(ارزیابی انواع مدلسازی رگرسیون در پیش‌بینی دما با استفاده از شرایط جوی بین المللی)

راشد شهابی
دانشجوی کارشناسی مهندسی کامپیوتر، دانشگاه بجنورد
r.shahabi2001@gmail.com

علی ایزدی
دانشجوی کارشناسی ارشد مهندسی کامپیوتر، دانشگاه بجنورد
ali.izadi.ce@gmail.com

چکیده

در این تحقیق، یک مدل پیش‌بینی دما بر اساس مجموعه داده‌های شرایط جوی بین‌المللی از دیتاست World Weather Repository توسعه داده شده است. هدف این مطالعه مقایسه و ارزیابی عملکرد چندین روش رگرسیون از جمله رگرسیون خطی، رگرسیون چندکی، رگرسیون ستیغی، رگرسیون لاسو، شبکه الاستیک، رگرسیون مؤلفه‌های اصلی، رگرسیون حداقل مربعات جزئی و رگرسیون بردار پشتیبان برای پیش‌بینی دمای هوا است. پس از پیش‌پردازش داده‌ها و انتخاب ویژگی‌های مؤثر، هر یک از مدل‌ها با استفاده از مجموعه داده‌های آموزش داده‌شده‌اند و نتایج با استفاده از معیارهای ارزیابی مانند میانگین مطلق خطا (MAE)، میانگین مربعات خطا (MSE)، ریشه میانگین مربعات خطا (RMSE) و ضریب تعیین (R^2) مورد ارزیابی قرار می گیرند. نمودارهای پراکندگی برای نمایش تطابق مقادیر پیش‌بینی شده و مقادیر واقعی رسم شدند. تحلیل نتایج نشان داد که مدل رگرسیون بردار پشتیبان به دلیل دقت بالا و انعطاف‌پذیری در مدل‌سازی روابط غیرخطی، عملکرد بهتری نسبت به سایر مدل‌ها دارد، هرچند برخی از روش‌های دیگر مانند رگرسیون خطی و ستیغی نیز دقت مناسبی را نشان دادند. این مطالعه نشان می‌دهد که روش‌های رگرسیون می‌توانند به‌طور مؤثر برای پیش‌بینی داده‌های پیوسته، به‌ویژه در حوزه‌هایی مانند پیش‌بینی دما، به کار برده شوند.
برای مشاهده کد پروژه، به این لینک مراجعه کنید.

کلمات کلیدی: رگرسیون، ارزیابی مدل‌های رگرسیون، پیش‌بینی دما، داده‌های جوی

۱. مقدمه

در دنیای امروز، تغییرات دما و تأثیرات آن بر زندگی انسان‌ها، صنایع، کشاورزی و اکوسیستم‌ها توجه بسیاری را به خود جلب کرده است. با پیچیده‌تر شدن شرایط جوی و افزایش اثرات تغییرات اقلیمی، پیش‌بینی دقیق دما و تحلیل شرایط آب‌وهوایی به یکی از چالش‌های مهم جهانی تبدیل شده است. (سمائی و احمدی، ۱۳۹۳) در سال‌های اخیر، تکنیک‌های یادگیری ماشین[1] و مدل‌های پیش‌بینی در این زمینه به‌طور قابل توجهی پیشرفت کرده‌اند؛ که به ما کمک می‌کند با بهره‌گیری از داده‌های جوی و مدل‌های رگرسیون[2]، شرایط دمایی را با دقت بیشتر و بهینه‌تر، پیش‌بینی کنیم. (Cifuentes, 2020)

در گذشته، علم هواشناسی بیشتر از مدل‌های سنتی و روش‌های آماری برای پیش‌بینی وضعیت جوی و دما استفاده می‌کرد. این روش‌ها به دلیل محدودیت‌های محاسباتی و کمبود داده‌های جهانی معمولاً از دقت کمتری برخوردار بودند و نمی‌توانستند تغییرات پیچیده جوی را به خوبی شبیه‌سازی کنند (Harper, 2012). با ظهور الگوریتم‌های یادگیری ماشین و دسترسی بیشتر به داده‌های جهانی، روش‌های پیش‌بینی در سال‌های اخیر از مدل‌های سنتی فراتر رفته و به سمت روش‌های پیشرفته‌ای مانند رگرسیون روی آورده‌اند. (Rajashekar, 2024)

رگرسیون، به عنوان یک روش کلیدی در یادگیری تحت نظارت در یادگیری ماشین، امکان پیش‌بینی مقادیر پیوسته را با استفاده از الگوهای کشف‌شده در داده‌های آموزشی فراهم می‌آورد. استفاده از رگرسیون به دلیل توانایی آن در پیش‌بینی و تحلیل تأثیر متغیرها بر یکدیگر، در دهه‌های اخیر به طرز چشمگیری گسترش یافته است. روش‌های مختلف رگرسیون مانند رگرسیون خطی (Yan and Su, 2007)، چندکی (Koenker and Hallock , 2001)، ستیغی (Saleh et al, 2019)، لاسو (Tibshirani, 1996)، شبکه الاستیک (Zhang et al, 2017)، بردار پشتیبان (Basak et al, 2007)، مؤلفه اصلی و کمترین مربعات جزئی (Ergon, 2014) برای تحلیل داده‌های پیچیده در حوزه‌های مختلفی از جمله پزشکی (اکبر بیگلریان و همکاران، ۱۳۹۰)، اقتصاد (Nunkoo et al, 2020) و مهندسی به‌کار گرفته می‌شوند. با وجود این، هر مدل با ویژگی‌ها و دقت متفاوتی همراه است و انتخاب مدل مناسب نیازمند ارزیابی دقیق بر اساس معیارهای مختلف است.

در حوزه هواشناسی و علوم جوی، مدل‌های پیش‌بینی دما و شرایط آب و هوایی، نقش مهمی در برنامه‌ریزی‌های کلان کشورها، مدیریت منابع آب، انرژی و کاهش خسارات ناشی از پدیده‌های جوی ایفا می‌کنند (Huang, 2007). پیشرفت در استفاده از مدل‌های رگرسیون در این حوزه به ما این امکان را داده که اثر عوامل مختلف جوی از جمله رطوبت، فشار، میزان ابرناکی و کیفیت هوا را بر دمای جهانی را با دقت بیشتری بررسی کرده و پیش‌بینی‌های جوی دقیق‌تری ارائه دهیم.

در این مقاله، ما از مدل‌های مختلف رگرسیون برای بررسی اثر شرایط جوی بر دمای جهانی استفاده کرده‌ایم. مدل‌هایی چون رگرسیون خطی[3]، رگرسیون چندکی[4]، رگرسیون ستیغی[5]، لاسو[6]، شبکه الاستیک[7]، بردار پشتیبان[8]، مولفه‌های اصلی[9] و کمترین مربعات جزئی[10] ارزیابی شده‌اند. هدف این پژوهش، مقایسه عملکرد این مدل‌ها و تعیین مناسب‌ترین روش برای پیش‌بینی دما در شرایط جوی جهانی است. بر اساس ارزیابی‌های انجام‌شده، انتظار می‌رود که مدل بردار پشتیبان به دلیل توانایی در مدل‌سازی روابط غیرخطی و دقت بالا، نتایج بهتری در این تحلیل ارائه دهد.

---

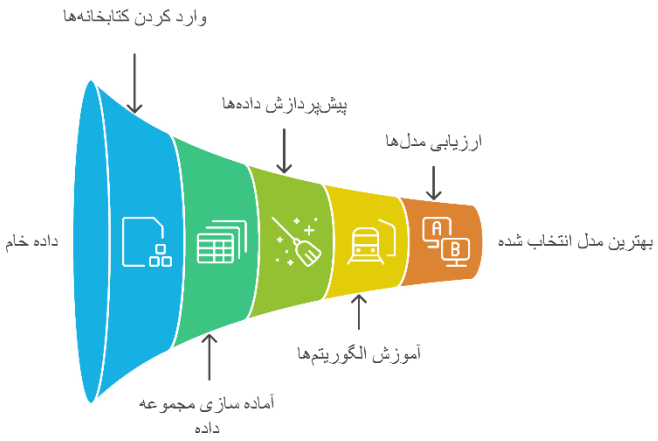
[1] Machine Learning
[2] Regression
[3] Linear Regression
[4] Quantile Regression
[5] Ridge Regression
[6] Lasso Regression
[7] Elastic Net Regression
[8] Support Vector Regression
[9] Principle Component Regression
[10] Partial Least Square Regression

۲. روش تحقیق

در این تحقیق، از یک مجموعه داده بین‌المللی شامل ۳۲۴۱۲ سطر و ۴۱ فیلد استفاده شده است که برای مدل‌سازی و پیش‌بینی دما در شرایط جوی به کار می‌رود. چندین مدل رگرسیون با استفاده از زبان برنامه‌نویسی پایتون برای این منظور ارزیابی شده‌اند. مراحل انجام کار در شکل ۱ آمده است. در مرحله اول، کتابخانه‌های مورد نیاز برای تحلیل داده و ساخت مدل‌ها وارد می‌شوند. در مرحله دوم، مجموعه داده آب‌وهوایی بین‌المللی به‌کار گرفته شده و سپس در مرحله سوم، داده‌ها پیش‌پردازش و برای مدل‌سازی آماده می‌شوند. مرحله چهارم به آموزش الگوریتم‌های مختلف یادگیری ماشین و مدل‌های رگرسیون اختصاص دارد. سپس در مرحله پنجم، نمودارهایی برای بررسی ارتباط ویژگی‌ها رسم می‌شود. در مرحله بعد، معیارهای ارزیابی مدل‌ها محاسبه و در نهایت، الگوریتم‌ها با هم مقایسه شده و بهترین مدل برای پیش‌بینی دما انتخاب می‌شود. در ادامه، هر یک از این مراحل به تفصیل توضیح داده می‌شود.

۲-۱. ورود کتابخانه‌های مورد نیاز

در فرآیند توسعه مدل‌های یادگیری ماشین و تحلیل داده‌ها، استفاده از کتابخانه‌های مختلف بسیار ضروری و تأثیرگذار است. این کتابخانه‌ها به توسعه‌دهندگان امکان می‌دهند تا تحلیل‌ها و مدل‌سازی‌ها را با سرعت و دقت بیشتری انجام دهند. اولین گام در این مسیر، وارد کردن کتابخانه‌های ضروری است که شامل کتابخانه‌های معروفی مانند Pandas، NumPy، Matplotlib و Scikit- Learn می‌شود.



شکل ۱. روند انجام کار

۲-۲. مجموعه داده

در این بخش از تحقیق، از مجموعه داده‌ای به نام GlobalWeatherRepository استفاده شده که از سایت Kaggle (Elgiriyewithana, 2024) دریافت شده است. از مجموعه داده‌ای استفاده شده که اطلاعاتی همچون موقعیت جغرافیایی، شرایط آب و هوایی، کیفیت هوا، و داده‌های نجومی مناطق مختلف جهان را در بر می‌گیرد. این مجموعه داده شامل ۳۲۴۱۲ رکورد و ۴۴ ویژگی است که در ادامه به طور کلی، به توصیف آن می‌پردازیم.

- نام کشور (Country): نام کشورهای جهان
- نام مکان (Location Name): نام مکان (پایتخت هر کشور)

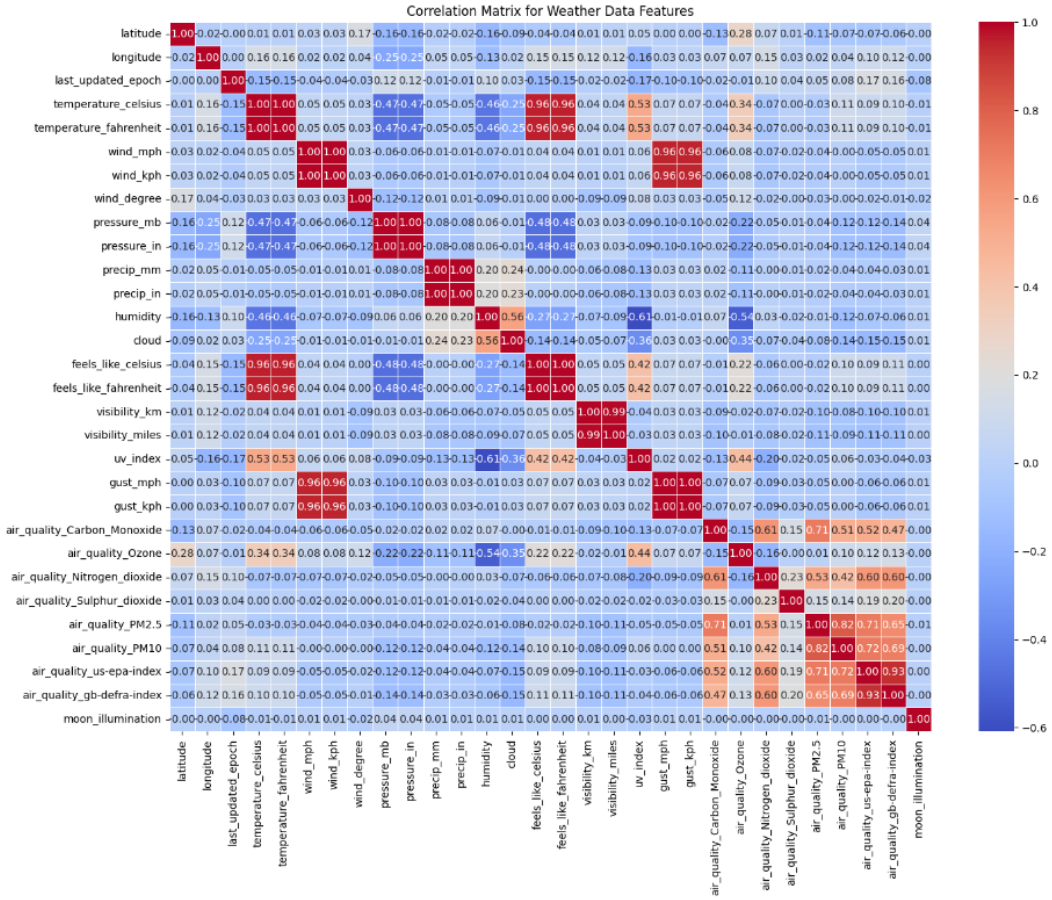
- عرض جغرافیایی (Latitude): مختصات عرض جغرافیایی مکان به درجه
- طول جغرافیایی (Longitude): مختصات طول جغرافیایی مکان به درجه
- منطقه زمانی (TimeZone): نشان‌دهنده منطقه زمانی هر مکان، بر اساس اختلاف زمانی نسبت به ساعت هماهنگ جهانی (UTC).
- آخرین به‌روزرسانی (Last Updated): زمان دقیق آخرین ثبت داده‌ها برای هر مکان
- دمای هوا (Temperature): دمای فعلی هوا به درجه سلسیوس و فارنهایت
- شرایط آب و هوایی (Weather Condition): توصیف شرایط جوی فعلی مانند صاف، ابری، نیمه‌ابری، بارانی و غیره
- سرعت باد (Wind Speed): سرعت باد در واحد کیلومتر بر ساعت و مایل بر ساعت
- جهت باد (Wind Direction): جهت وزش باد
- فشار هوا (Pressure): فشار فعلی هوا بر حسب میلی‌بار و اینچ جیوه
- بارش (Precipitation): میزان بارش فعلی به میلی‌متر و اینچ جیوه
- رطوبت نسبی (Humidity): درصد رطوبت نسبی موجود در هوا
- پوشش ابری (Cloud Cover): درصد پوشش ابرها در آسمان
- دید افقی (Visibility): مسافت دید افقی در واحد کیلومتر و مایل
- شاخص(UV Index): شاخص اشعه فرابنفش که شدت آن را مشخص می‌کند
- شاخص کیفیت هوا (Air Quality Index - AQI): شاخص کلی کیفیت هوا برای هر مکان
- طلوع خورشید (Sunrise): زمان طلوع خورشید بر اساس ساعت محلی
- غروب خورشید (Sunset): زمان غروب خورشید بر اساس ساعت محلی
- حالت ماه (Moon Phase): حالت فعلی ماه مانند هلال، نیمه و غیره
- دمای حس‌شده (Feels Like Temperature): دمایی که به دلیل تأثیرات رطوبت و باد، توسط انسان حس می‌شود

۲-۳. پیش‌پردازش داده‌ها

جهت پیش‌پردازش داده‌ها، مراحل زیر انجام می‌گیرند:

- حذف فیلدهای غیرضروری: در این مرحله، ابتدا ضریب همبستگی[11] بین تمامی ویژگی‌ها و متغیر هدف محاسبه شده و ماتریس همبستگی (شکل ۲) به ما نشان می‌دهد که کدام ویژگی‌ها رابطه قوی‌تر یا ضعیف‌تری با هدف دارند. از آنجایی که هدف ما کاهش تعداد ویژگی‌ها به مهم‌ترین و تأثیرگذارترین آنهاست، ویژگی‌هایی که مقدار همبستگی مطلق آنها با متغیر هدف کمتر از یک آستانه خاص بود (۰.۱)، حذف شدند. برای جلوگیری از بیش‌برازش[12] ویژگی دمای فعلی هوا به فارنهایت (temperature fahrenheit) نیز حذف شده است.

[11] Correlation Coefficients
[12] Overfitting



شکل ۲. ماتریس همبستگی

- تقسیم فیلدها به ویژگی‌ها و هدف: در این مجموعه داده، دمای فعلی به درجه سلسیوس به عنوان متغیر هدف در نظر گرفته شده است و ویژگی‌های مؤثر بر این متغیر نظیر شاخص کیفیت هوا (PM10T، Ozone)، پوشش ابری، رطوبت، فشار هوا، شاخص UV، دمای حس‌شده به درجه فارنهایت و سلسیوس و دمای هوا به درجه فارنهایت انتخاب شده‌اند. سایر فیلدهای غیرضروری برای اهداف این مطالعه حذف می‌شوند.

- تقسیم داده‌ها به مجموعه‌های آموزش و آزمون: برای ارزیابی مدل‌ها، داده‌ها به دو بخش آموزش و آزمون تقسیم می‌شوند. در این مجموعه داده، ۸۰ درصد از رکوردها برای آموزش و ۲۰ درصد برای آزمون در نظر گرفته می‌شوند.

۲-۴. آموزش مدل‌های مختلف رگرسیون

در این مرحله، داده‌ها با استفاده از الگوریتم‌های مختلف رگرسیون آموزش داده می‌شوند. این الگوریتم‌ها شامل رگرسیون خطی، رگرسیون چندکی، رگرسیون ستیغی، رگرسیون لاسو، رگرسیون شبکه الاستیک، رگرسیون مؤلفه‌های اصلی، رگرسیون کمترین مربعات جزئی و رگرسیون بردار پشتیبان هستند. هر یک از این مدل‌ها با توجه به داده‌های ورودی و ویژگی‌های تأثیرگذار، برای پیش‌بینی دمای فعلی آموزش می‌بینند. این فرایند به ما امکان می‌دهد تا نتایج هر مدل را بررسی و مقایسه کنیم.

۲-۵. نمودارها

پس از آموزش مدل‌های مختلف رگرسیون، نمودارهای پراکندگی[13] برای مقادیر واقعی و پیش‌بینی‌شده رسم می‌شوند. این نمودارها با نمایش یک نقطه برای هر جفت متغیر در مختصات دکارتی، رابطه بین متغیر پیش‌بینی‌شده و مقادیر واقعی را نشان می‌دهند. هدف از این نمودارها بررسی دقیق خطاهای مدل و تحلیل عملکرد هر یک از مدل‌ها در پیش‌بینی دما بر اساس شرایط جوی مختلف است. در بخش نتایج و یافته‌ها، این نمودارها به‌طور دقیق‌تر برای ارزیابی عملکرد مدل‌ها مورد تحلیل قرار خواهند گرفت.

۲-۶. محاسبه معیارهای ارزیابی

در این بخش، به بررسی و محاسبه معیارهای مختلف ارزیابی برای الگوریتم‌های به‌کار رفته می‌پردازیم. این معیارها به‌منظور سنجش عملکرد مدل‌های رگرسیونی و تعیین دقت پیش‌بینی آن‌ها استفاده می‌شوند. در اینجا y نمایانگر مقادیر واقعی، x نمایانگر مقادیر پیش‌بینی‌شده و n تعداد نمونه‌ها است.

- میانگین قدرمطلق خطا[14] (MAE): این معیار، اختلاف میان مقادیر پیش‌بینی‌شده و واقعی را به صورت قدرمطلق اندازه‌گیری می‌کند. این معیار، برای ارزیابی کلی خطا در پیش‌بینی‌ها مفید است و با استفاده از رابطه ۱ محاسبه می‌شود.

$$(۱)\quad MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i|$$

- میانگین مربع خطا[15] (MSE): این معیار معمولاً زمانی به کار می‌رود که هدف، تأکید بر خطاهای بزرگ‌تر باشد. در اینجا، تفاوت مقادیر واقعی و پیش‌بینی‌شده به توان دو می‌رسدو می‌تواند به تشخیص خطاهای جدی کمک کند. فرمول محاسبه میانگین مربع خطا در رابطه ۲ آورده شده است.

$$(۲)\quad MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2$$

- متوسط ریشه مربع خطا[16] (RMSE): این معیار مشابه میانگین مربع خطا است و برای زمانی که تمرکز بر مجازات خطاهای بزرگ‌تر در مقایسه با خطاهای کوچک‌تر است، به کار می‌رود همچنین به‌راحتی با واحدهای داده‌ها مقایسه می شود و به همین دلیل، تفسیر آن آسان‌تر است طبق رابطه ۳ محاسبه می‌شود.

$$(۳)\quad RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2}$$

- مربع R[17] (R^2): این معیار نشان‌دهنده تناسب مدل رگرسیون با داده‌ها است. بازه این معیار از ۰ تا ۱ بوده و به ما اطلاعاتی در مورد سهم تغییرات متغیر وابسته که توسط مدل توضیح داده می‌شود. هر چه مقدار آن بزرگ‌تر باشد، مدل تطابق بهتری با داده‌ها دارد. فرمول محاسبه مربع R در رابطه ۴ ارائه شده است که نمایانگر میانگین نمونه است.

$$(۴)\quad R^2 = 1 - \frac{\sum_i(y_i - x_i)^2}{\sum_i(y_i - y_m)^2}$$

---

[13] Scatter Plot
[14] Mean Absolute Error
[15] Mean Squared Error
[16] Root Mean Squared Error
[17] R-squared

- میانگین درصد خطای مطلق[18] (MAPE): این معیار، تفاوت میان مقادیر واقعی و پیش‌بینی‌شده را به صورت درصد اندازه‌گیری می‌کند این معیار، به‌ویژه برای مقایسه مدل‌ها در زمینه‌های مختلف کاربردی است و خروجی MAPE همیشه غیرمنفی بوده و مقدار صفر نشان‌دهنده بهترین عملکرد ممکن مدل است، که نشان می‌دهد پیش‌بینی بدون هیچ خطایی انجام شده است.

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - x_i}{y_i}\right| \quad (۵)$$

۲-۷. ارزیابی و انتخاب بهترین مدل

در این مرحله، مدل‌های مختلف رگرسیون با استفاده از معیارهای ارزیابی مقایسه می‌شوند تا دقت و کارایی هر یک در پیش‌بینی دما سنجیده شود. پس از تحلیل نتایج و بررسی این معیارها، مناسب‌ترین مدل از میان مدل‌های آزموده‌شده انتخاب خواهد شد.

۳. نتایج

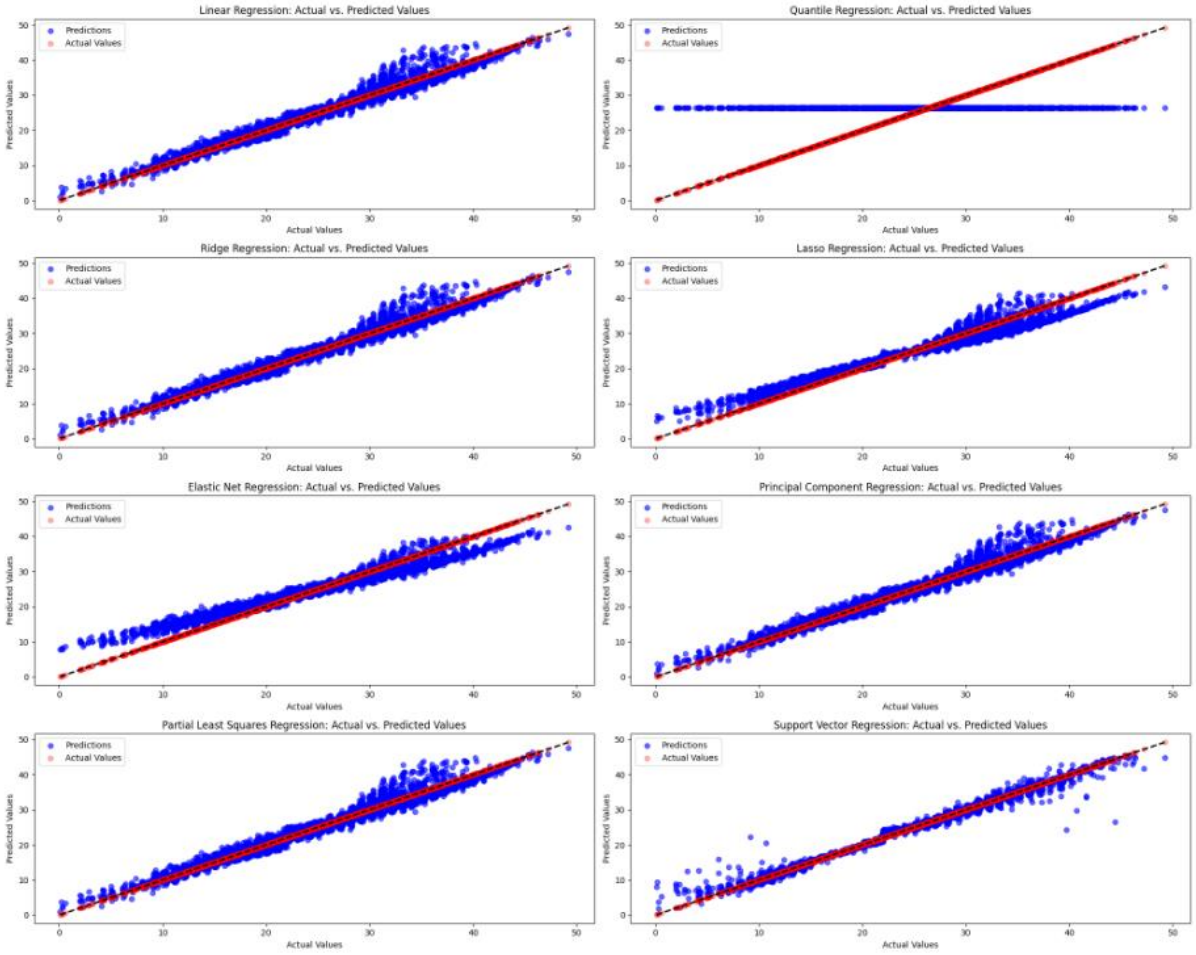
در این بخش، نتایج به‌دست‌آمده از مدل‌های مختلف رگرسیون تحلیل و بررسی می‌شوند. ابتدا نمودارهای پراکندگی مقادیر واقعی و پیش‌بینی‌شده ترسیم و تحلیل خواهند شد. سپس، مدل‌ها با استفاده از معیارهای ارزیابی گوناگون با یکدیگر مقایسه شده و دقیق‌ترین و کارآمدترین مدل انتخاب می‌شود. مدل‌های رگرسیونی مورد استفاده در این مطالعه شامل رگرسیون خطی، چندکی، ستیغی، لاسو، شبکه الاستیک، مؤلفه‌های اصلی، کمترین مربعات جزئی و بردار پشتیبان هستند.

۳-۱. نمودار پراکندگی

نمودارهای پراکندگی در شکل ۳ توزیع مقادیر پیش‌بینی‌شده را در مقابل مقادیر واقعی برای هر یک از مدل‌های رگرسیون نشان می‌دهند. در این مطالعه، ۲۰ درصد از داده‌ها، معادل ۶۴۸۲ رکورد، به‌عنوان مجموعه آزمون انتخاب شده است. محور افقی نمودارها دمای واقعی هر نمونه در مجموعه آزمون را نمایش می‌دهد، و محور عمودی دمای پیش‌بینی‌شده توسط هر مدل را نشان می‌دهد. هر نقطه در نمودار نشان‌دهنده یک نمونه از داده‌ها است.

بر اساس نمودارهای شکل ۳، مدل‌هایی مانند رگرسیون بردار پشتیبان، خطی، ستیغی ارتباط قوی‌تری بین مقادیر واقعی و پیش‌بینی شده را نشان می‌دهند، زیرا نقاط در این نمودارها به خط قطری نزدیک‌تر هستند. از سوی دیگر، مدلی مانند رگرسیون چندکی فاصله بیشتری بین مقادیر واقعی و پیش‌بینی شده دارد که بیانگر دقت کمتر آن در پیش‌بینی است.

[18] Mean Absolute Percentage Error



شکل ۳. نمودارهای پراکندگی برای ۸ مدل آموزشی مختلف

۳-۲. معیارهای ارزیابی

برای ارزیابی دقت و عملکرد هر یک از مدل‌های رگرسیون، از معیارهای ارزیابی مختلفی استفاده شده است که شامل میانگین قدر مطلق خطا (MAE)، میانگین مربع خطا (MSE)، ریشه میانگین مربع خطا (RMSE)، میانگین درصد خطای مطلق (MAPE) و ضریب تعیین (R-squared) می‌باشد. در جدول ۱، مقادیر این معیارها برای هر یک از مدل‌های رگرسیون محاسبه و ثبت شده است.

مقدار ضریب تعیین (R-squared) بین ۰ تا ۱ است، و هرچه این مقدار بیشتر باشد، مدل هماهنگی بیشتری با داده‌ها دارد و بهتر می‌تواند مقادیر واقعی را پیش‌بینی کند. مدل‌هایی که ضریب تعیین بیشتری دارند، دقت بیشتری در پیش‌بینی دما داشته‌اند. همچنین، برای سایر معیارها مانند MAE، MSE و RMSE، هر چه مقدار این خطاها کمتر باشد، مدل عملکرد بهتری دارد. بر اساس نتایج به‌دست‌آمده، مدل‌های رگرسیون بردار پشتیبان، خطی و ستیغی با کمترین میزان خطا و ضریب تعیین بالاتر، عملکرد بهتری در پیش‌بینی دما داشته‌اند، در حالی که مدل رگرسیون چندکی دارای خطای بیشتری بوده و به همین دلیل به‌عنوان مدل نامناسب شناسایی شده است.

جدول ۱. مقادیر معیارها برای ۸ مدل مختلف

Models	MAE	MSE	RMSE	R-Squared	MAPE
Linear Regression	0.0187641	0.0005699	0.0238728	0.9999889	0.0008973
Quantile Regression	5.4160333	52.2440098	7.2280017	-0.0154717	0.4081761
Ridge Regression	0.018809	0.0005769	0.0240196	0.9999887	0.000904
Lasso Regression	0.7616964	0.990137	0.9950563	0.9807546	0.05513
Elastic Net Regression	1.2615206	2.8399279	1.6852085	0.9448001	0.0905793
Principal Component Regression	0.7215246	0.9539551	0.9767062	0.9814578	0.0353279
Partial Least Squares Regression	0.0193828	0.0006032	0.0245614	0.9999882	0.0009072
Support Vector Regression	0.1405497	0.3355225	0.5792431	0.9934784	0.0327164

۳-۴. مقایسه مدل‌های مختلف

در این بخش، مدل‌های رگرسیون مختلف بر اساس معیارهای RMSE و MAE با هم مقایسه می‌شوند. جدول ۲ مدل‌ها را بر اساس مقدار RMSE مرتب کرده است؛ به این ترتیب که مدل با کمترین مقدار RMSE، یعنی رگرسیون بردار پشتیبان، بهترین عملکرد را نشان می‌دهد و در مقابل، رگرسیون چندکی بیشترین خطا را دارد. در جدول ۳ نیز مدل‌ها بر اساس مقدار MAEمرتب شده‌اند و دوباره مشخص شد که رگرسیون بردار پشتیبان کمترین خطا را داشته و دقیق‌ترین پیش‌بینی‌ها را ارائه می‌دهد، در حالی که رگرسیون چندکی دقت خیلی کمتری نشان داده است. با توجه به این مقایسه‌ها، مدل رگرسیون بردار پشتیبان به عنوان بهترین گزینه برای پیش‌بینی دما در شرایط مختلف جوی انتخاب می‌شود.

جدول ۲. رتبه‌بندی مدل‌ها بر اساس معیار RMSE

Rank	Models	RMSE
1	Support Vector Regression	0.8694194
2	Linear Regression	1.0374023
3	Ridge Regression	1.3911028
4	Partial Least Squares Regression	1.3916036
5	Principal Component Regression	1.3916116
6	Lasso Regression	1.9161798
7	Elastic Net Regression	2.1946457
8	Quantile Regression	7.2280017

جدول ۳. رتبه‌بندی مدل‌ها بر اساس معیار MAE

Rank	Models	MAE
1	Support Vector Regression	0.4883417
2	Ridge Regression	1.0372347
3	Linear Regression	1.0374023
4	Partial Least Squares Regression	1.0375614
5	Principal Component Regression	1.0376121
6	Lasso Regression	1.4231967
7	Elastic Net Regression	1.6648589
8	Quantile Regression	5.4160333

۴. نتایج

در این پژوهش، از داده‌های World Weather Repository که روزانه به‌روزرسانی می‌شود، برای پیش‌بینی دمای هوا بر اساس شرایط جوی جهانی بهره گرفتیم. مدل‌های رگرسیون ارزیابی‌شده شامل رگرسیون خطی، چندکی، ستیغی، لاسو، شبکه الاستیک، مؤلفه‌های اصلی، کمترین مربعات جزئی و بردار پشتیبان بوده‌اند. نتایج نشان دادند که رگرسیون بردار پشتیبان با دقت بالا و توانایی در مدیریت داده‌های غیرخطی، عملکرد بهتری نسبت به سایر مدل‌ها داشته است. همچنین، مدل‌های ستیغی، خطی، مولفه‌های اصلی و کمترین مربعات جزئی نیز به نتایج قابل قبولی دست یافتند و رقابت تنگاتنگی داشتند. در مقابل، نمودارهای پراکندگی و معیارهای ارزیابی نشان داده شد که رگرسیون چندکی برای مجموعه داده‌های مورد استفاده نتایج مطلوبی ندارد. دلایلی مانند تناسب پایین مدل با پیچیدگی داده‌ها و محدودیت‌های ذاتی رگرسیون چندکی، نشان می‌دهد که این مدل در اینجا نتوانسته است تغییرات کلی داده‌ها را به خوبی تفسیر کند. در حالی که مدل بردار پشتیبان بالاترین دقت را در پیش‌بینی دمای هوا ارائه کرده است.

منابع

سمائی، سیدرضا و بهدادفر، الهام،۱۴۰۲، تاثیرات تغییرات اقلیمی بر محیط زیست و انسان‌ها

Cifuentes, J., Marulanda, G., Bello, A., & Reneses, J. (2020). Air temperature forecasting using machine learning techniques: a review. Energies, 13(16), 4215.

Harper, K. C. (2012). Weather by the numbers: The genesis of modern meteorology. mit Press.

Rajashekar, P. (2024). Enhancing Weather Forecasting Precision through Advanced Machine Learning Techniques (Doctoral dissertation, CALIFORNIA STATE UNIVERSITY, NORTHRIDGE).

Yan, X., & Su, X. (2009). Linear regression analysis: theory and computing. world scientific.

Koenker, R., & Hallock, K. F. (2001). Quantile regression. Journal of economic perspectives, 15(4), 143-156.

Saleh, A. M. E., Arashi, M., & Kibria, B. G. (2019). Theory of ridge regression estimation with applications. John Wiley & Sons.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1), 267-288.

Zhang, Z., Lai, Z., Xu, Y., Shao, L., Wu, J., & Xie, G. S. (2017). Discriminative elastic-net regularized linear regression. IEEE Transactions on Image Processing, 26(3), 1466-1481.

Basak, D., Pal, S., & Patranabis, D. C. (2007). Support vector regression. Neural Information Processing-Letters and Reviews, 11(10), 203-224.

Ergon, R., Granato, D., & Ares, G. (2014). Principal component regression (PCR) and partial least squares regression (PLSR). Mathematical and statistical methods in food science and technology Wiley Blackwell, Chichester, 121-42.

بیگلریان، اکبر، بخشی، عنایت اله، رهگذر، مهدی، و کریملو، مسعود. (۱۳۹۰). مقایسه شبکه عصبی مصنوعی و رگرسیون لجستیک در پیش بینی پاسخ‌های دو حالتی مطالعات پزشکی. مجله دانشگاه علوم پزشکی خراسان شمالی، ۳(ویژه نامه آمار زیستی و اپیدمیولوژی)

Nunkoo, R., Seetanah, B., Jaffur, Z. R. K., Moraghen, P. G. W., & Sannassee, R. V. (2020). Tourism and economic growth: A meta-regression analysis. Journal of Travel Research, 59(3), 404-423.

Huang, M. (2020, July). Theory and Implementation of linear regression. In 2020 International conference on computer vision, image and deep learning (CVIDL) (pp. 210-217). IEEE.

https://www.kaggle.com/datasets/nelgiriyewithana/global-weather-repository

Abstract:
In this study, a temperature prediction model based on the international weather conditions dataset from the World Weather Repository has been developed. The aim of this research is to compare and evaluate the performance of several regression methods, including Linear Regression, Quantile Regression, Ridge Regression, Lasso Regression, Elastic Net, Principal Component Regression, Partial Least Squares Regression, and Support Vector Regression, for predicting air temperature. After data preprocessing and selection of influential features, each model is trained using the dataset, and results are evaluated using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). Scatter plots were drawn to display the alignment between predicted and actual values. The results analysis showed that Linear Regression performed better than other models due to its simplicity and efficiency, although other methods like Partial Least Squares, Support Vector, and Ridge Regression also demonstrated acceptable accuracy. This study indicates that regression methods can be effectively applied for predicting continuous data, particularly in areas such as temperature forecasting.