# به نام آفریدگار

## پروژه ، پیش‌بینی تقاضای بار برق با استفاده از داده‌های تاریخی بار واقعی و پیش‌بینی‌شده کشور لهستان

**Energy Load Project**

ENTSO-E Transparency Platform - Load Year Ahead

**با هدف** ؛ بررسی امکان‌پذیری پیش‌بینی تقاضای هفتگی برق با استفاده از داده‌های تاریخی و ارزیابی عملکرد مدل‌های مختلف در بازه چندساله.

## مقدمه

انرژی الکتریکی یکی از حیاتی‌ترین زیرساخت‌های هر کشور است و مدیریت صحیح تولید، توزیع و مصرف آن نیازمند پیش‌بینی دقیق تقاضا در بازه‌های زمانی مختلف است. هرچه دقت پیش‌بینی بار بیشتر باشد، تصمیم‌گیری در زمینه برنامه‌ریزی تولید، مدیریت شبکه، کاهش هزینه‌های عملیاتی و افزایش پایداری سیستم قدرت به شکل مؤثرتری انجام می‌شود.

در این پروژه، مجموعه‌ای از داده‌های هفتگی مرتبط با بار پیش‌بینی‌شده و بار واقعی برق مورد بررسی قرار می‌گیرد. این داده‌ها از پلتفرم ENTSO-E Transparency Platform استخراج شده‌اند و شامل اطلاعاتی درباره حداقل و حداکثر بار واقعی و پیش‌بینی‌شده در طول هفته‌های مختلف سال هستند. هدف اصلی پروژه آن است که ابتدا این داده‌ها در یک ساختار استاندارد و قابل اتکا تجمیع و پاک‌سازی شوند، سپس با انجام تحلیل اکتشافی، استخراج ویژگی و مدل‌سازی، امکان پیش‌بینی بهتر مصرف برق بررسی شود.

## بیان مسئله

پیش‌بینی تقاضای برق یک مسئله مهم در حوزه تحلیل داده و سیستم‌های انرژی است. هرچند سازمان‌ها و اپراتورهای شبکه از قبل مقادیر پیش‌بینی‌شده‌ای برای بار در اختیار دارند، اما همواره بین مقدار پیش‌بینی‌شده و مقدار واقعی اختلاف وجود دارد. این اختلاف می‌تواند ناشی از عوامل مختلفی مانند رفتار مصرف‌کنندگان، شرایط آب‌وهوایی، تغییرات فصلی، رویدادهای خاص و محدودیت‌های مدل‌های فعلی پیش‌بینی باشد.

**بنابراین، بررسی این موضوع اهمیت دارد که:**

- دقت پیش‌بینی در سال‌های مختلف چگونه تغییر کرده است،
- خطاها در چه بازه‌هایی بیشتر هستند،
- آیا الگوهای فصلی و هفتگی معناداری در داده‌ها وجود دارد،
- و آیا می‌توان با استفاده از روش‌های داده‌محور، مدل‌هایی ساخت که عملکرد بهتری نسبت به پیش‌بینی‌های موجود ارائه دهند.

## اهداف پروژه

**اهداف اصلی این پروژه عبارت‌اند از:**

- تجمیع و پاک‌سازی داده‌های چندساله بار برق
- طراحی یک پایگاه داده SQLite تمیز و منسجم
- تحلیل روندهای سالانه و هفتگی
- بررسی اختلاف بین مقادیر واقعی و پیش‌بینی‌شده
- ساخت ویژگی‌های مناسب برای مدل‌سازی
- مقایسه چند مدل پایه و یادگیری ماشین و سری زمانی
- ارائه نتایج در قالب گزارش نهایی و داشبورد

## معماری پروژه

بر اساس فایل معماری ارائه‌شده، ساختار پروژه شامل پوشه‌های داده خام، داده پردازش‌شده، نوت‌بوک‌ها، پایگاه داده، داشبورد، کدهای منبع، و گزارش نهایی است {معماری_داده.md } این معماری باعث می‌شود فرآیند توسعه پروژه منظم، قابل ردیابی، و مناسب برای ارائه دانشگاهی و گیت‌هاب باشد.

## یکپارچه‌سازی، پاک‌سازی و ذخیره‌سازی داده‌ها

در نخستین فاز اجرایی پروژه، تمرکز اصلی بر آماده‌سازی زیرساخت داده‌ای قرار گرفت. از آنجا که تمام مراحل بعدی از جمله تحلیل اکتشافی، استخراج ویژگی، مدل‌سازی و طراحی داشبورد به کیفیت و انسجام داده وابسته هستند، ابتدا فرایند تجمیع و پاک‌سازی داده‌ها انجام شد.

فایل‌های ورودی این پروژه به‌صورت CSV و با ساختاری شامل ستون‌های مربوط به بازه زمانی، شماره هفته، ناحیه جغرافیایی، حداقل بار پیش‌بینی‌شده، حداقل بار واقعی، حداکثر بار واقعی و حداکثر بار پیش‌بینی‌شده هستند [GUI_TOTAL_LOAD_YEARAHEAD_201812312300-201912312300.csv:1]. با توجه به اینکه ستون‌ها در حالت خام برای تحلیل مستقیم بهینه نبودند، مجموعه‌ای از تبدیل‌ها و استانداردسازی‌ها روی داده اعمال شد.

در گام نخست، همه فایل‌های CSV از پوشه داده {raw} خوانده شدند و در یک جدول یکپارچه ادغام شدند. سپس نام ستون‌ها به فرم استاندارد و سازگار با محیط SQL تبدیل شد. همچنین ستون بازه زمانی به دو بخش تاریخ شروع و تاریخ پایان تفکیک شد تا انجام تحلیل‌های زمانی در مراحل بعدی ممکن شود. ستون مربوط به هفته نیز از فرم متنی به یک متغیر عددی تبدیل شد. علاوه بر این، اطلاعات مربوط به ناحیه جغرافیایی به دو مؤلفه نام ناحیه و کد ناحیه شکسته شد.

در مرحله بعد، نوع داده ستون‌های عددی استاندارد شد تا تمام مقادیر بار به‌صورت عددی قابل تحلیل ذخیره شوند. سپس مجموعه‌ای از کنترل‌های کیفیت روی داده اجرا شد که شامل بررسی مقادیر گمشده، شناسایی ردیف‌های تکراری، کنترل بازه معتبر شماره هفته، بررسی صحت بازه‌های تاریخی، و ارزیابی سازگاری بین مقادیر حداقل و حداکثر واقعی و پیش‌بینی‌شده بود.

در پایان این مرحله، داده پاک‌سازی‌شده هم به‌صورت فایل CSV پردازش‌شده و هم به‌صورت جدول SQLite ذخیره شد. انتخاب SQLite به دلیل سبک بودن، سادگی استفاده، و مناسب بودن آن برای پروژه‌های تحلیلی و دانشگاهی انجام شد. این خروجی پایه اصلی مراحل بعدی پروژه را تشکیل می‌دهد.

## جمع‌بندی فاز ۱

خروجی فاز اول یک پایگاه داده تمیز، ساخت‌یافته و آماده تحلیل است که از چند فایل خام CSV تولید می‌شود. این ساختار باعث می‌شود در مراحل بعدی بتوان با اطمینان بیشتر به تحلیل روندها، اندازه‌گیری خطاها، ساخت ویژگی‌ها و آموزش مدل‌ها پرداخت.

## تحلیل اکتشافی داده‌ها

پس از تکمیل مرحله پاک‌سازی و ساخت پایگاه داده، در این فاز تحلیل اکتشافی داده‌ها به‌منظور شناخت بهتر رفتار بار برق و کیفیت پیش‌بینی‌ها انجام شد. هدف اصلی این مرحله، استخراج الگوهای اولیه، شناسایی روندهای سالانه و هفتگی، بررسی میزان اختلاف بین مقادیر واقعی و پیش‌بینی‌شده، و آماده‌سازی ذهنی و تحلیلی برای مراحل بعدی پروژه بود.

در ابتدای تحلیل، داده‌های ذخیره‌شده در پایگاه داده SQLite بازیابی شدند و چند متغیر کمکی برای اندازه‌گیری دامنه بار و خطای پیش‌بینی تعریف شد. از جمله این متغیرها می‌توان به اختلاف بین حداقل بار واقعی و پیش‌بینی‌شده، اختلاف بین حداکثر بار واقعی و پیش‌بینی‌شده، خطای مطلق، خطای درصدی، و دامنه بار واقعی و پیش‌بینی‌شده اشاره کرد.

### روندهای سالانه

تحلیل سالانه با هدف بررسی تغییرات میانگین بار واقعی و پیش‌بینی‌شده در طول سال‌های مختلف انجام شد. این بخش مشخص می‌کند که آیا سطح کلی مصرف یا بار برق در سال‌های مختلف روند افزایشی، کاهشی، یا نسبتاً ثابت داشته است. همچنین با مقایسه میانگین خطا در هر سال، می‌توان تغییرات دقت پیش‌بینی را در طول زمان ارزیابی کرد.

### روندهای هفتگی

برای تحلیل الگوی هفتگی، داده‌ها بر اساس شماره هفته گروه‌بندی شدند. این کار امکان بررسی رفتارهای تکرارشونده در طول هفته‌های سال را فراهم می‌کند. در صورت مشاهده الگوهای مشابه در هفته‌های مشخص، می‌توان وجود الگوی فصلی یا دوره‌ای در داده را مطرح کرد.

### مقایسه مقادیر واقعی و پیش‌بینی‌شده

یکی از اهداف اصلی این فاز، مقایسه مستقیم مقادیر واقعی و پیش‌بینی‌شده برای حداقل و حداکثر بار بود. این مقایسه به کمک نمودارهای خطی و پراکندگی انجام شد تا مشخص شود پیش‌بینی‌ها تا چه حد به مقادیر واقعی نزدیک بوده‌اند. هرچه نقاط در نمودارهای پراکندگی به الگوی خطی نزدیک‌تر باشند، کیفیت پیش‌بینی بهتر ارزیابی می‌شود.

### تحلیل اختلاف‌ها و خطاها

برای سنجش دقیق‌تر عملکرد پیش‌بینی، اختلاف بین مقادیر واقعی و پیش‌بینی‌شده در دو سطح حداقل و حداکثر بار محاسبه شد. سپس خطاها از منظر میانگین خطای مطلق، خطای درصدی و توزیع کلی خطاها مورد بررسی قرار گرفتند. این تحلیل نشان می‌دهد آیا مدل پیش‌بینی موجود تمایل به کم‌برآوردی یا بیش‌برآوردی دارد و در چه بازه‌هایی خطا بزرگ‌تر می‌شود.

### تحلیل نوسان‌پذیری

به‌منظور بررسی پایداری یا بی‌ثباتی بار برق، نوسان‌پذیری داده‌ها از طریق انحراف معیار و دامنه بار اندازه‌گیری شد. این بخش می‌تواند نشان دهد که بار واقعی در مقایسه با بار پیش‌بینی‌شده رفتار قابل پیش‌بینی‌تری دارد یا خیر. همچنین می‌توان تشخیص داد کدام بازه‌های زمانی ناپایدارتر هستند.

## جمع‌بندی فاز ۲

نتایج تحلیل اکتشافی داده‌ها پایه‌ای برای طراحی سوالات پژوهشی و انتخاب ویژگی‌ها و مدل‌های مناسب در مراحل بعدی فراهم می‌کند. این فاز کمک می‌کند پیش از ورود به مدل‌سازی، شناخت روشن‌تری از ساختار داده، میزان خطا، نوسان‌پذیری، و الگوهای زمانی موجود در مجموعه داده حاصل شود.

# استخراج ویژگی‌ها و آماده‌سازی برای مدل‌سازی

پس از انجام پاک‌سازی داده‌ها و تحلیل اکتشافی، در این مرحله تمرکز پروژه بر استخراج ویژگی‌های مناسب برای مدل‌سازی قرار گرفت. هدف اصلی این فاز آن بود که داده خام هفتگی به یک مجموعه‌داده غنی و ساختاریافته تبدیل شود؛ به‌گونه‌ای که الگوهای زمانی، روندها، نوسان‌ها و رفتارهای تکرارشونده موجود در بار برق برای مدل‌های پیش‌بینی قابل استفاده شوند.

در داده‌های سری زمانی، استفاده صرف از مقادیر خام معمولاً برای ساخت مدل‌های دقیق کافی نیست. به همین دلیل، مجموعه‌ای از ویژگی‌های جدید بر اساس زمان، رفتار گذشته، میانگین‌های متحرک، نوسان‌های اخیر و شاخص‌های فصلی تولید شد. این ویژگی‌ها به مدل کمک می‌کنند تا وابستگی‌های زمانی و تغییرات ساختاری موجود در داده را بهتر یاد بگیرد.

## ویژگی‌های زمانی

در گام نخست، ویژگی‌های پایه زمانی شامل سال، ماه، فصل، و شماره هفته استخراج شدند. این متغیرها اطلاعات اولیه‌ای درباره موقعیت زمانی هر مشاهده در اختیار مدل قرار می‌دهند و برای شناسایی روندهای سالانه و الگوهای فصلی مفید هستند.

## ویژگی‌های تأخیری(Lag Features)

به‌منظور استفاده از اطلاعات گذشته، ویژگی‌های تأخیری برای متغیرهای اصلی داده ساخته شد. این ویژگی‌ها نشان می‌دهند که مقدار واقعی یا پیش‌بینی‌شده بار در یک تا چند هفته قبل چه بوده است. از آنجا که در داده‌های زمانی، گذشته معمولاً یکی از مهم‌ترین منابع اطلاعات برای پیش‌بینی آینده است، استفاده ازlag ها نقش مهمی در افزایش توان مدل خواهد داشت.

## ویژگی‌های میانگین و انحراف معیار متحرک

برای ثبت رفتار کوتاه‌مدت و میان‌مدت داده، میانگین متحرک و انحراف معیار متحرک در پنجره‌های زمانی مختلف محاسبه شد. میانگین متحرک نمایی از سطح کلی بار در هفته‌های اخیر ارائه می‌دهد، در حالی که انحراف معیار متحرک میزان نوسان‌پذیری اخیر را نشان می‌دهد. این دو نوع ویژگی برای درک وضعیت پایدار یا ناپایدار سری زمانی اهمیت دارند.

## ویژگی‌های تفاضلی

در ادامه، تغییرات مقدار بار نسبت به هفته قبل محاسبه شد. این ویژگی‌ها نشان می‌دهند بار برق چه میزان افزایش یا کاهش داشته است و می‌توانند در شناسایی جهش‌ها، افت‌ها و تغییرات روند مؤثر باشند. علاوه بر اختلاف ساده، تغییرات درصدی نیز در برخی متغیرها محاسبه شدند تا تغییرات به‌صورت نسبی نیز قابل بررسی باشند.

## شاخص‌های فصلی

به‌منظور ثبت اثر فصل و ماهیت چرخه‌ای هفته‌های سال، شاخص‌های فصلی نیز تعریف شدند. برای این منظور علاوه بر تعریف فصل‌های سال، از تبدیل‌های سینوسی و کسینوسی بر روی شماره هفته استفاده شد. این روش باعث می‌شود مدل رابطه چرخه‌ای بین هفته‌های ابتدایی و انتهایی سال را بهتر درک کند.

## نتیجه این فاز

خروجی این مرحله یک مجموعه‌داده توسعه‌یافته برای مدل‌سازی است که در آن علاوه بر متغیرهای اصلی، مجموعه‌ای از ویژگی‌های مهندسی‌شده نیز وجود دارد. این دیتاست می‌تواند مستقیماً در فاز بعدی برای آموزش و مقایسه مدل‌های مختلف مورد استفاده قرار گیرد.

## ملاحظات فنی

لازم به ذکر است که برخی ویژگی‌های تأخیری و متحرک در هفته‌های ابتدایی هر سری زمانی دارای مقادیر تهی هستند. این موضوع کاملاً طبیعی است، زیرا برای محاسبه این ویژگی‌ها نیاز به داده‌های گذشته وجود دارد. در مرحله مدل‌سازی، این مقادیر با رویکرد مناسب مدیریت خواهند شد.

## مدل‌سازی و مقایسه مدل‌ها

در این فاز، پس از آماده‌سازی مجموعه‌داده ویژگی‌سازی‌شده، چند مدل مختلف برای پیش‌بینی بار برق مورد بررسی و مقایسه قرار گرفتند. هدف اصلی این مرحله، ارزیابی عملکرد رویکردهای مختلف یادگیری ماشین و مدل‌های سری زمانی در پیش‌بینی متغیر هدف، یعنیactual_max_mw ، بود.

برای حفظ ماهیت زمانی داده‌ها، تقسیم داده‌ها به مجموعه آموزش و آزمون به‌صورت ترتیبی انجام شد و از اختلاط تصادفی داده‌ها جلوگیری گردید. این موضوع از آن جهت اهمیت دارد که در مسائل پیش‌بینی زمانی، استفاده از اطلاعات آینده در فرآیند آموزش منجر به ارزیابی غیرواقعی و خوش‌بینانه مدل خواهد شد.

### مدل‌های انتخاب‌شده

در این پروژه، چهار رویکرد اصلی برای مقایسه در نظر گرفته شد:

#### 1. مدلBaseline

این مدل ساده‌ترین مرجع برای مقایسه است و مقدار آخرین مشاهده در داده آموزش را به‌عنوان پیش‌بینی برای داده‌های آزمون در نظر می‌گیرد. هدف از استفاده از این مدل، ایجاد یک حداقل سطح عملکرد برای سنجش ارزش افزوده مدل‌های پیشرفته‌تر است.

#### 2. Random Forest

این مدل یکی از روش‌های قدرتمند یادگیری ماشین برای مسائل رگرسیون است که با استفاده از مجموعه‌ای از درخت‌های تصمیم، روابط غیرخطی موجود در داده را یاد می‌گیرد Random Forest .در عین سادگی نسبی، معمولاً عملکرد پایداری در داده‌های جدولی دارد.

#### 3. XGBoost

این مدل یک روش تقویتی مبتنی بر درخت است که در بسیاری از مسائل پیش‌بینی، عملکرد بسیار خوبی از خود نشان می‌دهد . XGBoostبه‌ویژه در شرایطی که داده شامل ویژگی‌های مهندسی‌شده متنوع باشد، می‌تواند الگوهای پیچیده‌تری را استخراج کند.

#### 4. ARIMA

این مدل به‌عنوان نماینده‌ای از روش‌های کلاسیک سری زمانی انتخاب شد ARIMA .با تکیه بر ساختار خودهمبستگی و روندهای موجود در سری زمانی، امکان پیش‌بینی آینده را فراهم می‌کند و برای مقایسه با روش‌های مبتنی بر ویژگی بسیار مناسب است.

### معیارهای ارزیابی

برای مقایسه مدل‌ها از چهار شاخص اصلی استفاده شد:

MAE : میانگین قدرمطلق خطا

RMSE : ریشه میانگین مربعات خطا

MAPE : میانگین درصد خطای مطلق

$R^2$ : ضریب تعیین

این شاخص‌ها به‌صورت مکمل، هم میزان خطای مطلق، هم حساسیت به خطاهای بزرگ، و هم قدرت توضیح مدل را نشان می‌دهند.

## نتایج مقایسه

پس از آموزش مدل‌ها و اجرای پیش‌بینی روی داده‌های آزمون، عملکرد آن‌ها بر اساس شاخص‌های فوق مقایسه شد. انتظار می‌رود مدل Baseline ضعیف‌ترین عملکرد را داشته باشد، زیرا از هیچ ساختار پیچیده‌ای در داده استفاده نمی‌کند. در مقابل، مدل‌های Random Forestو XGBoost به دلیل بهره‌گیری از ویژگی‌های مهندسی‌شده، معمولاً عملکرد بهتری ارائه می‌دهند. همچنین مدل ARIMAتصویری از توانایی روش‌های کلاسیک سری زمانی در این مسئله فراهم می‌کند.

## جمع‌بندی

هدف اصلی این فاز، صرفاً یافتن یک مدل برتر نبود، بلکه مقایسه منظم چند رویکرد متفاوت برای درک بهتر رفتار داده و سنجش قابلیت پیش‌بینی آن بود. نتایج به‌دست‌آمده مبنای مناسبی برای انتخاب مدل نهایی و توسعه فازهای بعدی پروژه فراهم می‌کنند.