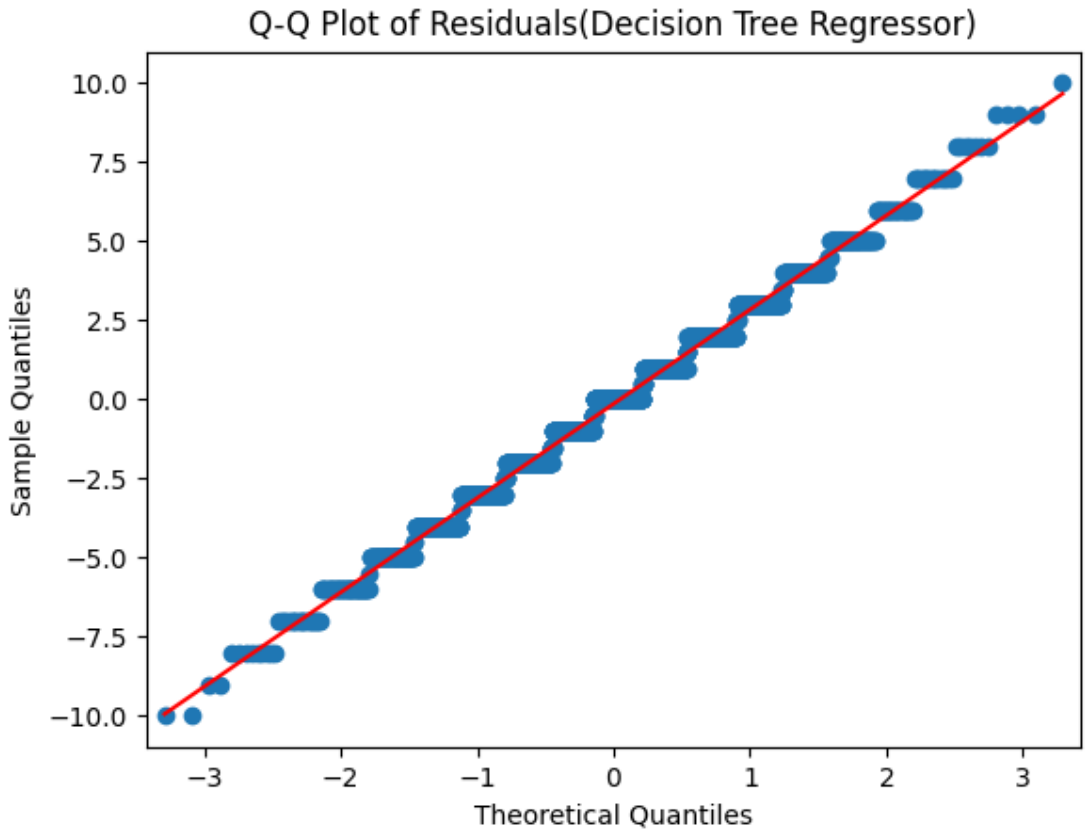
نمودار Residplot:

نمودار Residplot برای این مدل پراکندگی غیرتصادفی و الگوهای مشخصی را نشان می‌دهد، که بیانگر آن است که مدل خطاهایی وابسته به مقدار پیش‌بینی‌شده دارد. این رفتار معمولاً به دلیل ماهیت مدل‌های درختی در ایجاد مرزهای پیش‌بینی گسسته و غیرخطی است.



نمودار qqplot :

نقاط در نمودار QQPlot به طور قابل ملاحظه‌ای از خط مرجع فاصله دارند، که نشان می‌دهد باقی‌مانده‌ها از توزیع نرمال تبعیت نمی‌کنند. این نتیجه قابل پیش‌بینی است، زیرا مدل‌های مبتنی بر درخت معمولاً خروجی‌های گسسته یا غیرخطی تولید می‌کنند.



تست Shapiro-Wilk :

Shapiro-Wilk Test p-value: 0.00

مقدار p بسیار کوچک است، که نشان می‌دهد باقی‌مانده‌ها از توزیع نرمال پیروی نمی‌کنند.

تست D'Agostino :

D'Agostino Test p-value: 1.00

مقدار p بسیار بزرگ است، که نشان می‌دهد داده‌ها از نظر چولگی و کشیدگی با توزیع نرمال هم‌خوانی دارند.
این تناقض نشان می‌دهد ممکن است داده‌ها دارای ویژگی‌های خاصی باشند.

مقدار RMSE :

Root Mean Squared Error: 2.98

خطای پیش‌بینی این مدل کوچک است، اما نسبت به Bayesian Ridge دقت کمتری دارد.

مقدار R^2 :

R^2 Score: 0.9760
این مقدار نشان می‌دهد که مدل 97.6٪ از واریانس متغیر هدف را توضیح می‌دهد، که عملکرد بسیار خوبی است.

6. خلاصه نتیجه گیری

نتیجه کلی از نمودارها:

مدل Bayesian Ridge Regression بهترین عملکرد را دارد، زیرا هم در QQPlot و هم در Residplot رفتارهای تصادفی و توزیع نرمال باقی‌مانده‌ها را نشان می‌دهد. مدل SVR در مقادیر انتهایی مشکل دارد، و Decision Tree Regressor به دلیل ساختار ذاتی خود، الگوهای غیرتصادفی در باقی‌مانده‌ها نشان می‌دهد که می‌تواند در داده‌های پیچیده‌تر مشکل ایجاد کند.

- مدل Bayesian Ridge Regression با کمترین مقدار RMSE (2.02) و بیشترین مقدار R^2 (0.9890)، بهترین عملکرد را در بین مدل‌ها نشان می‌دهد. همچنین باقی‌مانده‌های آن نرمال هستند، که نشان‌دهنده مدل‌سازی دقیق و معتبر است.

- مدل SVR عملکرد خوبی دارد، اما RMSE بالاتر (5.75) و R^2 کمتر (0.9109) آن نشان می‌دهد که این مدل به خوبی Bayesian Ridge عمل نمی‌کند.

- مدل Decision Tree Regressor با R^2 = 0.9760 عملکرد خوبی در توضیح واریانس داده‌ها دارد، اما نرمال نبودن باقی‌مانده‌ها و مقدار کمی بالاتر RMSE (2.98) نشان می‌دهد که در پیش‌بینی ممکن است خطای بیشتری داشته باشد.

نتیجه کلی: برای این پروژه، مدل Bayesian Ridge Regression به دلیل دقت بالا، نرمال بودن باقی‌مانده‌ها و عملکرد بهتر در پیش‌بینی، مناسب‌ترین مدل است.