پیشبینی برق تولیدی پنل های خورشیدی

پروژه ی داده کاوی

اعضای گروه: حدیث غفوری، زهرا قربانی

استاد درس: دکتر حمیدرضا حکیم داودی

فهرست مطالب

- مقدمه
  - نیروگاه خورشیدی چگونه کار می کند؟
  - توضیح مسئله
  - داده های جمع آوری شده
  - نحوه حل مسئله
  - نحوه ارزیابی
  - سوال داده کاوی
- اکتشافات داده ای
- مدل سازی
- ارزیابی و نتیجه گیری
- منابع

فهرست مطالب

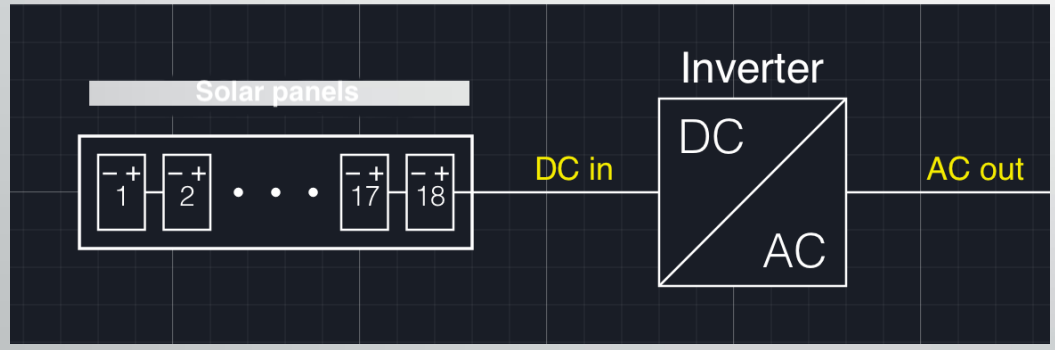
- مقدمه
  - نیروگاه خورشیدی چگونه کار می کند؟
  - توضیح مسئله
  - داده های جمع آوری شده
  - نحوه حل مسئله
  - نحوه ارزیابی
  - سوال داده کاوی
- اکتشافات داده ای
- مدل سازی
- ارزیابی و نتیجه گیری
- منابع

نیروگاه خورشیدی چگونه کار می کند؟

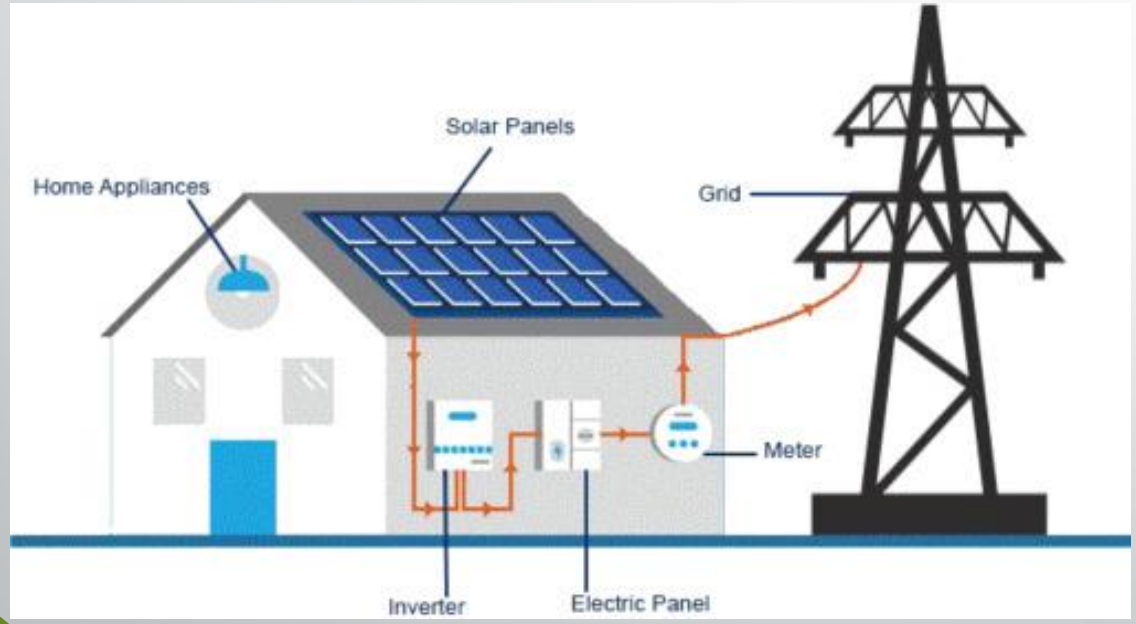
- هنگامی که یک فوتون به سطح سلول فتوولتائیک برخورد می کند، انرژی آن به الکترون های موجود در سلول سیلیکونی منتقل می شود.
- این الکترون ها "تحریک" می شوند و شروع به جریان در مدار می کنند و جریان الکتریکی تولید می کنند.
- یک پنل خورشیدی انرژی جریان مستقیم (DC)تولید می کند.
- سپس، این بر عهده اینورتر است که آن را به جریان متناوب تبدیل کند تا آن را انتقال دهد و در شبکه های توزیع ما استفاده کند.
- در واقع ساختمان های خانگی و صنعتی برای انتقال و استفاده از جریان متناوب طراحی شده اند.

نیروگاه خورشیدی چگونه کار می کند؟

- دو جزء اصلی تشکیل دهنده نیروگاه خورشیدی :
  - ماژول هایی که نور خورشید را به الکتریسیته تبدیل می کنند.
  - یک یا چند اینورتر - دستگاه هایی که جریان مستقیم را به جریان متناوب تبدیل می کنند.
- عملکرد یک نیروگاه خورشیدی وابسته به عوامل:
  - درجه حرارت
  - آلودگی
  - راندمان اینورترها
  - قدمت اینورترها یا پانل ها



توضیح مسئله



داده های جمع اوری شده از نیروگاه خورشیدی

- داده های مربوط به حسگر
- داده های مربوط به اینورترها

سوال مهم و اساسی

- پیشبینی برق تولیدی روز های آینده
- تشخیص پنل ها و تجهیزات معیوب و نیاز به تعمیر (به طور کلی عملکرد نامناسب)

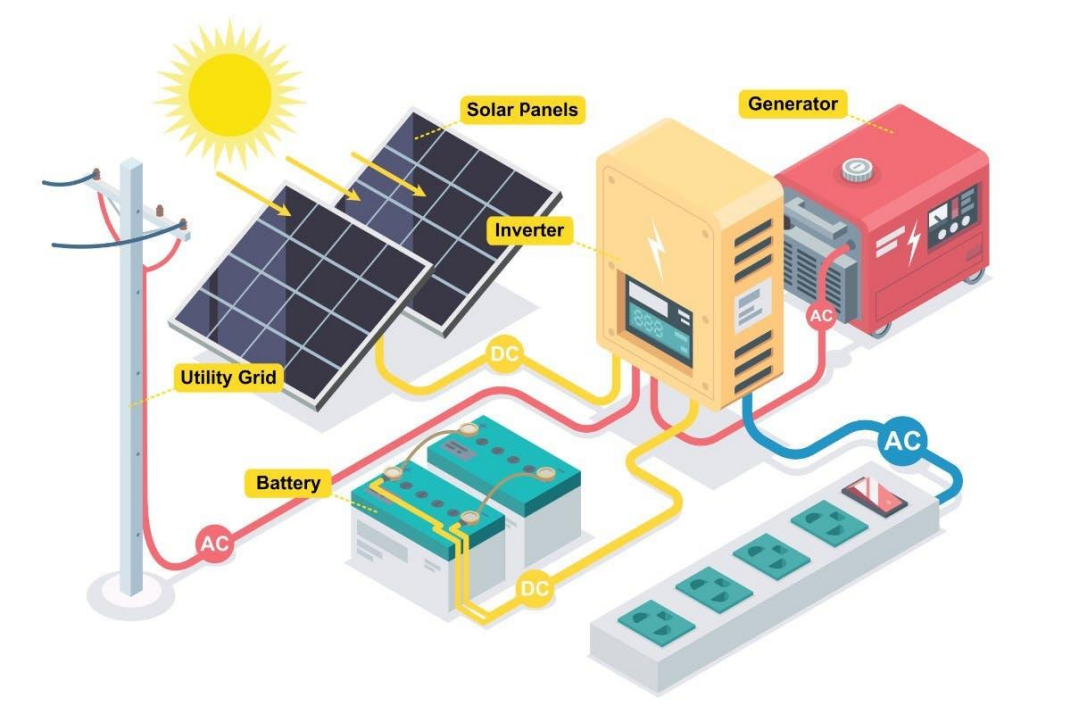
پیشبینی برق تولیدی روز آینده

- مدیریت کارآمد شبکه برق و تجارت برق
- انعطاف بیشتر شبکه های برق و سازگاری با شرایط
- به حداقل رساند اختلالات و مشکلات احتمالی

داده های جمع آوری شده

- دو دسته رکورد در بازه های 15 دققه ای
  - اطلاعات مربوط به آب و هوایی نیروگاه
  - اطلاعات مربوط به برق تولیدی اینورترها
- اطلاعات حسگر
  - شدت تابش
  - دمای ماژول ها
  - دمای محیط

داده های جمع آوری شده



- اطلاعات برق تولیدی اینورترها
  - AC power
  - DC power
  - Daily Yield
  - Total Yield
- بررسی دیتاست
  - عدم وجود مقادیر Null
  - عدم وجود رکورد های Duplicate

نحوه حل مسئله

- پیشبینی برق تولیدی هر 15 دقیقه برای هر اینورتر
  - با توجه به شرایط آب و هوایی و برق تولیدی روز های قبل
  - عدم اگاهی از شرایط اب و هوایی برای روز های آینده
- استفاده از شرایط آب و هوایی روز های قبل
  - سختی پیشبینی شرایط آب و هوایی روز های آینده
  - ساخت رکورد های جدید با اطلاعات روز های قبل

نحوه حل مسئله

- رکورد های جدید
  - لیبل برق تولیدی امروز
  - فیچر ها، فیچرهای سه روز قبل
- حل مسئله رگرشن
  - آموزش مدل با مدل های مربوط به رگرسیون
  - ارزیابی مدل ها با استفاده از ولیدیشن ست
  - انتخاب بهترین مدل
  - گزارش نتیجه داده های تست به روی مدل نهایی

نحوه ارزیابی

- معیار های ارزیابی متفاوت
  - MSE
  - MAE
  - Accuracy
- استفاده از کراس ولیدیشن
  - K-Fold

سوال داده کاوی

پیش بینی مقدار AC POWERتولیدی یک روز آینده بر حسب کیلو وات

پس یک مسئله رگرشن داریم

داده ها مربوط به هر اینورتر و در هر 15 دقیقه ای از روز است، ما برق تولیدی برای هر اینورتر در هر 15 دقیقه از روز را پیشبینی میکنیم و در نهایت با استفاده از مجموع آنها برق تولیدی کل روز را بیان میکنیم.

فهرست مطالب

- مقدمه
  - توضیح مسئله
  - داده های جمع آوری شده
  - نحوه حل مسئله
  - نحوه ارزیابی
  - سوال داده کاوی
- اکتشافات داده ای
- مدل سازی
- ارزیابی و نتیجه گیری
- منابع

اکتشافات داده ای

```
number of null value:
DATE_TIME      0
PLANT_ID       0
SOURCE_KEY     0
DC_POWER       0
AC_POWER       0
DAILY_YIELD    0
TOTAL_YIELD    0
dtype: int64
```

```
DATE_TIME :     3158
PLANT_ID :     1
SOURCE_KEY :     22
DC_POWER :     32909
AC_POWER :     32686
DAILY_YIELD :     29900
TOTAL_YIELD :     37267
```

- بررسی کلی داده های جنریشن
  - بررسی مقادیر نال
  - بررسی تعداد مقادیر یکتا
  - بررسی داپلیکیت رکورد
  - نتیجه: پلنت آیدی یکتا است پس همه ی رکوردها مربوط به یک نیروگاه است و تعداد اینورتر ها 22 تا است.

اکتشافات داده ای(ادامه)

```
number of null value:
DATE_TIME                0
PLANT_ID                 0
SOURCE_KEY               0
AMBIENT_TEMPERATURE      0
MODULE_TEMPERATURE       0
IRRADIATION              0
dtype: int64
```

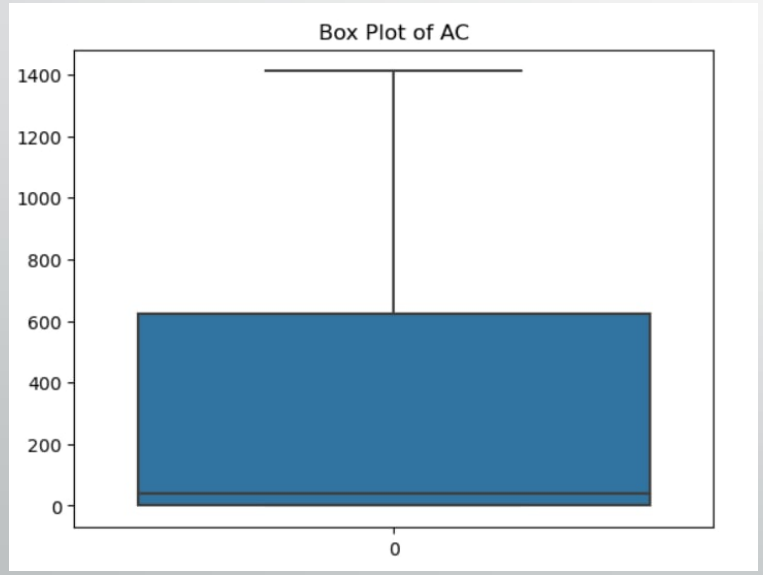
```
DATE_TIME :     3182

PLANT_ID :     1

SOURCE_KEY :     1

AMBIENT_TEMPERATURE :     3182

MODULE_TEMPERATURE :     3182

IRRADIATION :     1758
```

- بررسی کلی داده های سنسور
  - مقادیر نال
  - داپلیکیت رکورد
  - مقادیر یکتا
  - نتیجه : سورس کی و پلنت آیدی یکتاست پس در سطح کل نیروگاه یه سنسور وجود داشته

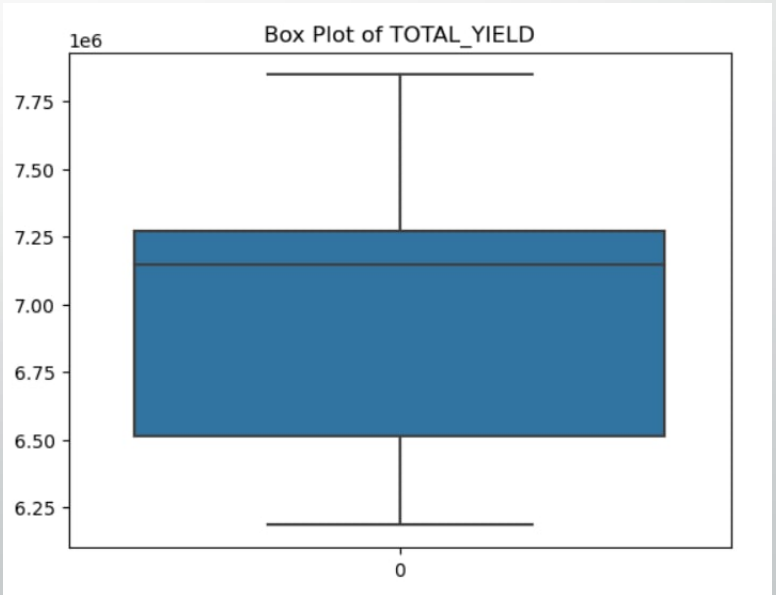
اکتشافات داده ای(ادامه)

- شناسایی داده های پرت(outlier)
- دو روش Z-Score و IQR

AC_POWER

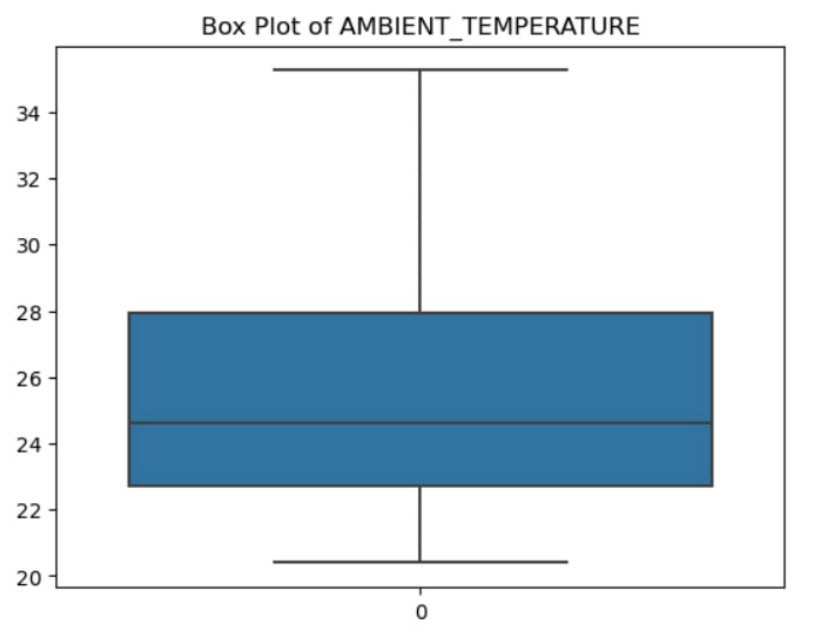


TOTAL_YIELD

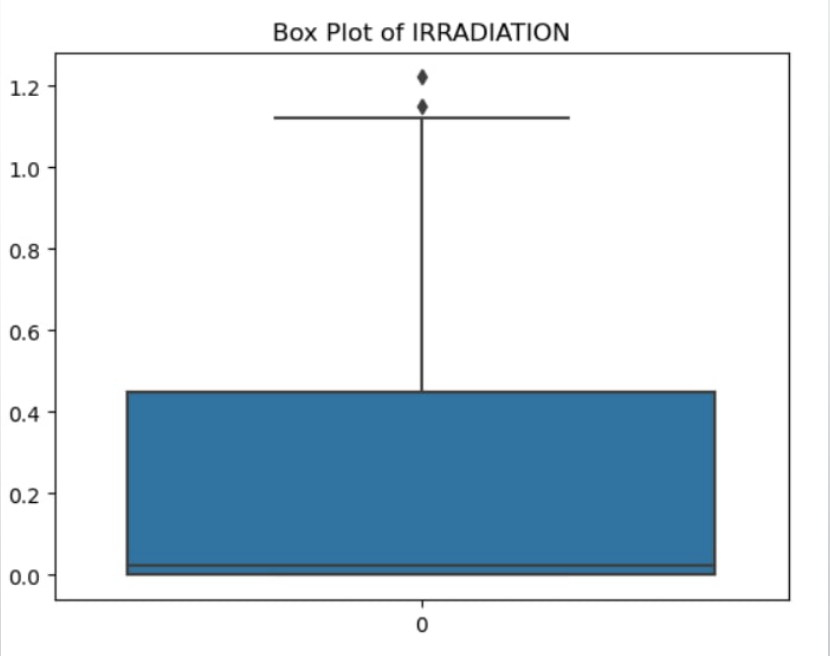


اکتشافات داده ای(ادامه)

TEMPERATURE



IRRADIATION

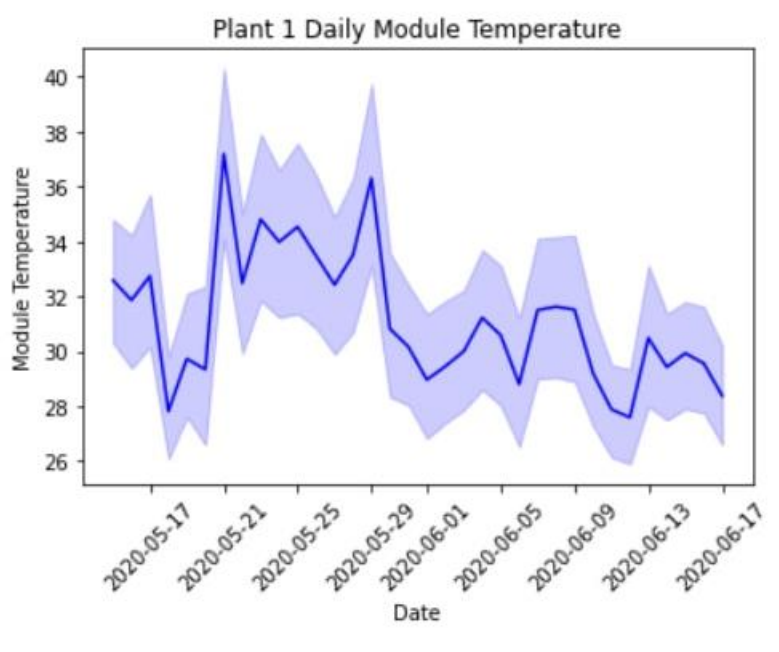


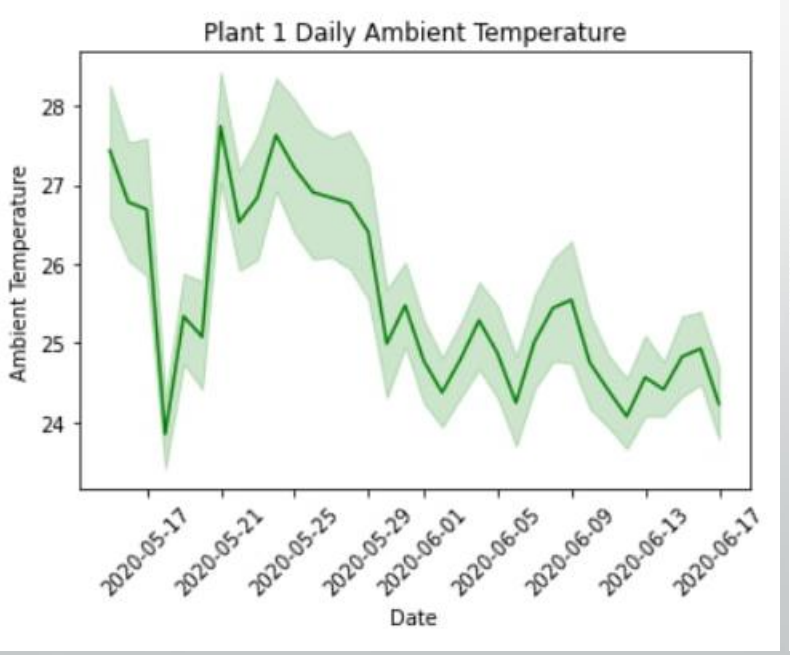
نتیجه: فقط فیلد شدت تابش دو Outlier دارد

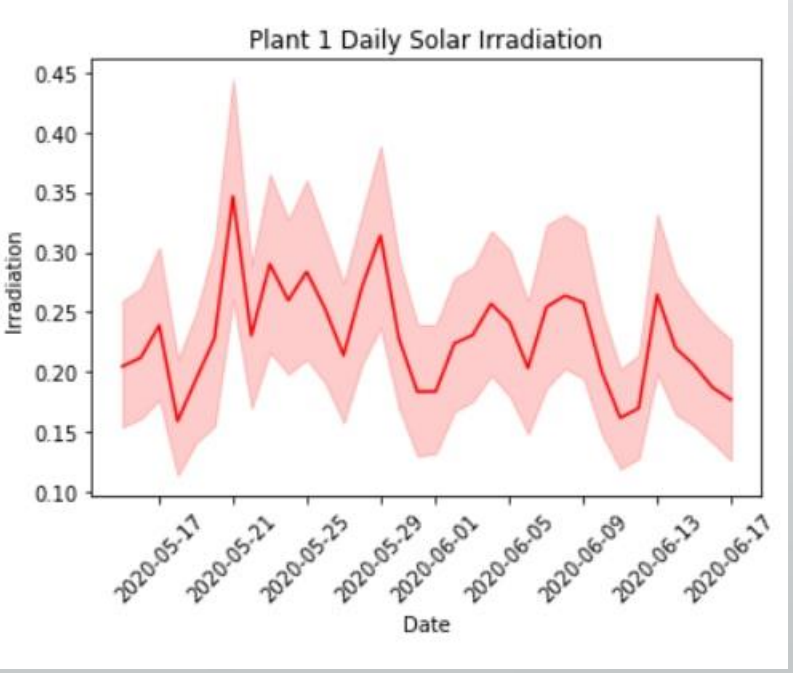
اکتشافات داده ای(ادامه)

"Presence of sunlight" is dictated by the intensity of sunlight and the wavelength of sunlight that hits the PV cells.

- EDA





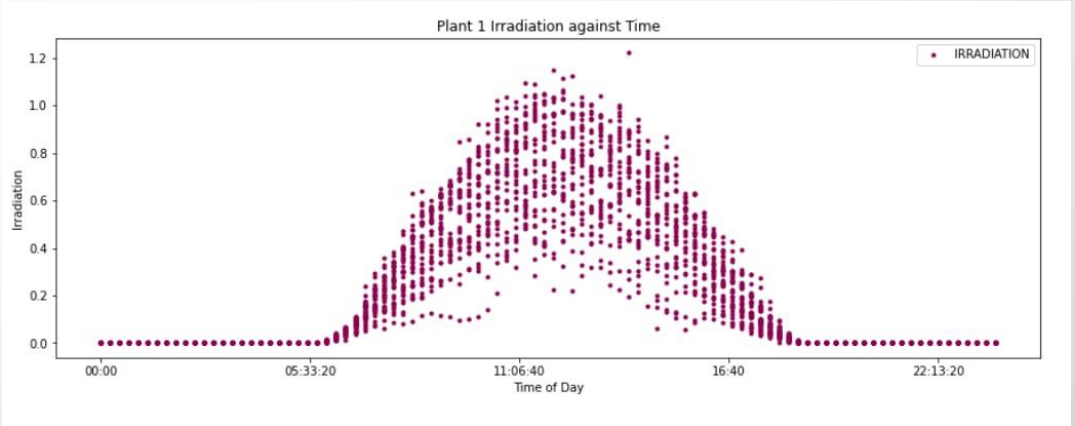


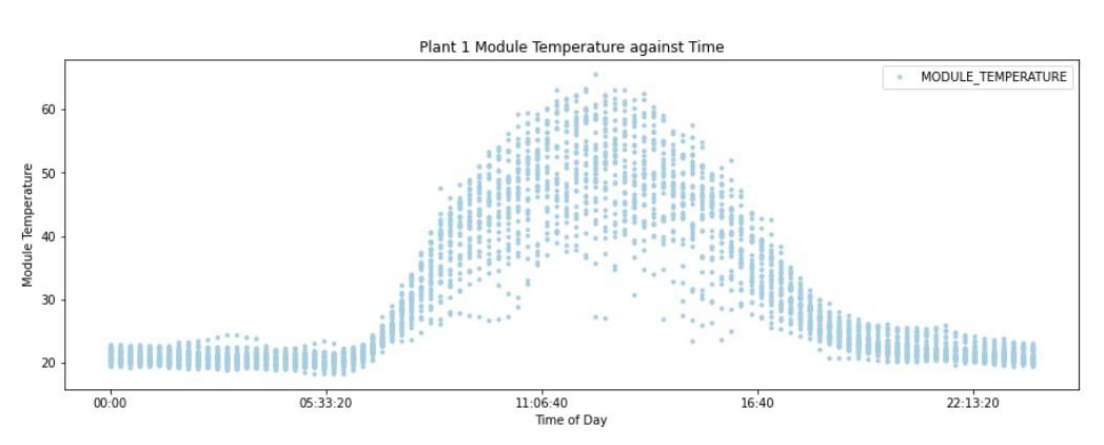
اکتشافات داده ای(ادامه)

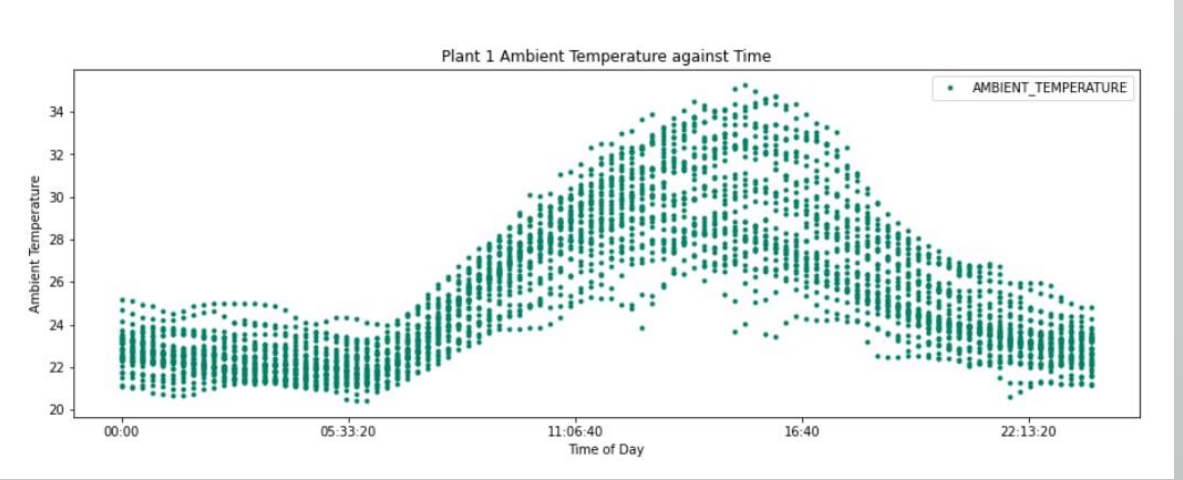
EDA

Power output is generated with the presence of sunlight, which starts at around 0540hrs and ends at around 1800hrs

This means that even though there may still be sunlight at 1800hrs, they are diffused sunlight and scattered sunlight that do not have adequate range of wavelength for power generation.







اکتشافات داده ای(ادامه)

داده های جنریشن

DATE	
2020-01-06	2112
2020-02-06	2088
2020-03-06	2094
2020-04-06	2052
2020-05-06	2080
2020-05-15	1954
2020-05-16	1934
2020-05-17	2112
2020-05-18	2112
2020-05-19	1990
2020-05-20	1672
2020-05-21	1368
2020-05-22	2028
2020-05-23	1958
2020-05-24	2112
2020-05-25	2060
2020-05-26	2072
2020-05-27	2052
2020-05-28	1980
2020-05-29	1490
2020-05-30	2112
2020-05-31	2106
2020-06-06	2052
2020-06-13	2112
2020-06-14	2112
2020-06-15	2112
2020-06-16	2112
2020-06-17	2068
2020-07-06	2112
2020-08-06	2112
2020-09-06	2112
2020-10-06	2112
2020-11-06	2112
2020-12-06	2112

داده های سنسور

DATE	
2020-05-15	93
2020-05-16	88
2020-05-17	96
2020-05-18	96
2020-05-19	93
2020-05-20	80
2020-05-21	68
2020-05-22	96
2020-05-23	90
2020-05-24	96
2020-05-25	96
2020-05-26	96
2020-05-27	96
2020-05-28	96
2020-05-29	79
2020-05-30	96
2020-05-31	96
2020-06-01	96
2020-06-02	96
2020-06-03	95
2020-06-04	96
2020-06-05	96
2020-06-06	96
2020-06-07	96
2020-06-08	96
2020-06-09	96
2020-06-10	96
2020-06-11	96
2020-06-12	96
2020-06-13	96
2020-06-14	96
2020-06-15	96
2020-06-16	96
2020-06-17	96

- بررسی دقیقتر داده ها
  - Group Byروی داده ها بر اساس تاریخ
  - تعداد رکورد های هر گروه برای داده های سنسور باید 24 * 4 = 96 باشد.
  - تعداد رکورد های هر گروه برای داده های جنریشن باید 24 * 4 * 22 = 2112 باشد.

اکتشافات داده ای(ادامه)

- مشکلات
  - در بعضی از 15 دقیقه ها داده های سنسور ثبت نشده است
  - در بعضی از 15 دقیقه ها داده های بعضی از اینورتر ها ثبت نشده است
  - مشکل بزرگتر: فرمت تاریخ در داده های جنریشن
- بررسی دقیقتر تاریخ ها
  - بازه 34 روزه - شروع 15 ماه پنجم و پایان 17 ماه ششم
  - تاریخ در فرمت استرینگ
  - تبدیل به فرمت ابجکت
  - اصلاح مشکل مربوط به جابجایی ماه و روز

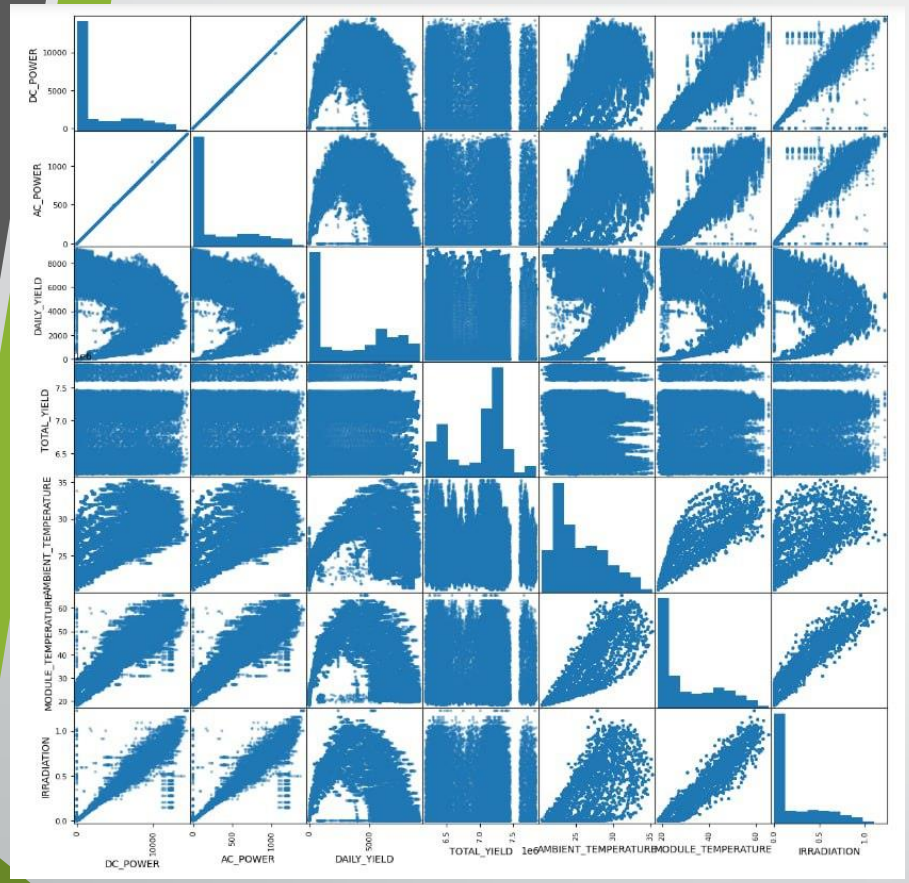
نتیجه بعد از اصلاح تاریخ ها

2020-05-15	1954
2020-05-16	1934
2020-05-17	2112
2020-05-18	2112
2020-05-19	1990
2020-05-20	1672
2020-05-21	1368
2020-05-22	2028
2020-05-23	1958
2020-05-24	2112
2020-05-25	2060
2020-05-26	2072
2020-05-27	2052
2020-05-28	1980
2020-05-29	1490
2020-05-30	2112
2020-05-31	2106
2020-06-01	2112
2020-06-02	2088
2020-06-03	2094
2020-06-04	2052
2020-06-05	2080
2020-06-06	2052
2020-06-07	2112
2020-06-08	2112
2020-06-09	2112
2020-06-10	2112
2020-06-11	2112
2020-06-12	2112
2020-06-13	2112
2020-06-14	2112
2020-06-15	2112
2020-06-16	2112
2020-06-17	2068

اکتشافات داده ای(ادامه)

- بررسی عدم وجود رکورد ها
  - در داده های سنسور عدم وجود رکورد برای 82 بازه 15 دقیقه
  - در داده های جنریشن عدم وجود رکورد برا 207 بازه 15 دقیقه و اینورتر
  - اضافه کردن رکورد های جدید
  - مرج دو دسته دیتا بر اساس ستون دیت تایم

اکتشافات داده ای(ادامه)



- بررسی کورلیشن بین ستونها

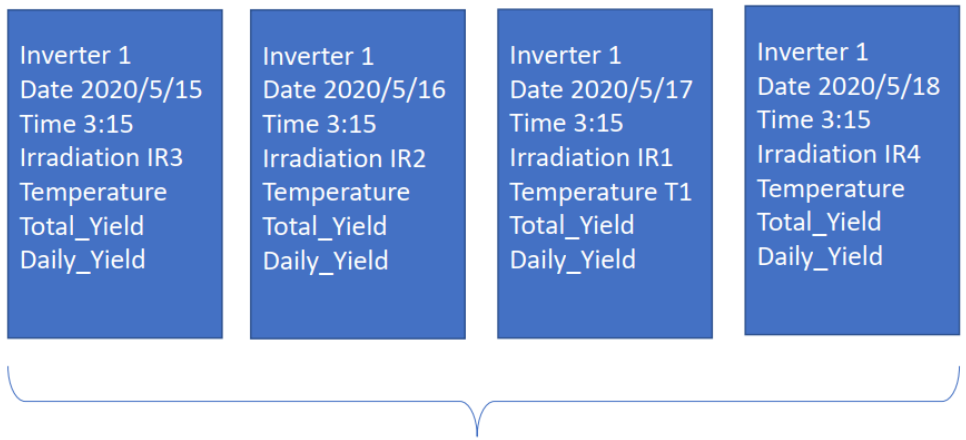
	DC_POWER	AC_POWER	DAILY_YIELD	TOTAL_YIELD	AMBIENT_TEMPERATURE	MODULE_TEMPERATURE	IRRADIATION
DC_POWER	1.000000	0.999996	0.054778	0.006352	0.729097	0.952115	0.983778
AC_POWER	0.999996	1.000000	0.054685	0.006348	0.729330	0.952260	0.983773
DAILY_YIELD	0.054778	0.054685	1.000000	0.003920	0.426140	0.160042	0.048180
TOTAL_YIELD	0.006352	0.006348	0.003920	1.000000	-0.034470	-0.012157	-0.001855
AMBIENT_TEMPERATURE	0.729097	0.729330	0.426140	-0.034470	1.000000	0.856131	0.726685
MODULE_TEMPERATURE	0.952115	0.952260	0.160042	-0.012157	0.856131	1.000000	0.962001
IRRADIATION	0.983778	0.983773	0.048180	-0.001855	0.726685	0.962001	1.000000

نتیجه
کورلیشن بین شدت تابش و دمای ماژول خیلی بالاست

همچنین ای سی پاور و دی سی پاور هم کورلیشن زیادی دارند
(انتخاب اس سی پاور به عنوان لیبل)

کورلیشن شدت تابش و ای سی پاور نیز زیاد است
که نشان دهنده رابطه خطی مثبت بین شدت تابش و ای سی پاور است

اکتشافات داده ای(ادامه)



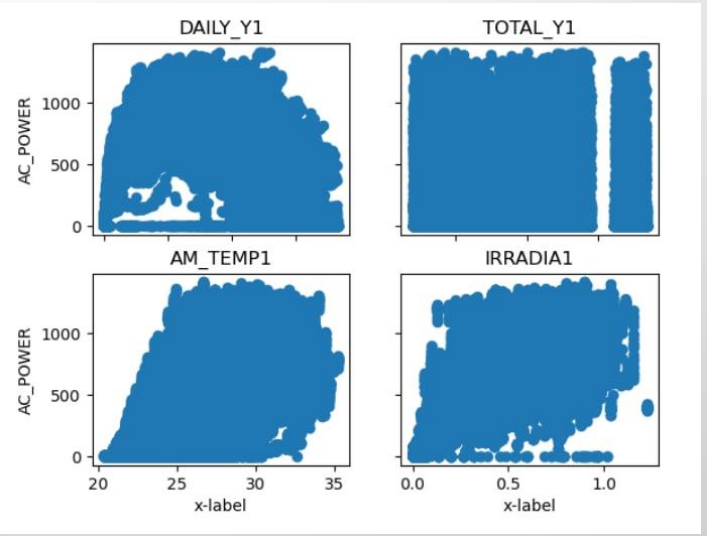
id	DATE	Time	Irradiation1	Irradiation2	Irradiation3	Temp1	...
Inverter 1	2020/5/18	3:15	IR1	IR2	IR3	T1	...

- ساخت رکوردها
  - استفاده از پنجره های 4 تایی
  - هر پنجره شامل رکورد های یک 15 دقیقه خاص برای یک اینورتر خاص در چهار روز متوالی است.
  - لیبل ای سی پاور روز چهارم و همه فیچر های سه روز اول به عنوان فیچر رکورد جدید استفاده میشوند.

اکتشافات داده ای(ادامه)

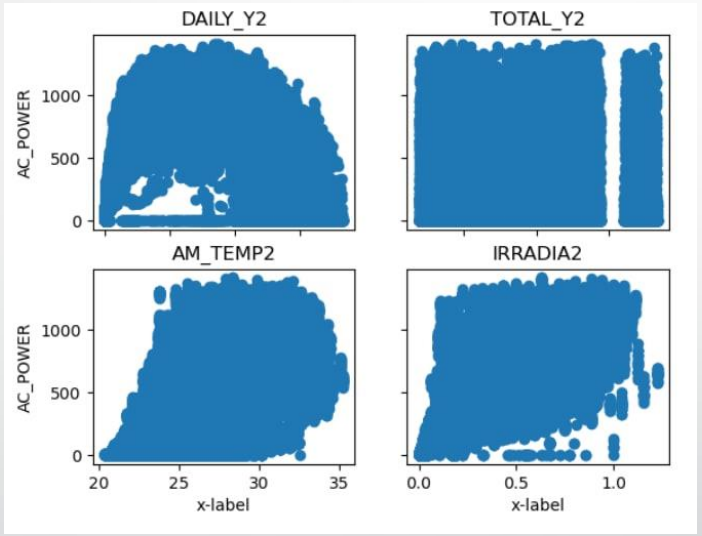
- بررسی روابط بین فیچر ها و متغیر هدف

فیچر های مربوط به روز قبل



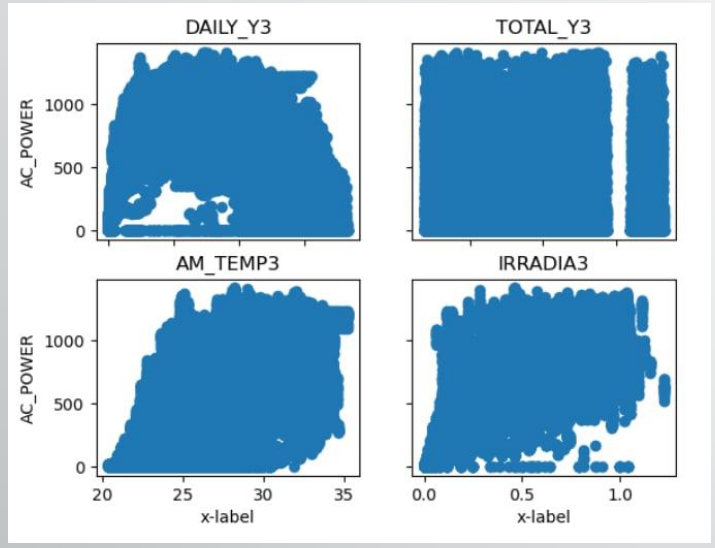
(1)

فیچر های مربوط به 2 روز قبل



(2)

فیچر های مربوط به 3 روز قبل



(3)

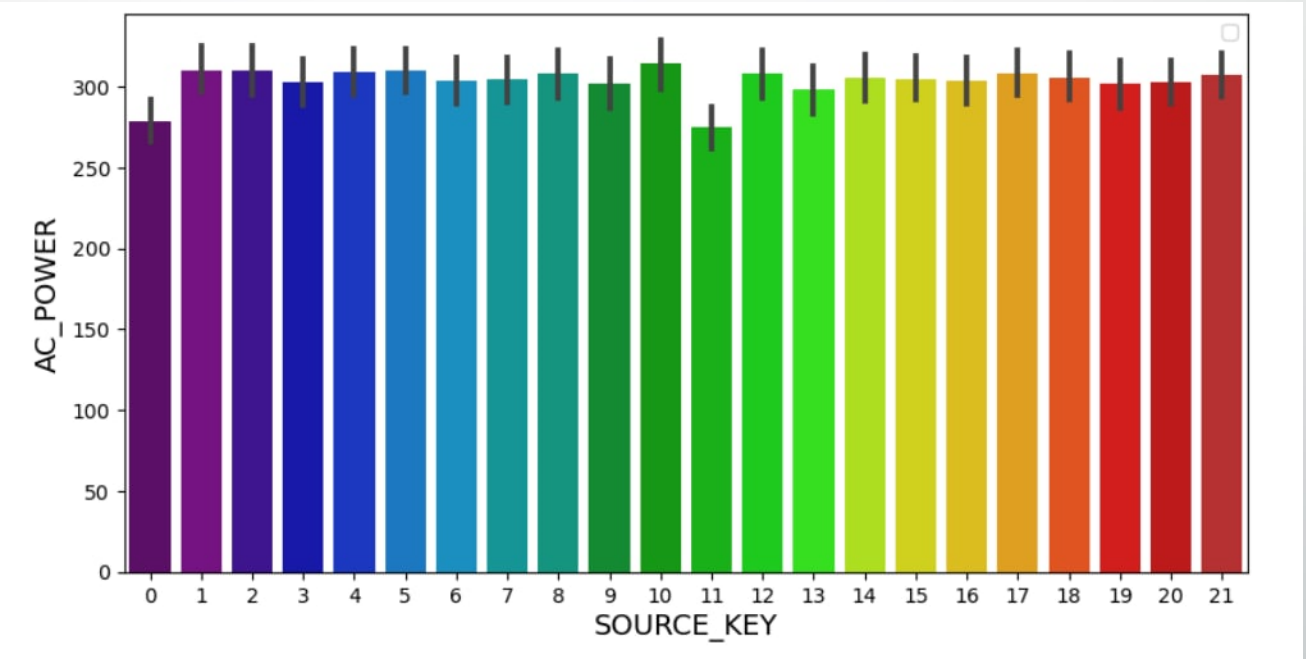
نتیجه:

1.شدت تابش و دمای محیط با AC_POWER رابطه تقریبا خطی دارند.

2.رابطه DAYLY_YEALDو AC_POWER تقریبا به شکل سهمی است.

اکتشافات داده ای(ادامه)

- بررسی روابط بین فیچر ها و متغیر هدف

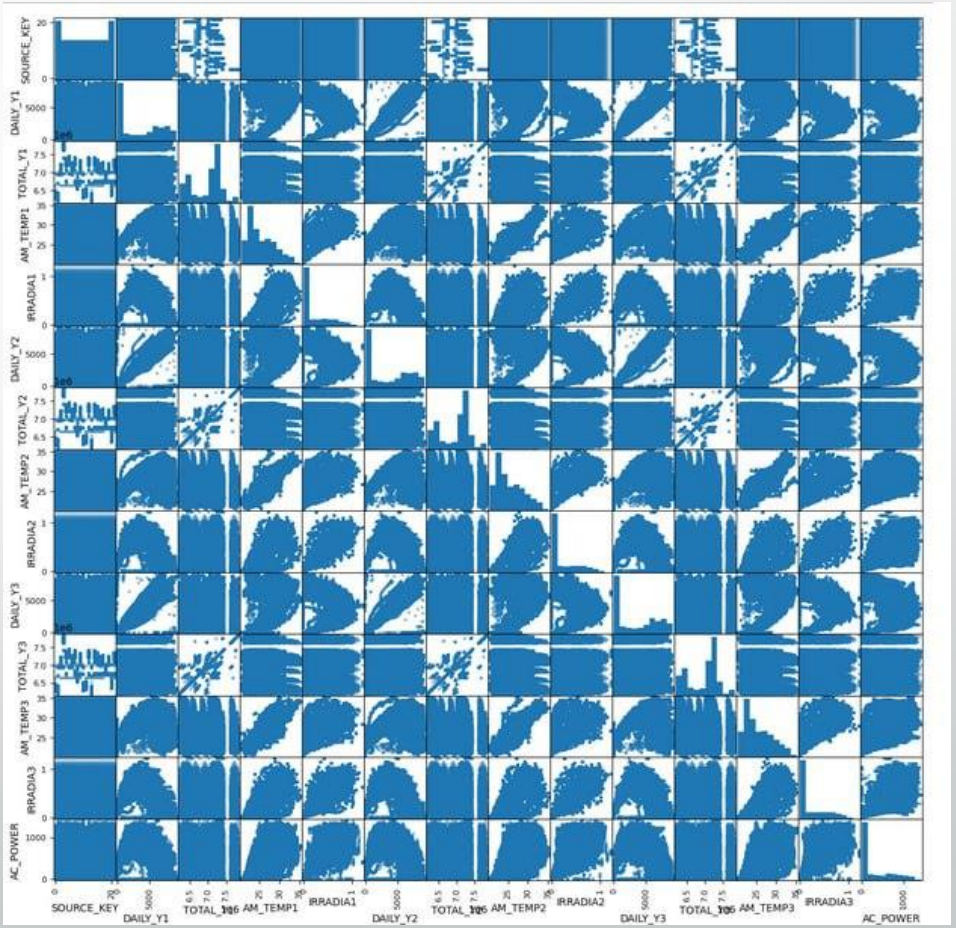


برق تولیدی هر INVERTER

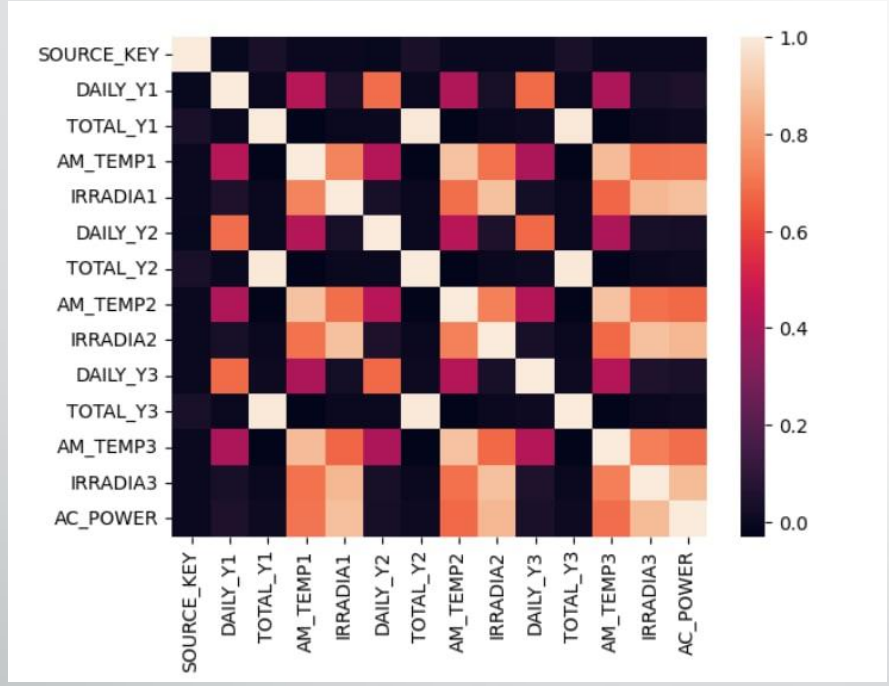
نتیجه :
دو Inverter صفر و ده برق کمتری نسبت به بقیه inverterها تولید کرده اند.
این میتواند دلیلی بر مشکل داشتن آنها باشد

اکتشافات داده ای(ادامه)

- بررسی دوباره کورلیشن بین فیچر ها



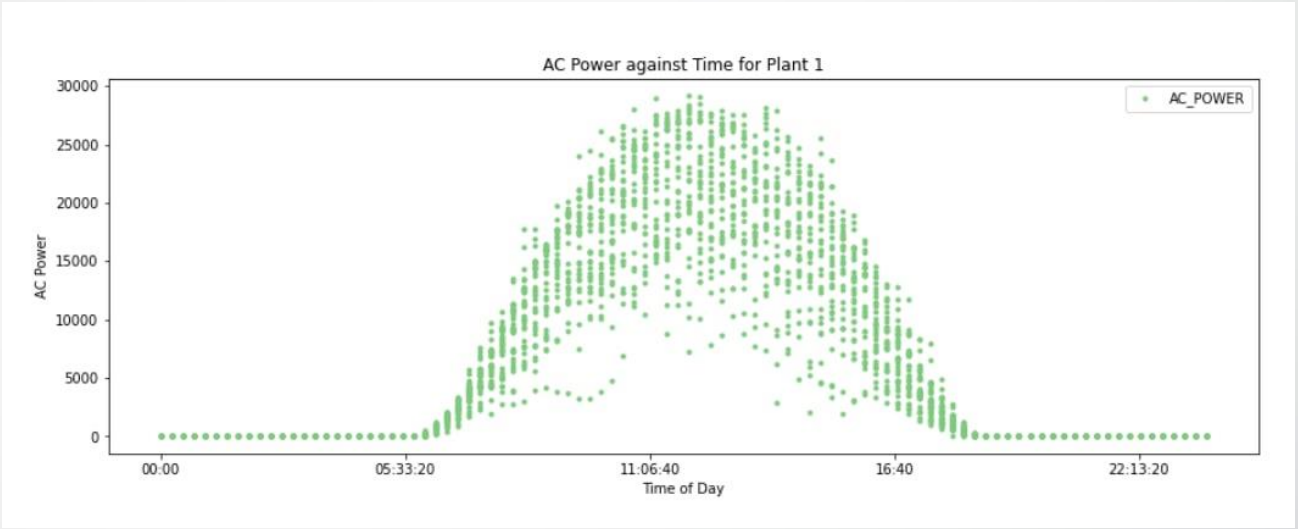
بررسی دو به دو متغیر ها

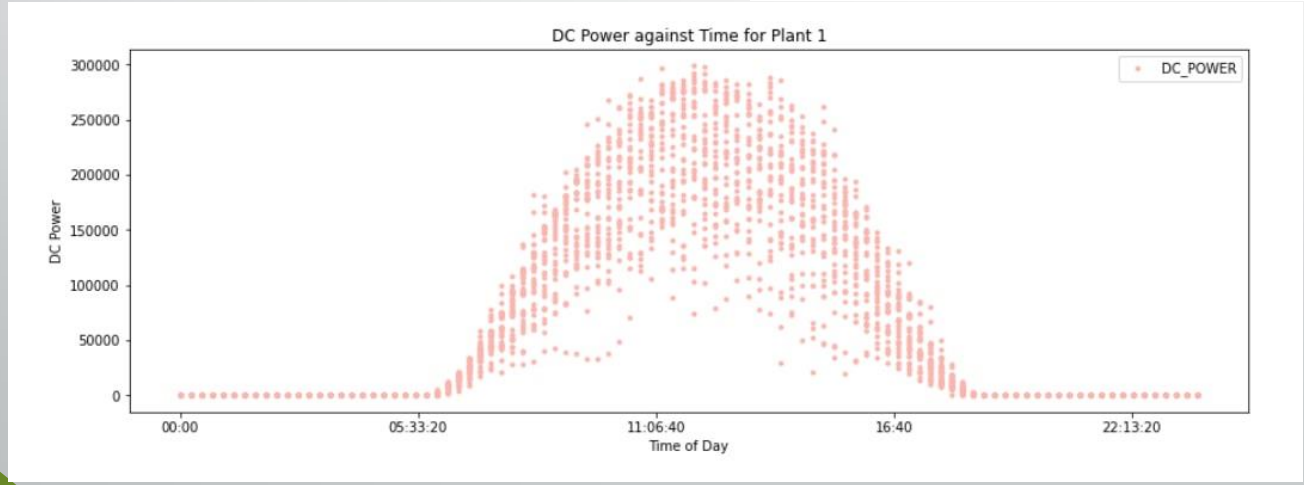


نتیجه :
زیاد بودن شباهت بین متغیر های TOTAL_YEALDسه روز

اکتشافات داده ای(ادامه)

نتیجه
ای سی پاور و دی سی پاور توزیع یکسانی دارند و تفاوت آنها در مقیاسشان است





فهرست مطالب

- مقدمه
  - توضیح مسئله
  - داده های جمع آوری شده
  - نحوه حل مسئله
  - نحوه ارزیابی
  - سوال داده کاوی
- اکتشافات داده ای
- مدل سازی
- ارزیابی و نتیجه گیری
- منابع

مدل سازی

- تقسیم داده ها به سه بخش Testو Train وValidation
  - 70درصد داده ها برای Train
  - 15درصد داده ها برای Validation
  - 15درصد برای Test
- مقایسه روش scale مناسب
  - استفاده از روش MinmaxScaler
  - استفاده از روش StandardScaler
  - آموزش هر دسته از داده ها به روی مدل
  - پیشبینی داده های Validation
  - مقایسه MSE هر دو مدل برای داده های Validation

انتخاب Scalerمناسب

```
Validation MSE: 2221.5986
r2_score Validation: 0.99
Training Accuracy: 0.998
Validation Accuracy: 0.986

StandardScaler Validation MSE: 2238.6894
StandardScaler r2_score Validation: 0.99
StandardScaler Training Accuracy: 0.998
StandardScaler Validation Accuracy: 0.986

MinMaxScaler Validation MSE: 2160.7444
MinMaxScaler r2_score Validation: 0.99
MinMaxScaler Training Accuracy: 0.998
MinMaxScaler Validation Accuracy: 0.986
```

مدل سازی(ادامه)

- اضافه کردن فیچر جدید
  - کورلیشن زیاد بیاد توتال یلد روز قبل دو روز قبل و سه روز قبل
  - بررسی انتخاب بین دو مورد
    - 1) انتخاب توتال یلد روز قبل و حذف بقیه
    - 2) اضافه کردن فیچر جدید از میانگین سه توتال یلد و حذف هر سه توتال یلد
  - آموزش مدل با استفاده از دو دسته دیتا
  - پیشبینی داده های ولیدیشن با مدلها
  - بررسی mseمدل روی داده های ولیدیشن و انتخاب مدل

نتایج هردو مورد

```
Mean of Total_Yeild Validation MSE: 2137.5345
Mean of Total_Yeild r2_score Validation: 0.99
Mean of Total_Yeild Training Accuracy: 0.998
Mean of Total_Yeild Validation Accuracy: 0.986

Yesterday Total_Yeild Validation MSE: 2160.7444
Yesterday Total_Yeild r2_score Validation: 0.99
Yesterday Total_Yeild Training Accuracy: 0.998
Yesterday Total_Yeild Validation Accuracy: 0.986
```

با بررسی نتایج هردو مورد میتوان دید که انتخاب روش میانیگن بهتر است

مدل سازی(ادامه)

- پیش مدل
  - انتخاب اسکیلر مین مکس و اضافه کردن فیچر میانگین توتال یلد ها
  - تقسیم داده ها به سه قسمت تست و ترین و ولیدیشن
  - استفاده از کا فولد برای Cross Validation
  - استفاده از dummy regressorبا استراتژی میانگین به عنوان Baseline
  - همه مدل ها بهتر از Baselineعمل میکنند.

عملکرد Baseline

```
MSE Training is 159019.0208
MAE Training is 345.2382
r2_score Training: 0.00
Score Training is 0.0000 %

MSE Validation is 154890.0923
MAE Validation is 341.8530
r2_score Validation: -0.00
Score validation is -0.0201 %
```

مدل سازی(ادامه)

- مدل سازی
- استفاده از الگوریتم های Linear Regression, Decision Tree, Random Forest

نتایج Linear Regression

```
MSE Training is 26977.7008
MAE Training is 95.6362
r2_score Training: 0.83
Score Training is 83.0349 %

MSE Validation is 26942.6367
MAE Validation is 95.2451
r2_score Validation: 0.83
Score validation is 82.6018 %
```

نتایج Decision Tree

```
MSE Training is 0.0000
MAE Training is 0.0000
r2_score Training: 1.00
Score Training is 100.0000 %

MSE Validation is 3791.7975
MAE Validation is 18.3424
r2_score Validation: 0.98
Score validation is 97.5515 %
```

نتایج Random Forest

```
MSE Training is 275.2243
MAE Training is 5.3911
r2_score Training: 1.00
Score Training is 99.8269 %

MSE Validation is 2175.5750
MAE Validation is 14.8534
r2_score Validation: 0.99
Score validation is 98.5951 %
```

مدل سازی(ادامه)

- استفاده از K-Foldبرای ارزیابی بهتر

KFold-Linear Regression

```
Fold 1: Training MSE = 26946.22, Validation MSE = 26645.43
Fold 2: Training MSE = 26986.03, Validation MSE = 26578.84
Fold 3: Training MSE = 26582.03, Validation MSE = 27376.85

Average Training MSE = 26838.09, Average Validation MSE = 26867.04
```

KFold- Decision Tree

```
Fold 1: Training MSE = 0.00, Validation MSE = 3517.69
Fold 2: Training MSE = 0.00, Validation MSE = 3729.86
Fold 3: Training MSE = 0.00, Validation MSE = 3670.57

Average Training MSE = 0.00, Average Validation MSE = 3639.37
```

KFold- Random Forest

```
Fold 1: Training MSE = 268.28, Validation MSE = 1996.86
Fold 2: Training MSE = 273.87, Validation MSE = 1970.95
Fold 3: Training MSE = 272.35, Validation MSE = 1977.44

Average Training MSE = 271.50, Average Validation MSE = 1981.75
```

مدل سازی(ادامه)

- تنظیم hyper parameterها
  - Decision Tree

تنظیم پارامترهای Decision Tree

```
{'max_depth': 40,
 'min_samples_leaf': 1,
 'min_samples_split': 10,
 'splitter': 'random'}
```

نتایج بهترین مدل Decision Tree

```
MSE Training is 1084.0477
MAE Training is 10.2390
r2_score Training: 0.99
Score Training is 99.3183 %

MSE Validation is 2995.7469
MAE Validation is 17.0125
r2_score Validation: 0.98
Score validation is 98.0655 %
```

فهرست مطالب

- مقدمه
  - توضیح مسئله
  - داده های جمع آوری شده
  - نحوه حل مسئله
  - نحوه ارزیابی
  - سوال داده کاوی
- اکتشافات داده ای
- مدل سازی
- ارزیابی و نتیجه گیری
- منابع

ارزیابی و نتیجه گیری

- معیار های ارزیابی
  - معیار Accuracy
  - معیار MSE
  - معیار MAE
  - معیار R2 score

ارزیابی و نتیجه گیری

- مقایسه مدل ها
  - بهترین نتایج روی Validation: مدل Random Forest
  - بعد از آن، مدل Decision Tree
  - مدل درخت بدون هایپرپارامتر روی داده های ترین فیت میشود.
  - مدل Linear Regressionعملکرد ضعیفی نسبت به دو مدل قبلی دارد که نشان دهنده ی غیرخطی بودن داده ها است.

نتایج بهترین مدل بر داده های تست

```
MSE Test is 1656.7831
MAE Test is 14.0781
r2_score Test: 0.99
Score Test is 98.9712 %
```

مراجع

- https://www.kaggle.com/datasets/anikannal/solar-power-generation-data
- https://www.kaggle.com/code/shumaylasmawi/solar-power-plant-eda-and-output-prediction
- https://www.kaggle.com/code/virosky/how-to-manage-a-solar-power-plant