دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مدیریت علم و فناوری

گزارش کار هفته دهم

رگرسیون

نگارش

رضا اکبری مقدم

استاد

دکتر مهدی قطعی

دی ماه ۹۹

فهرست مطالب

فهرست اشکال

مقدمه

رگرسیون یک روش آماری بوده و در اقتصاد، برنامه نویسی و فعالیت های دیگر استفاده می شود. هدف رگرسیون شناسایی قدرت و خواص یک متغیر وابسته نسبت به متغیرهای دیگر (که به متغیر های مستقل معروفند) می‌باشد.

رگرسیون را می‌توان در سه دسته تقسیم‌بندی کرد که عبارتند از رگرسیون خطی، غیرخطی و رگرسیون خطی چندگانه (چند متغیره). در روش خطی برای توصیف یا پیش‌بینی خروجی ناشی از چند متغیر مستقل، تنها بر اساس یک متغیر وابسته انجام می‌گیرد. در رگرسیون چند متغیره برای همین منظور از دو یا چند متغیر وابسته استفاده می‌شود.

رگرسیون غیر خطی برای مواردی استفاده می شود که به دنبال یافتن رابطه‌ای غیر خطی بین متغیرها هستیم. از این روش در محاسبات بسیار پیچیده استفاده می‌شود.

معرفی مجموعه داده

مجموعه داده Combined Cycle Power Plant که اطلاعات چرخه نیروگاه ها طی ۶ سال زمانی که با حداکثر توان مشغول به کار بودند را جمع آوری نموده است. ویژگی ها از متغیرهای متوسط ساعتی دما (T) ، فشار محیط (AP) ، رطوبت نسبی (RH) و خلا اگزوز (V) برای پیش بینی خالص انرژی الکتریکی ساعتی (EP) نیروگاه تشکیل شده است.

مراحل انجام کار

در این گزارش با استفاده از ابزار پایتون به بررسی این مجموعه داده میپردازیم.

پس از اضافه کردن مجموعه داده نمایی از جدول این مجموعه داده میگیریم.

	AT	V	AP	RH	PE
0	14.96	41.76	1024.07	73.17	463.26
1	25.18	62.96	1020.04	59.08	444.37
2	5.11	39.40	1012.16	92.14	488.56
3	20.86	57.32	1010.24	76.64	446.48
4	10.82	37.50	1009.23	96.62	473.90
...	...	...	...	...	...
9563	16.65	49.69	1014.01	91.00	460.03
9564	13.19	39.18	1023.67	66.78	469.62
9565	31.32	74.33	1012.92	36.48	429.57
9566	24.48	69.45	1013.86	62.39	435.74
9567	21.60	62.52	1017.23	67.87	453.28

9568 rows × 5 columns

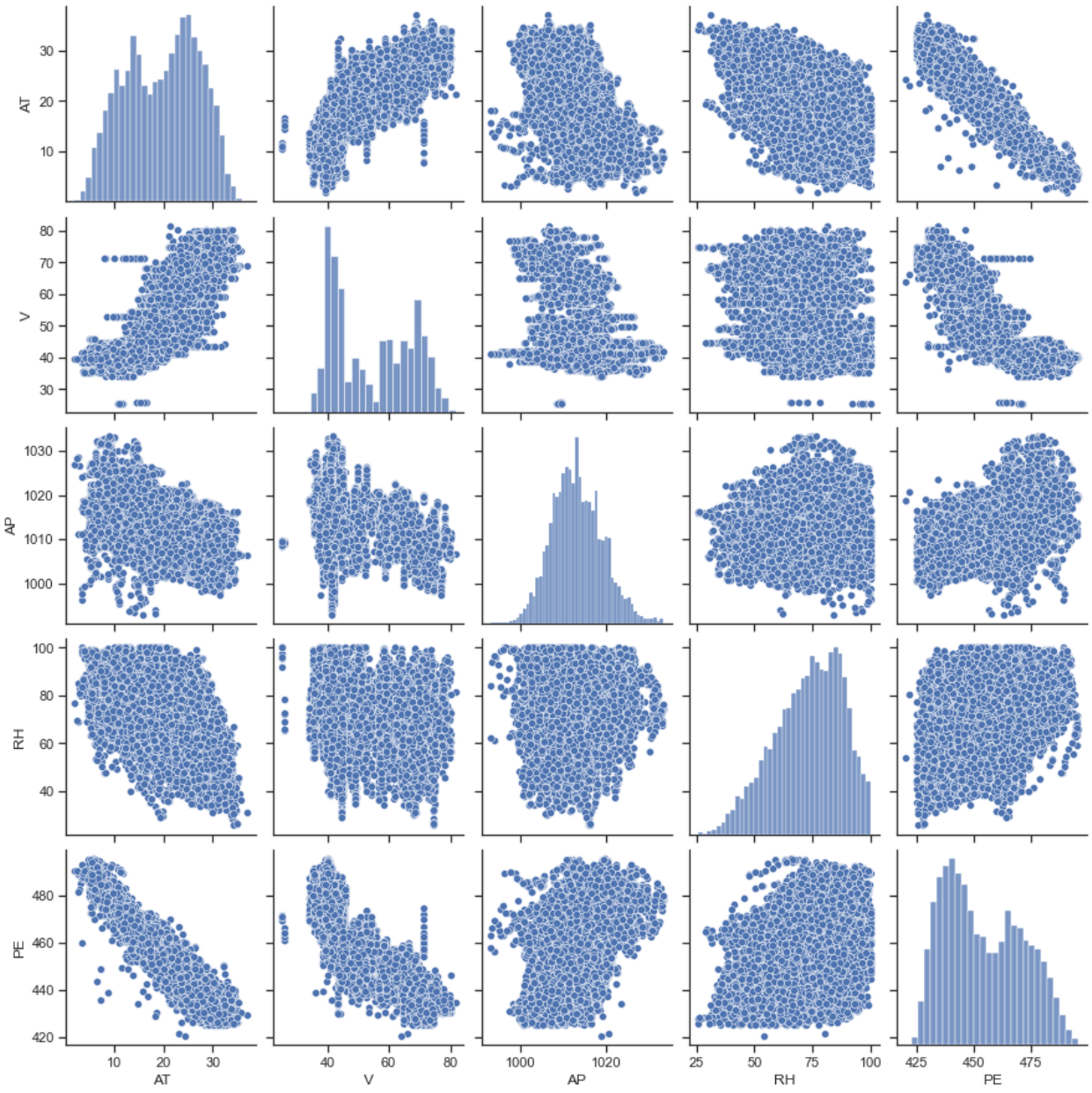
شکل ۱

سپس یک شمای کلی از وضعیت آماری داده ها و ویژگی های این مجموعه داده به نمایش گذاشته شده است.

	AT	V	AP	RH	PE
count	9568.000000	9568.000000	9568.000000	9568.000000	9568.000000
mean	19.651231	54.305804	1013.259078	73.308978	454.365009
std	7.452473	12.707893	5.938784	14.600269	17.066995
min	1.810000	25.360000	992.890000	25.560000	420.260000
25%	13.510000	41.740000	1009.100000	63.327500	439.750000
50%	20.345000	52.080000	1012.940000	74.975000	451.550000
75%	25.720000	66.540000	1017.260000	84.830000	468.430000
max	37.110000	81.560000	1033.300000	100.160000	495.760000

شکل ۲

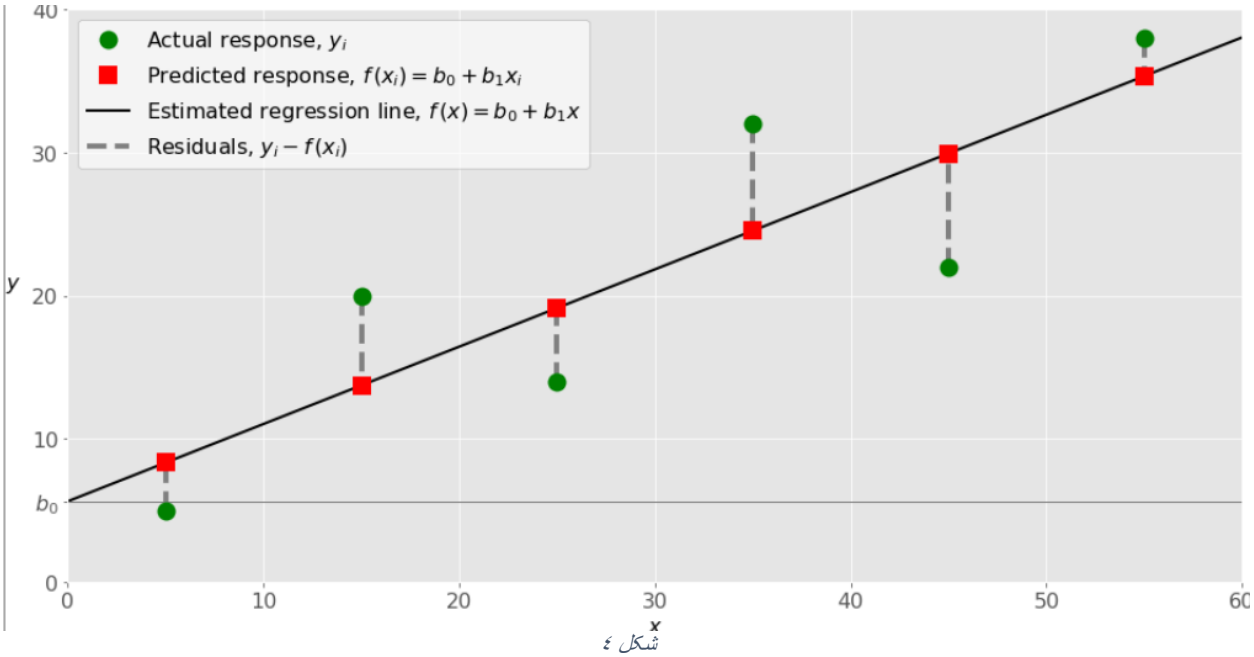
در این شمای کلی از آمار ویژگی ها مشاهده میشود که متوسط دمای این مجموعه داده ۱۹,۵ درجه سانتی گراد و متوسط فشار ۱۰۱۳ میلی بار میباشد.



شکل ۳

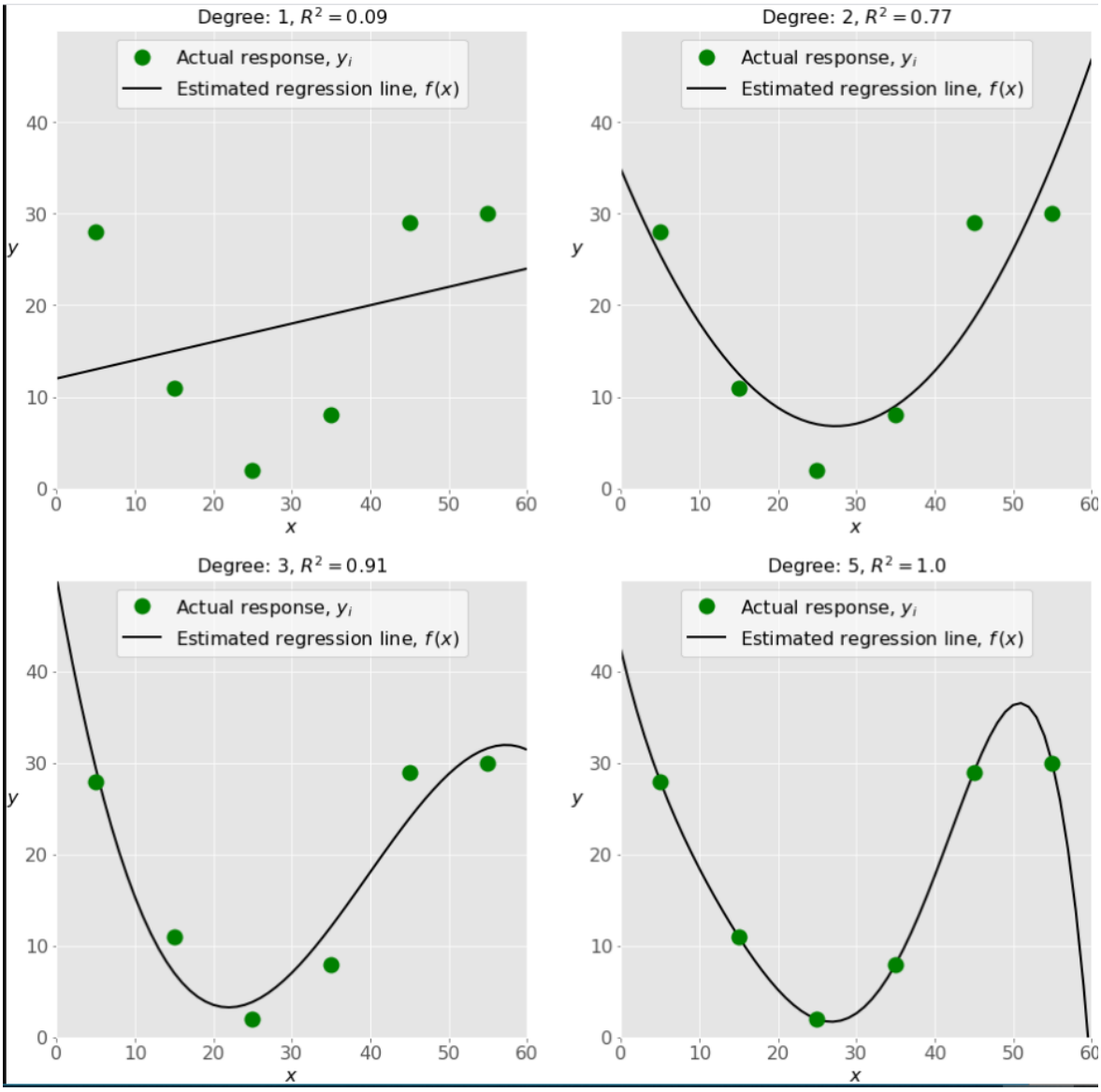
در این شکل نمودار پراکندگی هر کدام از ویژگی ها نمایش داده شده است.

در این قسمت میخواهیم با استفاده از `LinearRegression` به بررسی رگرسیون خطی ساده بپردازیم.



شکل ٤

همانطور که در شکل مشاهده میشود در رگرسیون خطی ساده یک خط به عنوان پیشبینی رگرسیون در نظر گرفته میشود که معادله مربوط آن در تصویر مشاهده میشود.



شکل ۵

همچنین در معیار R2 مشاهده میشود که هر چه این مقدار آن به یک نزدیکتر باشد پیشبینی مدل ما دقیقتر است اما در مقادیر نزدیک ۱ مشاهده میشود که اورفیت رخ میدهد.

ویژگی PE که توان خروجی نیروگاه میباشد را به عنوان هدف و نسبت به بقیه ویژگی ها بررسی میکنیم.

در اینجا فقط به بررسی یکی از ویژگی ها میپردازیم و ویژگی دما را نسبت به توان خروجی بررسی میکنیم.

پس از اعمال مدل بر روی این بخش از داده ها معیار R2 را محاسبه میکنیم.

```
r_sq = model.score(w, y)
print('R Square:', r_sq)
```

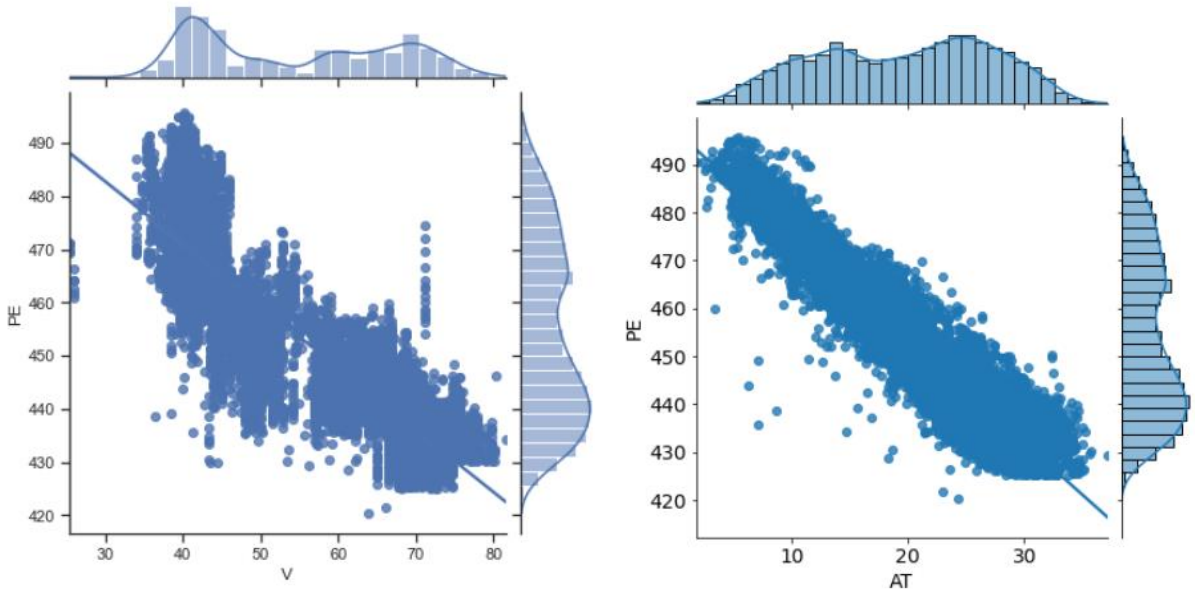
```
R Square: 0.8989475964148236
```

شکل ۶

مشاهده میشود که مقدار R2 برای این بخش از داده برابر با ۰,۸۹ میباشد که مقدار قابل قبولی است.
همچنین مقادیر intercept و slope را برای این بخش محاسبه میکنیم.

```
intercept: [497.03411989]
slope: [[-2.17131996]]
```

شکل ۷



شکل ۸

که در شکل بالا نمایی از نمودار پراکندگی و این مدل برای دو ویژگی AT و V نمایش داده شده است.
در شکل زیر یک خلاصه ای از کارایی این مدل آورده شده است که نشان میدهد میزان خطاها مقدار قابل توجهی کم میباشد.

```
Performance of Simple Linear Regression :
MSE :  3.222
Absloute Error :  6.784445
Correlation :  0.8912554
R Square :  0.898947
```

شکل ۹

Multiple Linear Regression

حال رگرسیون چندگانه این مجموعه داده را بر اساس تمامی ویژگی ها محاسبه میکنیم.

در شکل زیر مشاهده میشود که مقدار R2 برای این مدل ۰,۹۳۱ و همینطور جدولی از سایر معیارها مانند coefficient و Pvalue و standard error نمایش داده شده است.

مشاده میشود که مقدار Pvalue برای این تمامی ویژگی ها صفر میباشد که نشان میدهد این مدل مناسب میباشد و تمامی متغیر ها در پیشبینی مدل تاثیرگذار هستند.

همچنین از بررسی معیار Coef که برای اکثر ویژگی ها کم میباشد و متغیر AT بر روی این مجموعه داده تاثیر زیادی دارد و نسبت به بقیه ویژگی ها برای پیشبینی مدل مناسبتر میباشد.

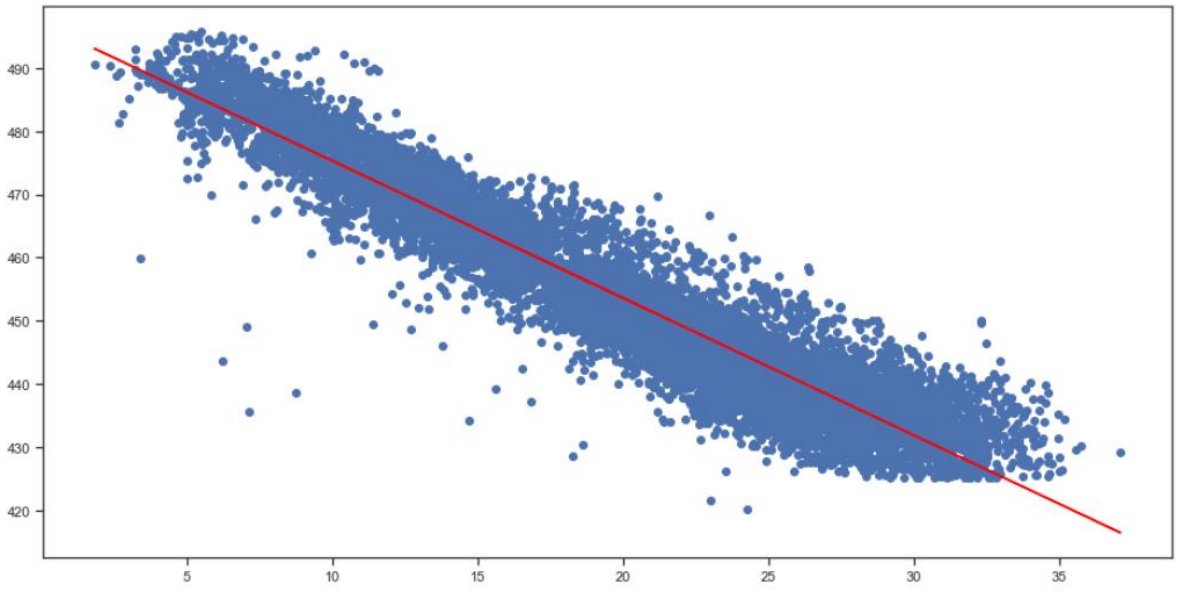
OLS Regression Results

Dep. Variable:	PE	**R-squared:**	0.931
Model:	OLS	**Adj. R-squared:**	0.931
Method:	Least Squares	**F-statistic:**	2.564e+04
Date:	Wed, 06 Jan 2021	**Prob (F-statistic):**	0.00
Time:	18:00:31	**Log-Likelihood:**	-22352.
No. Observations:	7654	**AIC:**	4.471e+04
Df Residuals:	7649	**BIC:**	4.475e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	460.2499	10.790	42.656	0.000	439.099	481.401
AT	-1.9858	0.017	-116.830	0.000	-2.019	-1.952
V	-0.2296	0.008	-28.380	0.000	-0.245	-0.214
AP	0.0565	0.010	5.399	0.000	0.036	0.077
RH	-0.1587	0.005	-34.549	0.000	-0.168	-0.150

Omnibus:	335.497	**Durbin-Watson:**	2.000
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	909.695
Skew:	-0.204	**Prob(JB):**	2.90e-198
Kurtosis:	4.639	**Cond. No.**	2.14e+05

شکل ۱۰



شکل ۱۱

در شکل بالا نمودار پراکندگی این مدل نشان داده شده است.

همچنین در شکل زیر بررسی جامعی از کارایی این مدل آورده شده است.

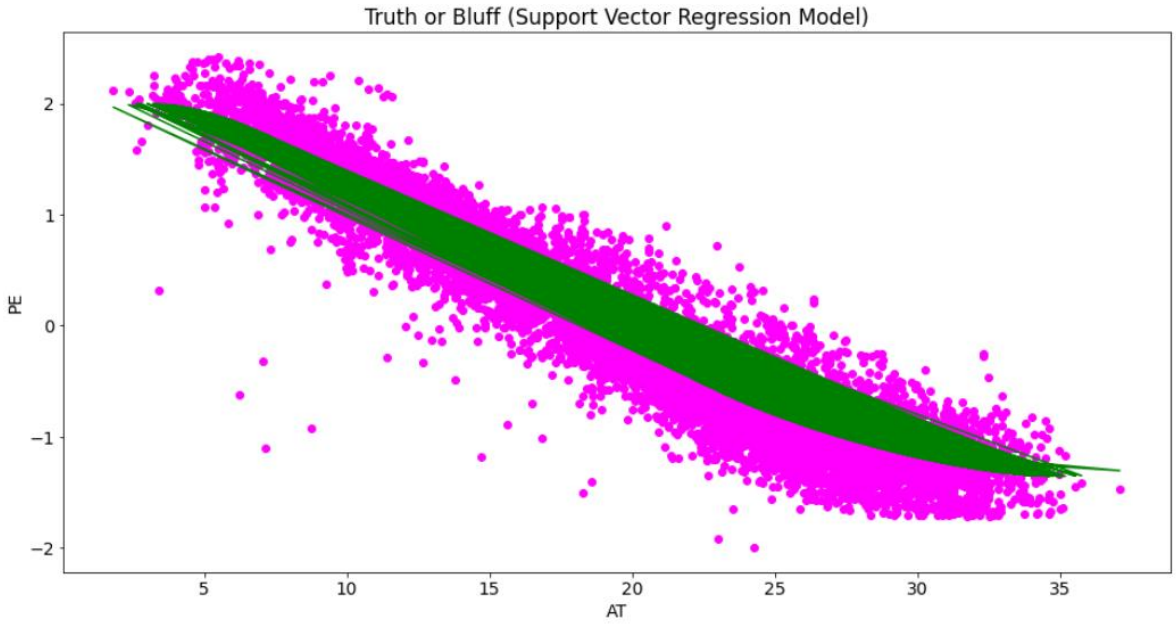
```
Performance of Multiple Linear Regression :
MSE :  10.32
Absloute Error :  9.4875
Correlation :  0.920553
R Square :  0.931
```

شکل ۱۲

SVM

»ماشین بردار پشتیبان (SVM) «یک الگوریتم نظارت‌شده یادگیری ماشین است که هم برای مسائل طبقه‌بندی و هم مسائل رگرسیون قابل استفاده است؛ با این حال از آن بیشتر در مسائل طبقه‌بندی استفاده می‌شود. در الگوریتمSVM ، هر نمونه داده را به عنوان یک نقطه در فضای n- بعدی روی نمودار پراکندگی داده‌ها ترسیم کرده (n تعداد ویژگی‌هایی است که یک نمونه داده دارد) و مقدار هر ویژگی مربوط به داده‌ها، یکی از مؤلفه‌های مختصات نقطه روی نمودار را مشخص می‌کند. سپس، با ترسیم یک خط راست، داده‌های

مختلف و متمایز از یکدیگر را دسته‌بندی می‌کند. بردارهای پشتیبان در واقع مختصات یک مشاهده منفرد هستند. ماشین بردار پشتیبان مرزی است که به بهترین شکل دسته‌های داده‌ها را از یکدیگر جدا می‌کند.



شکل ۱۳

در شکل بالا نمایی از خروجی این مدل به صورت نموداری نشان داده شده است.

```
Performance of SVM :
MSE :  5.732
Absloute Error :  4.564
Correlation :  0.9601
R Square :  0.89966651
```

شکل ۱۴

در شکل بالا آمار کلی از اطلاعات این مدل آورده شده است.

جدول نتیجه گیری

با توجه به خروجی تمامی مدل ها یک جدول کلی از عملکرد کارایی این مدل ها در جدول زیر آورده شده است تا به بررسی جامعی از کارایی این الگوریتم ها بپردازیم.

عملکرد / مدل	MSE	AE	Corr	R2
SLR	۳,۲۲۲	۶,۷۸۴۴۴۵	۰,۸۹۱۲۵۵۴	۰,۸۹۸۹۴۷
MLR	۱۰,۳۲	۹,۴۸۷۵	۰,۹۲۰۵۵۳	۰,۹۳۱
SVM	۵,۷۳۲	۴,۵۶۴	۰,۹۶۰۱	۰,۸۹۹۶۶۶۵۱

همانطور که در جدول بالا مشاهده میشود مدل رگرسیون چندگانه بهترین ضریب R2 را دارد ولی از نظر میانگین استاندارد خطاها نسبت به بقیه مدل ها بایشتر میباشد و همچنین مدل رگرسیون ساده کمترین میانگین استاندارد خطا را دارد اما ضریب R2 و correlation آن نسبت به بقیه کمتر میباشد اما از نظر عملکردی SVM متعادلتر بوده و خروجی های بهتری را داشته است که میتوان آن را بهترین مدل انتخاب نمود.

منابع

۱- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
۲- https://realpython.com/logistic-regression-python/
۳- https://analica.ir/intro-regression-for-managers/
٤- https://www.analyticsvidhya.com/blog/۲۰۲۰/۰۳/polynomial-regression-python/