به نام خدا

1928
K. N. Toosi University of Technology

دانشگاه صنعتی خواجه نصیرالدین طوسی

دانشکده برق

شناسایی سیستم

گزارش تمرین شماره ۱

[علیرضا یاحقی]

[۴۰۰۱۰۴۱۳]

استاد : آقای دکتر مهدی علیاری

اردیبهشت ۱۴۰۴

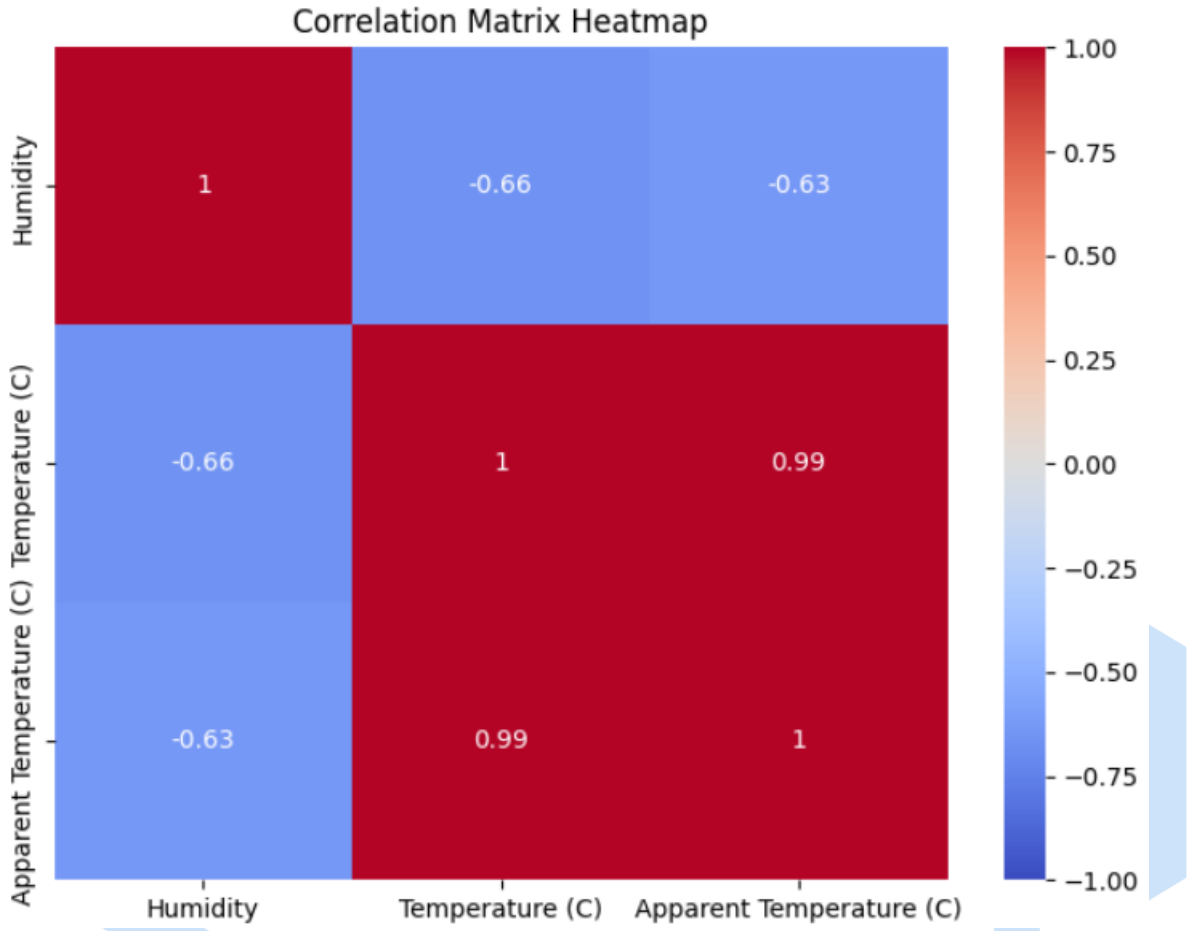
فهرست مطالب

سوال ۱

یک دیتاست در زمینهٔ آب و هوا با نام Szeged in Weather ۲۰۰۶-۲۰۱۶ را در نظر بگیرید. در این دیتاست هدف آن است که ارتباط بین Humidity با Temperature و هم چنین ارتباط بین Humidity و Temperature Apparent پیدا شده و با کمک داده های Humidity و Temperature تخمین انجام شود.

۱. ابتدا هیت مپ ماتریس همبستگی و هیستوگرام پراکندگی ویژگی ها را رسم و تحلیل کنید.
۲. روی این دیتاست، تخمین LS و RLS را با تنظیم پارامترهای مناسب عمال کنید. نتایج به دست آمده را با محاسبهٔ خطاها و رسم نمودارهای مناسب برای هر دو مدل با هم مقایسه و تحلیل کنید.
۳. در مورد Square Least Weighted توضیح دهید و آن را روی دیتاست داده شده عمال کنید.

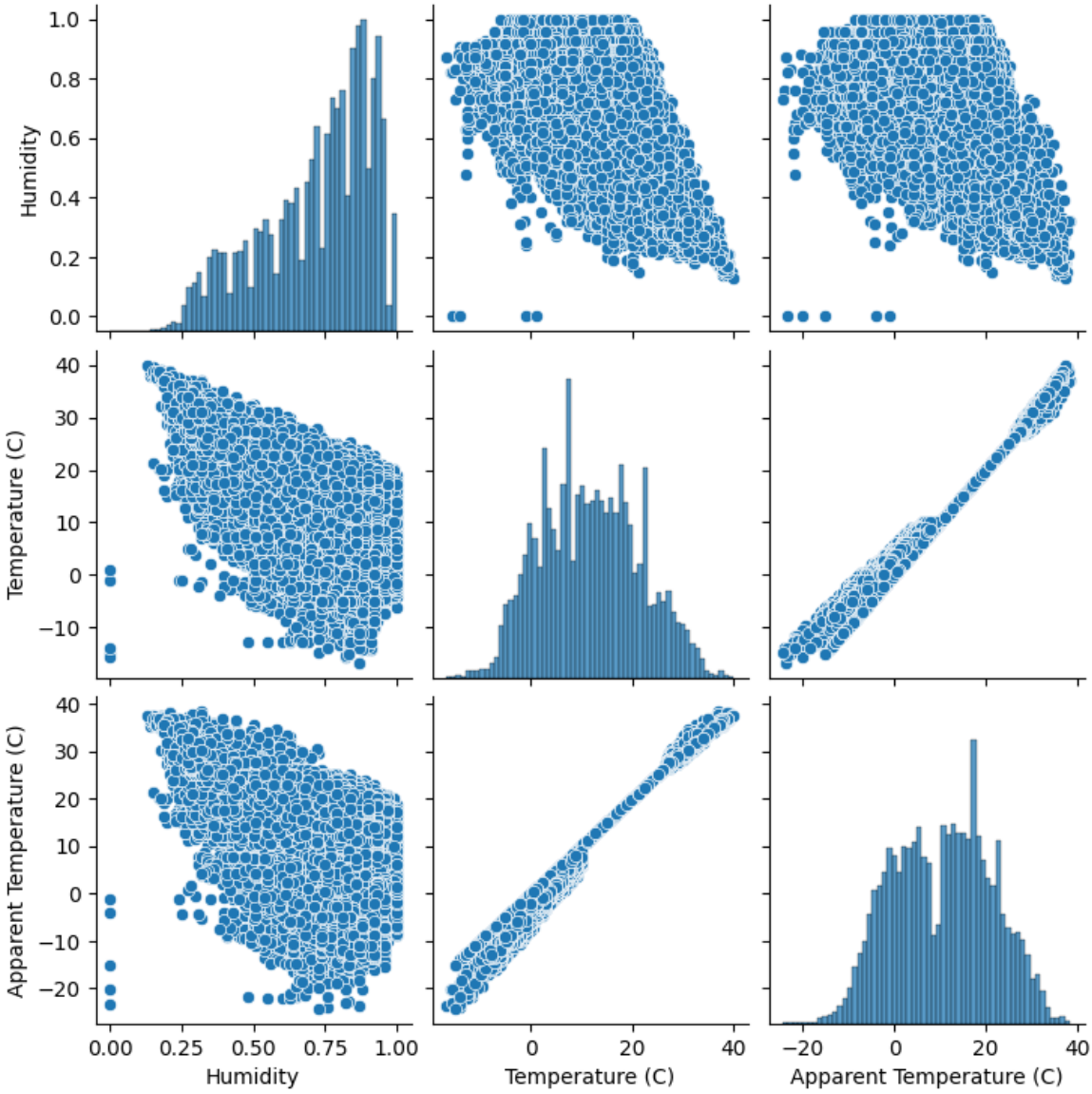
پاسخ بخش ۱



شکل ۱ هیت مپ ماتریس هم بستگی

انتظار می‌رفت که Humidity با Temperature و Apparent Temperature همبستگی منفی داشته باشد، زیرا معمولاً با افزایش دما، رطوبت نسبی کاهش می‌یابد که باوجه به هیت مپ این اتفاق افتاد.

همبستگی بین Temperature و Apparent Temperature بدرستی نزدیک یک و تریبا یک می باشد، زیرا Apparent Temperatureمعمولاً تابعی از Temperature و رطوبت است.



همانگونه که مشاهده می کنیم دما و دمای محسوس تقریبا توزیع نرمال دارند. نکته مهم دیگری که میشه استخراج کرد این هست Temperature (C) و Apparent Temperature (C)
به وضوح دارای رابطهٔ خطی بسیار قوی و مثبت هستند(نزدیک به خط مستقیم با شیب مثبت).

هر دوی این متغیرها رابطه منفی نسبتاً واضحی با Humidityدارند.

اگر هدف مدلسازی رطوبت (Humidity) با استفاده از ویژگی‌های دیگر است، بهتر است یکی از دو متغیر Temperatureیا Apparent Temperature را حذف کنیم تا از چندخطی بودن جلوگیری شود.

راهکار هایی که داریم عبارتند از:

۱. حذف یکی از دو متغیر (معمولاً ساده‌ترین و مؤثرترین راه)

۲. ترکیب این دو ویژگی (مثلاً گرفتن میانگین یا استفاده از PCA برای کاهش ابعاد)

۳. استفاده از (RLS) یا WLS که نسبت به چند خطی بودن مقاوم‌تر هستند.

کاری که انجام دادم این بود یکبار با PCA، LS زدم و یکبار بدون PCA و نتایج را با هم مقایسه کردم.

--- LS with PCA (Test Set) ---

RMSE: 0.1544458868659783

MAE : 0.124490377614216

R2 : 0.3783888617362897

--- LS without PCA (Test Set) ---

RMSE: 0.1464905139362151

MAE : 0.11707451906625574

R2 : 0.4407769042906242

مدل LS بدون PCA عملکرد بهتری دارد.

تمام معیارهای خطا در این مدل بهتر است (کمتر یا بالاتر در مورد R^2)، چون اطلاعات کامل‌تری از دو ویژگی اصلی در دسترس دارد. کاهش بعد با PCA موجب کاهش دقت شده است.

PCA فقط مؤلفه‌ای از بیشترین واریانس را حفظ می‌کند، اما ممکن است تمام اطلاعات مؤثر در پیش‌بینی را منتقل نکند. با این حال، مدل LS با PCA هنوز نتایج قابل قبولی دارد و مزیت آن سادگی (فقط ۱ ویژگی) و حذف چندخطی بودن است.

اما چون هدف ما فقط پیش‌بینی است و چندخطی بودن مشکلی ایجاد نمی‌کند، مدل بدون PCA را انتخاب می کنیم.

پاسخ بخش ۲

مدلسازی با روش LS:

ابتدا داده ها را نرمالسازی می کنیم و سپس یک ستون ۱ به عنوان بایاس به دیتا اضافه می کنیم و سپس دیتا را به دو بخش آموزش(۷۰٪) و تست (۳۰٪) تقسیم می کنیم و LS می زنیم.

نتایج:

(theta) from LS: [0.7345964 -0.44860222 0.32697116]

بایاس ($\theta 0$)= ۰.۷۳۴۶ : نشان دهنده مقدار پایه رطوبت پیش بینی شده است.

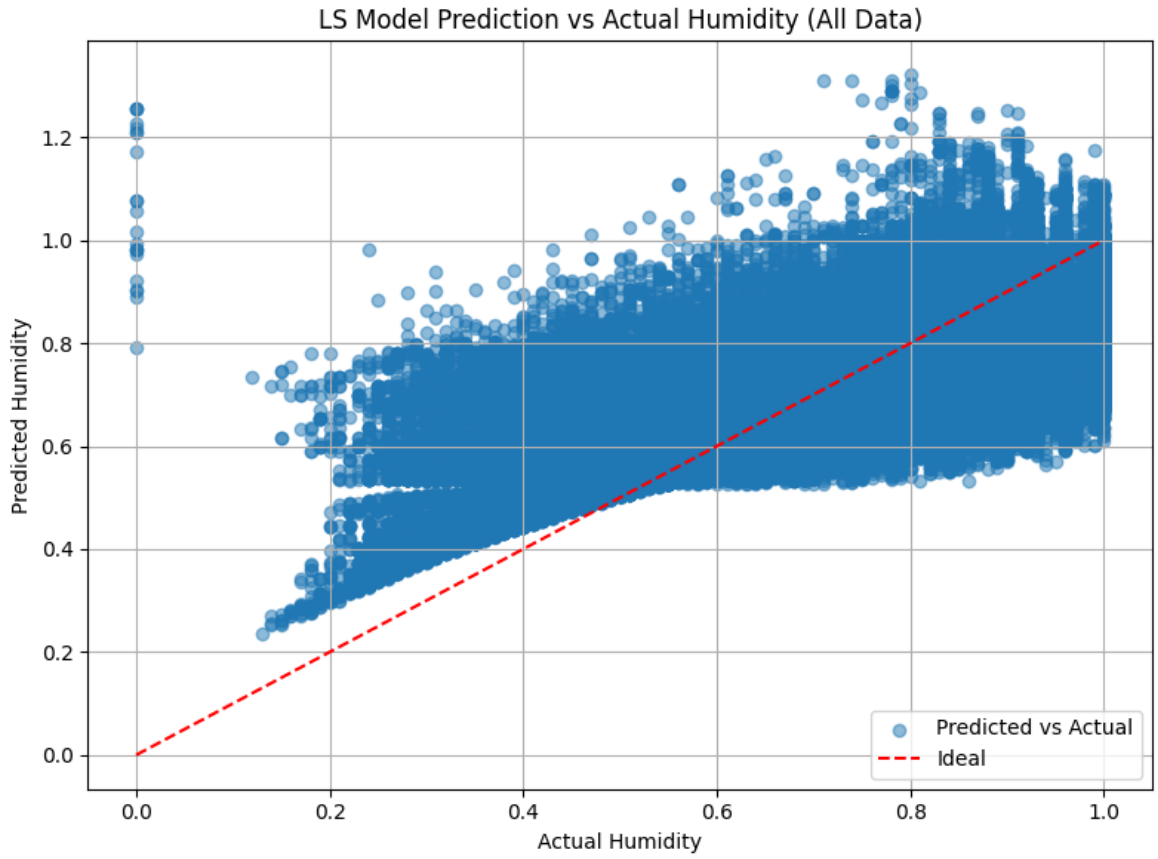
$\theta 1$ =-0.4486 : نشان دهنده تأثیر منفی دما روی رطوبت است، که منطقی است زیرا افزایش دما معمولاً رطوبت نسبی را کاهش می دهد.

$\theta 2$ =0.3270 : نشان دهنده تأثیر مثبت دمای محسوس است، که ممکن است به دلیل تأثیر رطوبت و باد در دمای محسوس باشد.

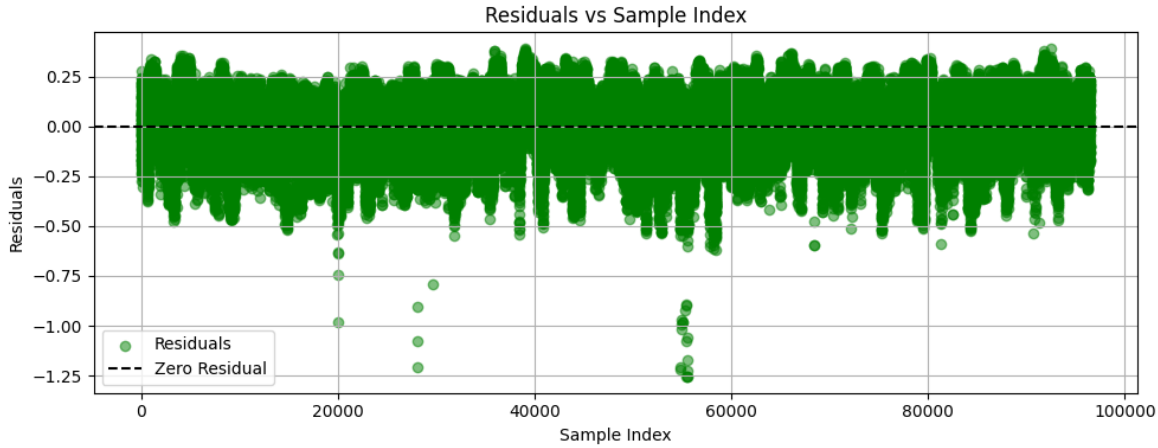
ضرایب به نظر معقول هستند، اما علامت های مخالف بین $\theta 1$ و $\theta 2$ ممکن است به هم خطی بین Temperature (C) و Apparent Temperature(C) اشاره کند.

	Train	Test	All Data
MSE	0.0213	0.0214	0.0213
MAE	0.1165	0.1170	0.1166
R2	0.4444	0.4374	0.4424

- اختلاف کم بین Train و Test نشان‌دهنده عدم بیش‌برازش است.
- مدل LS عملکرد متوسطی دارد (R ~ 0.442). خطاها (MSE و MAE) پایین هستند، اما R۲ نشان می دهد که بخش قابل توجهی از واریانس دادهها (حدود ۵۶٪) توضیح داده نشده است. این ممکن است به دلیل نویز یا رابطه غیر خطی باشد.



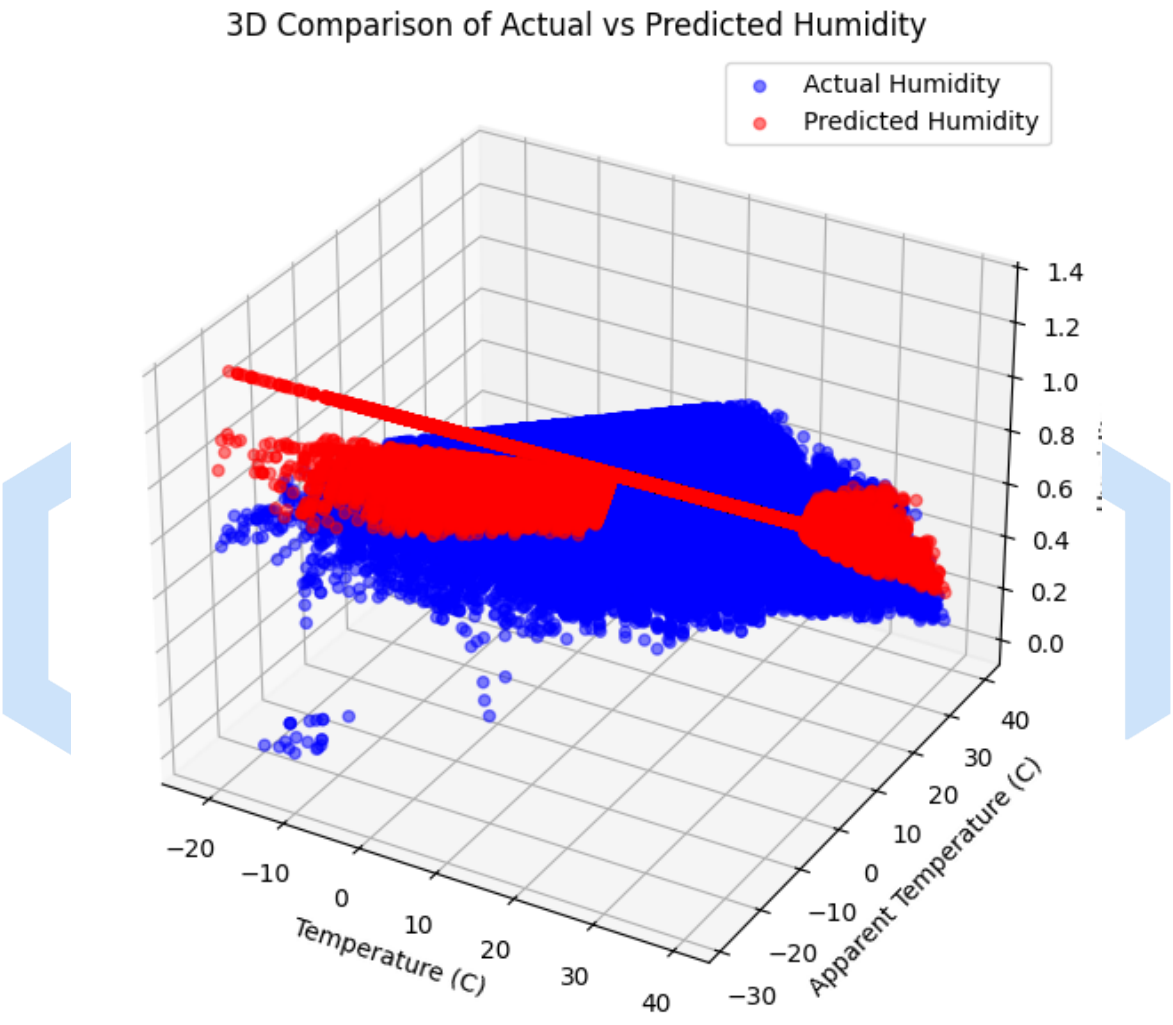
- تطابق نسبی: نزدیکی نقاط به خط ایده‌آل در محدوده متوسط (۰.۴ تا ۰.۸) نشان‌دهنده عملکرد خوب مدل در این محدوده است.
- انحراف در لبه‌ها: فاصله نقاط در رطوبت‌های پایین و بالا نشان‌دهنده ضعف مدل در پیش‌بینی مقادیر افراطی است، که ممکن است به پرت‌ها یا ناتوانی مدل خطی در مدل‌سازی روابط غیرخطی مربوط باشد.



میانگین نزدیک به صفر: این نشان‌دهنده تعادل کلی مدل است، که با طراحی LS سازگار است.

نوسانات: دامنه ±۰.۷۵ نشان‌دهنده خطاهای قابل‌توجه در برخی نقاط است، که ممکن است به پرت‌ها یا تغییرات فصلی مربوط باشد.

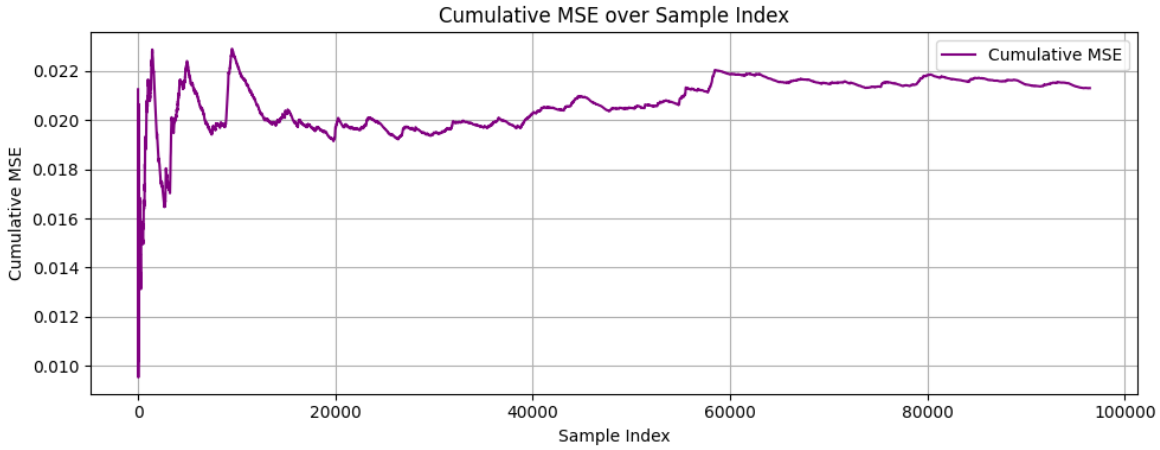
الگوهای خوشه‌ای: خوشه‌های باقیمانده‌های بزرگ (مثلاً نزدیک ۶۰۰۰۰) ممکن است نشان‌دهنده ناهمسانی واریانس (Heteroscedasticity) یا تأثیر پرت‌ها باشد.



نقاط آبی (رطوبت واقعی) وقرمز (رطوبت پیش بینی شده) در برخی نواحی نزدیک به هم هستند، اما اختلاف قابل توجهی در برخی نقاط وجود دارد.

ابر نقاط نشان دهنده تمرکز دادهها در محدوده دماهای ۰ تا ۳۰درجه سانتی کراد ودماهای محسوس مشابه است.

یک خط قرمز برجسته در وسط نمودار نشان دهنده انحراف سیستماتیک در پیش بینی ها ممکن است باشد.



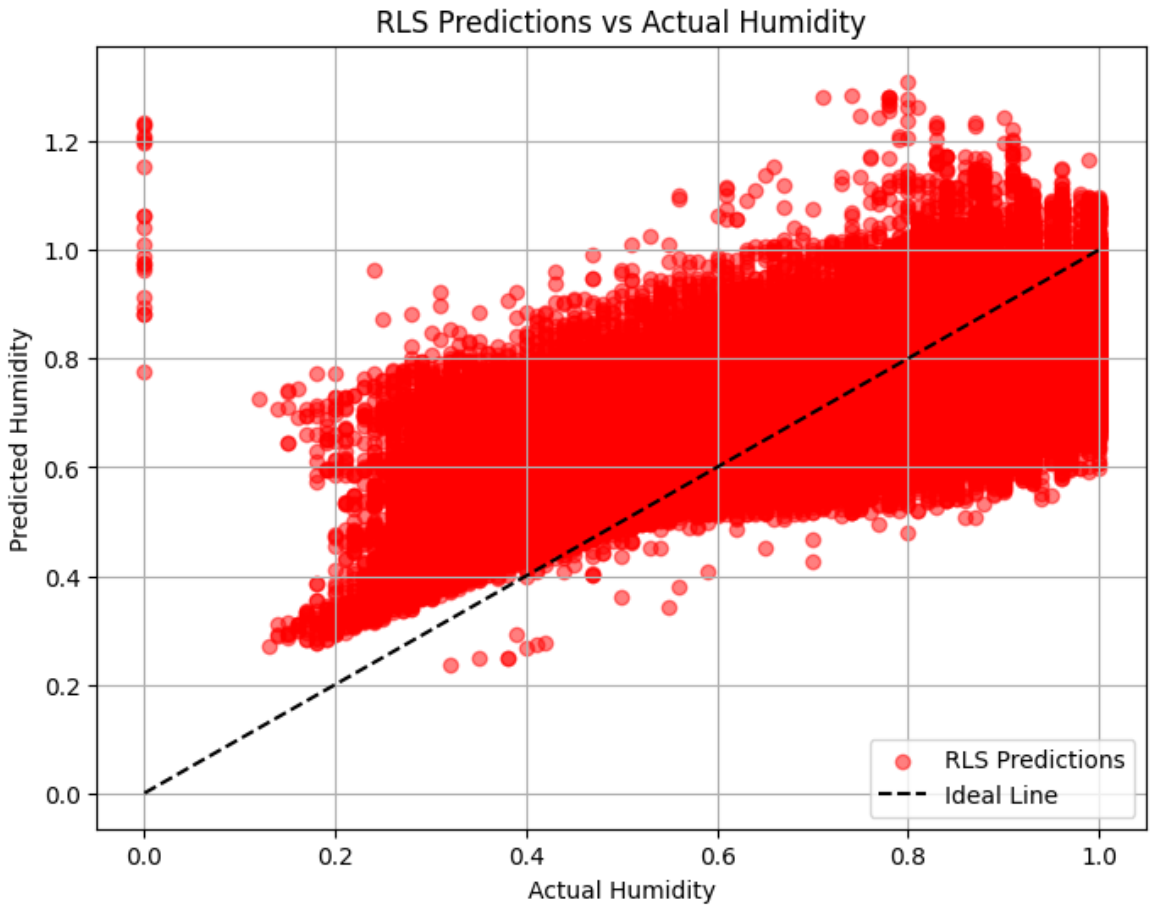
نمودار به یک مقدار ثابت (۰.۰۲۱) همگرا می شود، مدل پایدار است. نوسانات بزرگ نشان‌دهنده تأثیر پرت‌ها است.

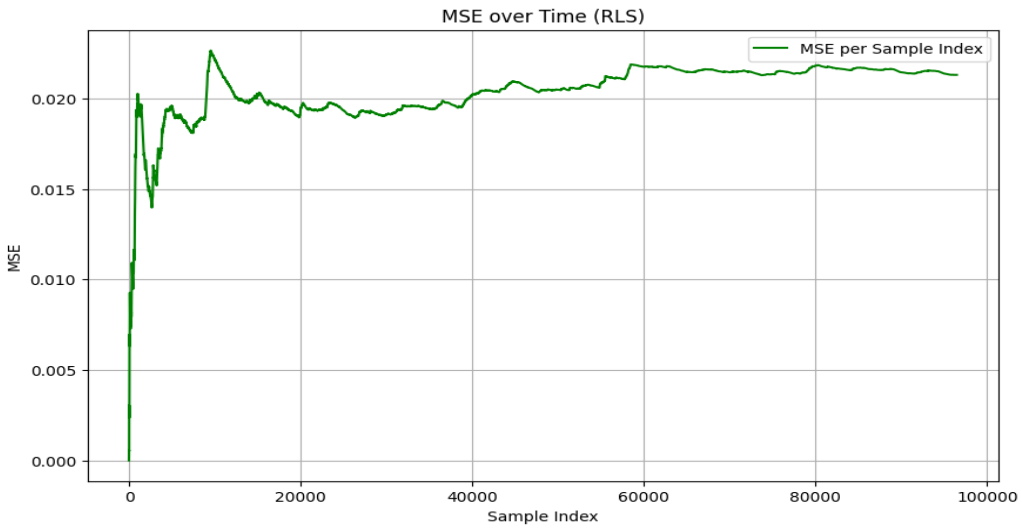
مدلسازی با روش RLS:

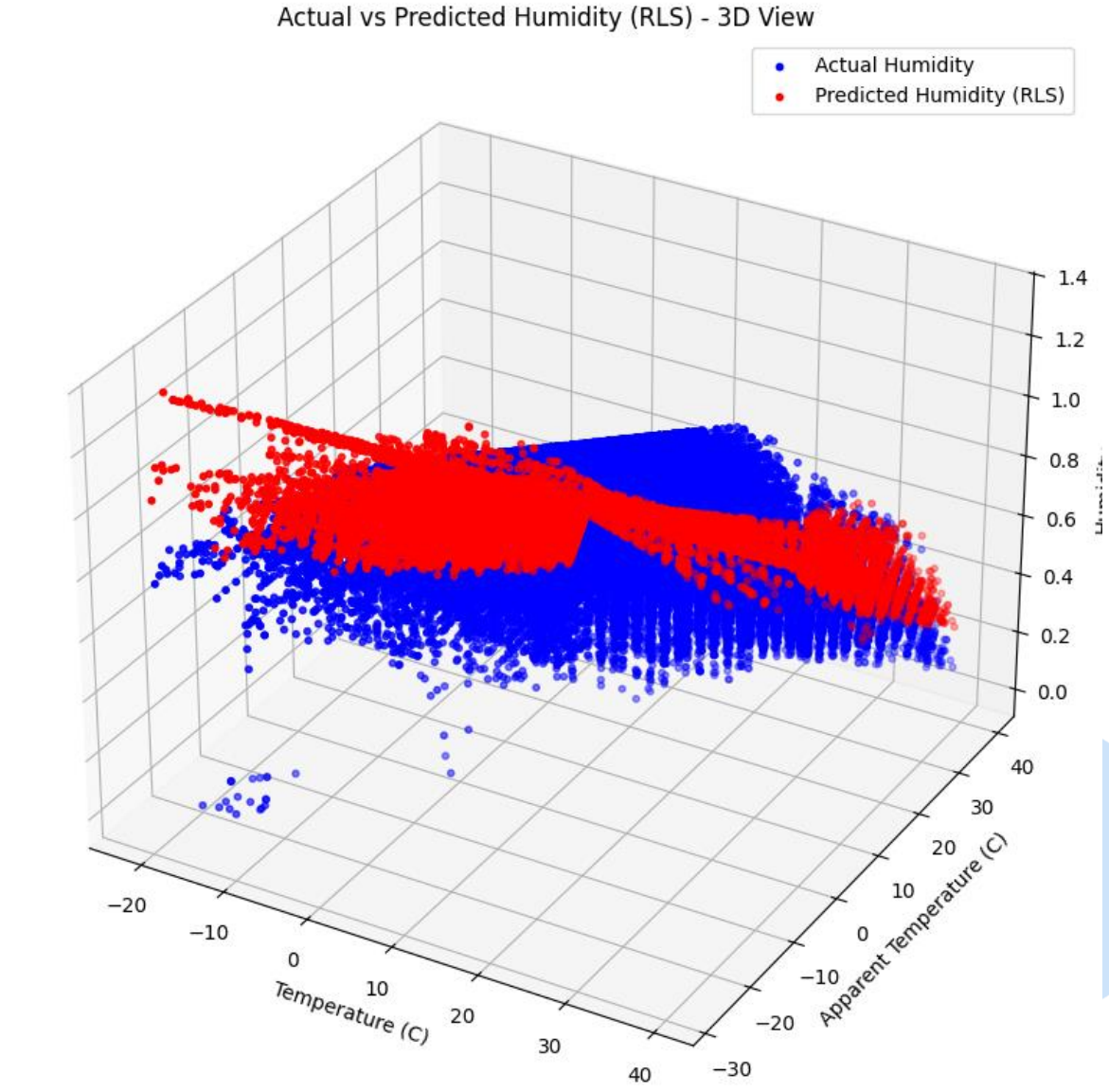
MSE: 0.0213

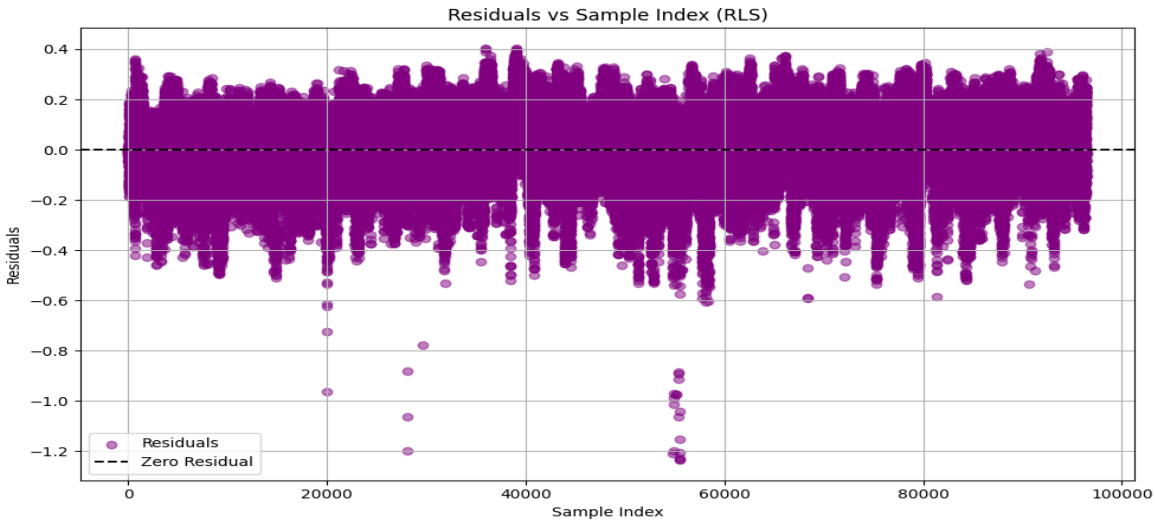
MAE: 0.1169

R2: 0.4425






Actual vs Predicted Humidity (RLS) - 3D View
Actual Humidity
Predicted Humidity (RLS)
Temperature (C)
Apparent Temperature (C)


Residuals vs Sample Index (RLS)
Residuals
Zero Residual
Residuals
Sample Index

شباهت با LS از آنجا که فاکتور فراموشی (λ) در کد مشخص نشده (به‌صورت پیش‌فرض λ=1فرض می‌شود(RLSبه LS کلاسیک تقلیل می‌یابد). این توضیح‌دهنده شباهت معیارها است.

پاسخ بخش ۳

مدل WLS (Weighted Least Squares) یا کمترین مربعات وزنی، نسخه‌ای از روش کمترین مربعات (LS) است که در آن فرض نمی‌شود همه مشاهدات دارای واریانس یکسانی باشند. در عوض، به هر داده یک وزن خاص داده می‌شود، تا تأثیر آن داده در تخمین پارامترها کم یا زیاد شود.

$$\hat{\theta} = (X^T Q X)^{-1} X^T Q y$$

$$J = \sum_{i=1}^{N} (e^i)^2 q^i = E^T Q E$$

$$Q = diag(q^i), q^i \in (0,1)$$

روش وزن دهی که انتخاب کردم بر اساس میزان پراکندگی بود

یعنی داده‌هایی که از میانگین فاصله زیادی دارند، پرت هستند در نتیجه وزن کمتر می‌گیرند.

در نتیجه، این روش باعث می‌شود مدل کمتر تحت تأثیر داده‌های پرت قرار بگیرد.

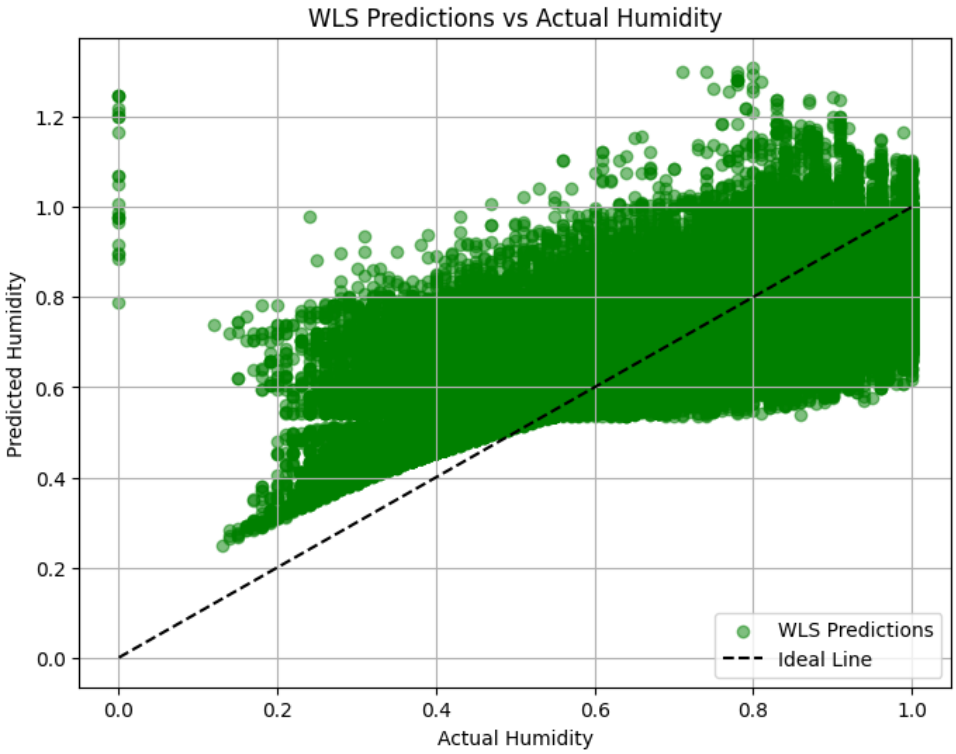
نتایج WLS:

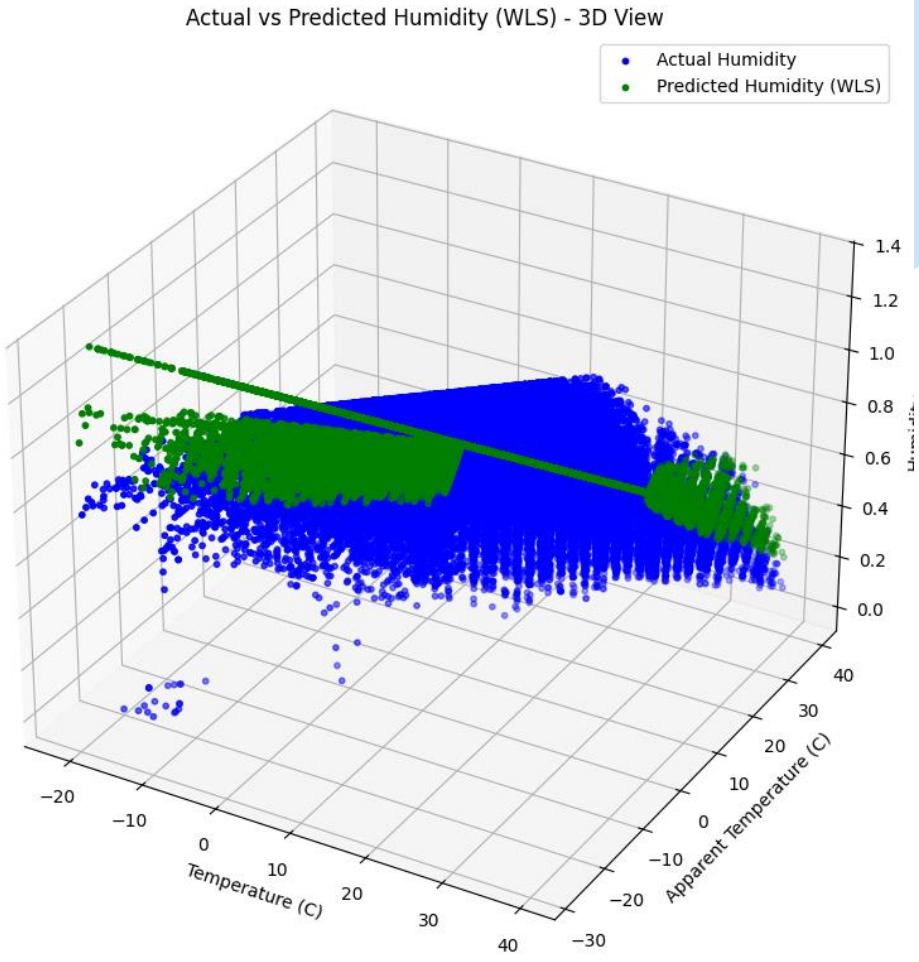
MSE: 0.0213

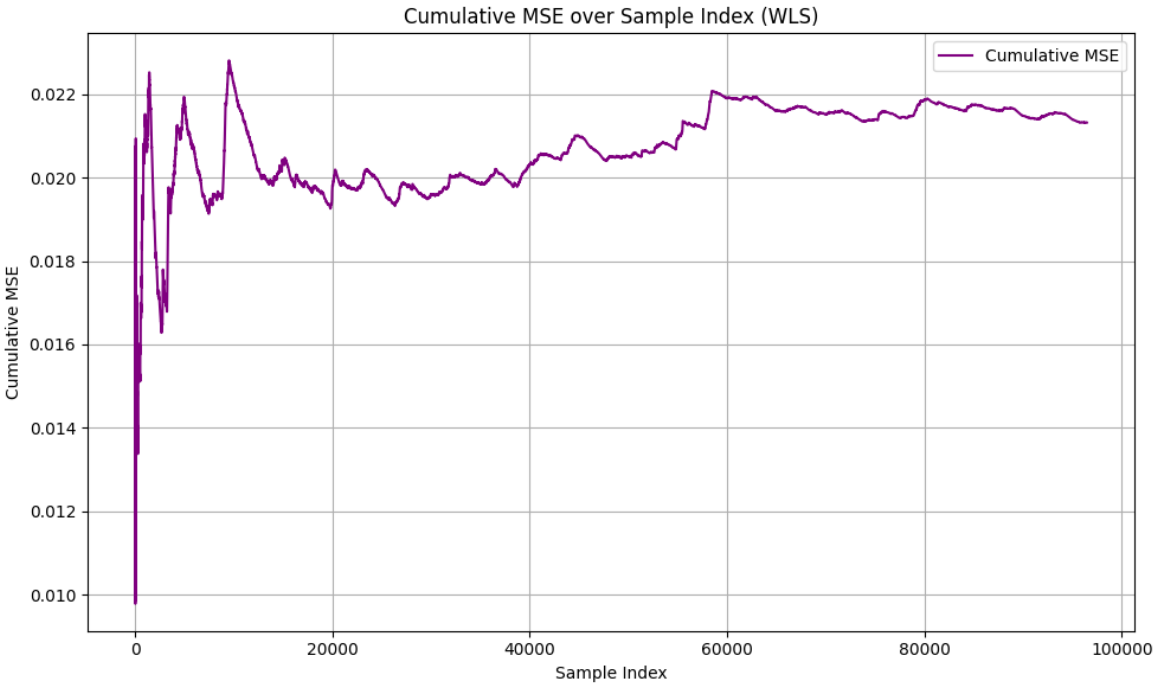
MAE: 0.1169

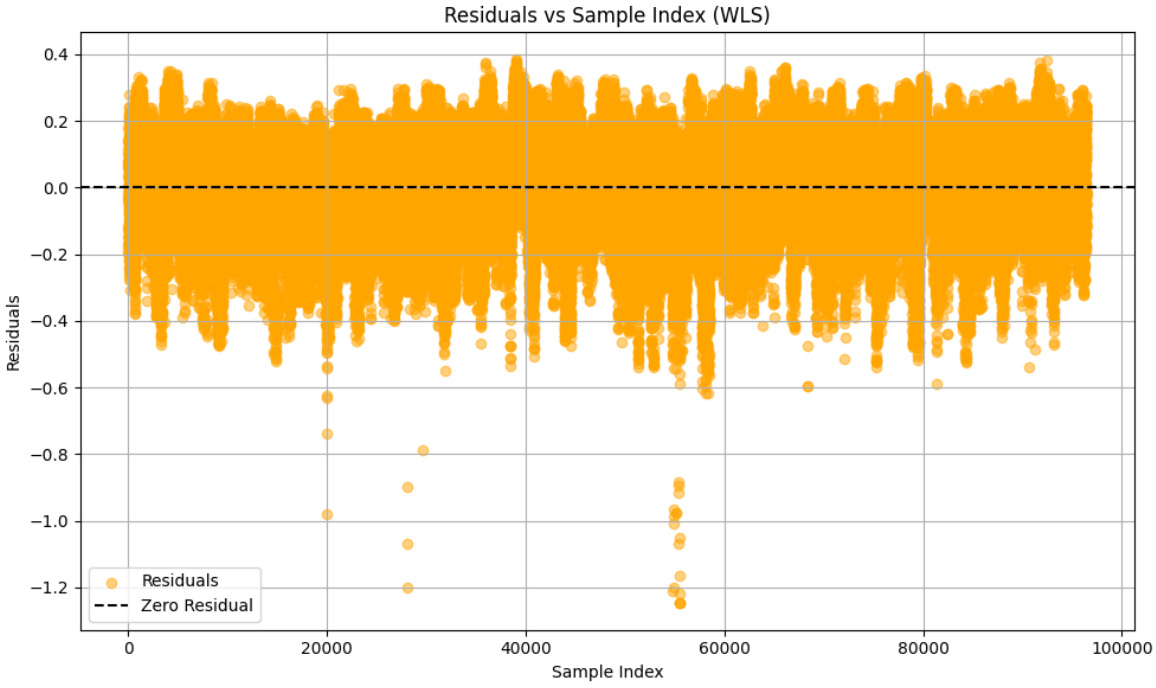
R2: 0.4420

WLS Coefficients (θ): [0.73720773 -0.44379557 0.32575899]


WLS Predictions vs Actual Humidity
Predicted Humidity
Actual Humidity
WLS Predictions
Ideal Line
0.0
0.2
0.4
0.6
0.8
1.0
1.2


Actual vs Predicted Humidity (WLS) - 3D View
Actual Humidity
Predicted Humidity (WLS)
Temperature (C)
Apparent Temperature (C)
−20
−10
0
10
20
30
40
−30
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0





همانگونه که می بینیم با استفاده از وزن دهی هم به نتایج بهتری دست پیدا نکردیم برای همین آمدم یک روش دیگر وزن دهی استفاده کردن باتوجه به مربعات خطای هر نمونه در بخش LS که باز هم به نتیجه بهتری نرسید برای همین مشکل چیزی فارغ از نحوه وزن دهی بود، احتمال هم خطی بودن دما و دمای محسوس هست که باعث می شود نتیجه بهتری نگیریم این احتمال را با دستور VIFچک کردم و

عدد ۶۸ را دریافت کردم که مقداری بسیار بزرگ تر از ۱۰ بود و نشان از هم خطی بودن دما و دمای محسوس هست و آمدم یکی را حذف کردم و یا یک بار متغیر جدید در نظر گرفتم که میانگین دما و دمای محسوس بود و در دو حالت هم باز به نتایج بهتری دست پیدا نکردم و تنها فرضی که باقی می ماند این است که یا دیتا ها داده پرت بسیار دارند یا کلا را بطه دما و رطوبت غیر خطی هست که با روش های خطی به خوبی نمی توان مدلسازی و پیش بینی کرد.