پیش پردازش:

به منظور تحلیل بهتر روی داده ها، ویژگی date ابتدا نیاز است به فرمت timestamp تبدیل شده و سپس برای تحلیل با نمودار correlation، به فرمت عددی تبدیل شود اما داده null در این ویژگی وجود ندارد بنابراین نیازی به حذف یا تخمین داده نیست.
برای ویژگی weather station که به محل ایستگاه جمع آوری و رصد شرایط آب و هوایی اشاره دارد تنها نیاز است که با روش one-hot-encoding نام هر ایستگاه به شماره آن map شود.
ویژگی های rainfall، minimum temperature و maximum temperature که داده های عددی پیوسته هستند شامل تعداد کمی داده null بوده که به دلیل رابطه تقریبا خطی بین داده های موجود میتوان از interpolation خطی برای تخمین داده های پوچ با استفاده از داده ها موجود استفاده کرد.
در رابطه با ویژگی evaporation که در حدود 12000 داده null وجود دارد استفاده از میانه، میانگین و یا مد ممکن است دقت مدل را کاهش دهد بنابراین در ابتدا میتوان با تحلیل نمودار وابستگی، بیشترین وابستگی بین ویژگی فعلی با ویژگی دیگری که داده های آن کامل است که در اینجا max temperature است با آموزش یک مدل رگرسیون خطی برای پیشبینی داده های پوچ evaporation استفاده کرد اما در نهایت به دلیل وابستگی زیاد آن برای افزایش دقت مدل میتوان آنرا با دیگر ویژگی ها ترکیب یا حذف کرد.
اما برای ویژگی air velocity، cloudiness و sunshine به دلیل مشابهت زیاد داده های کنار هم، میتوان از روش forward fill, backward که داده های پوچ را با توجه به داده های قبلی و بعدی تکمیل میکند استفاده کرد. اما برای ویژگی Gust trajectory که یک ویژگی categorical بوده و مختصات جهت باد را نشان میدهد برای تبدیل به داده عددی با تعریف یک dictionary که جهت های جغرافیایی را به مختصات بر حسب درجه تبدیل میکند، استفاده میکنیم. ویژگی recorded temperature نیز به دلیل correlation زیاد با max , min temperature میتوانیم حذف کنیم که منجر به ساده تر شدن داده ها و افزایش سرعت و کارایی مدل میشود.

تحلیل نمودار box plot:

بارش در همان روز بر اســـاس هدف (بارش روز بعد) :نمودار نشـــان می‌دهد که اگر روز بعد باران ببارد (هدف = 1) ، احتمال زیادی وجود دارد که آن روز نیز باران ببارد. در حالی که اگر روز بعد باران نبارد (هدف = 0)، احتمال بارش آن روز کمتر است. این امر در ناحیه وسیع‌تر مستطیل برای هدف = 1 و نقاط خارج از محدوده برای هدف = 0 دیده می‌شود.

سطح رطوبت در ساعت 3 بعد از ظهر بر اساس هدف :این نمودار نشان می‌دهد که در صورت پیش‌بینی بارش روز بعد (هدف = 1)، سـطح رطوبت در سـاعت 3 بعد از ظهر به طور میانگین بالاتر اسـت. در حالی که در روزهایی که بارش روز بعد پیش‌بینی نشـده اسـت (هدف = 0)، سـطح رطوبت پایین‌تری دارند. این امر از میانگین‌های بالاتر و توزیـع‌های گسترده‌تر برای هدف = 1 مشخص است.

سرعت هوا بر اساس هدف :این نمودار نشان می‌دهد که در روزهایی که بارش روز بعد پیش‌بینی شده است (هدف = 1)، سرعت هوا به طور میانگین بالاتر است. در حالی که در روزهایی که بارش روز بعد پیش‌بینی نشده است (هدف = 0)، سرعت هوا پایین‌تر است و نقاط خارج از محدوده بیشتری وجود دارد.

دمای حداقل بر اساس هدف :این نمودار نشان می‌دهد که در روزهایی که بارش روز بعد پیش‌بینی شده است (هدف = 1)، دمای حداقل به طور میانگین بالاتر است. در حالی که در روزهایی که بارش روز بعد پیش‌بینی نشده است (هدف = 0)، دمای حداقل پایین‌تری دارند. این امر از میانگین‌های بالاتر برای هدف = 1 و توزیع‌های نزدیک‌تر به میانه در نمودار قابل مشاهده است.

تحیل نمودار Histogram:

توزیع سرعت هوا را نشان می‌دهد. توزیع به نظر دو قله‌ای است، با یک قله حدود ۳۰ و قله دیگری بالاتر حدود ۷۰، که نشان‌دهنده دو گروه یا الگوی متفاوت در داده‌های سرعت هوا است.

توزیع مسیر تندباد در ساعات ۳ بعدازظهر یک توزیع چند قله‌ای با چندین قله است، که نشان‌دهنده وجود چندین مقادیر یا دسته‌های متداول در داده‌های مسیر تندباد در آن زمان است. توزیع سرعت هوا در ساعت ۳ بعدازظهر نشان می‌دهد که این توزیع چندین قله دارد و به نظر می‌رسد نسبت به توزیع کلی سرعت هوا، پراکنده‌تر است، که ممکن است بازتابی از تغییرپذیری سرعت هوا در آن زمان خاص از روز باشد.

توزیع فشار جوی در ساعت ۹ صبح یک توزیع به نظر تقریباً نرمال یا به شکل زنگ است، که نشان‌دهنده آن است که فشار جوی در آن زمان از الگوی توزیع نرمال معمول پیروی می‌کند.

توزیع ابری بودن در ساعات ۳ بعدازظهر نشان می‌دهد که توزیع به شدت به سمت راست کج است، با فراوانی بسیار بالا در مقادیر ابری پایین (۰ و ۱) و دنباله طولانی به سمت مقادیر ابری بالاتر. این الگو نشان‌دهنده آن است که شرایط آسمان صاف یا عمدتاً صاف رایج‌تر بوده است.