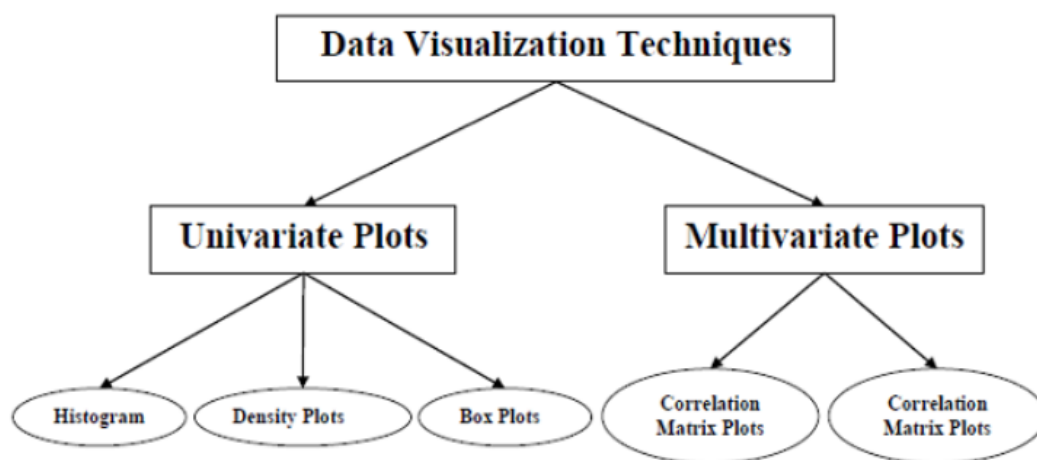


درک داده با مصورسازی (Visualization) :

مقدمه:

در بخش قبل، در رابطه با اهمیت داده برای الگوریتم های یادگیری ماشین و همچنین برخی از دستورالعمل های پایتون جهت فهمیدن داده به همراه آمار و ارقام، صحبت کردیم. روشی دیگر به نام مصورسازی (Visualization) نیز برای درک داده وجود دارد.

با کمک تجسم داده، می توان دید که داده ها چگونه به نظر می رسند و چه نوع همبستگی (correlation) بین ویژگی های داده وجود دارد. این سریع ترین راه برای فهمیدن این است که آیا ویژگی ها با خروجی مطابقت دارند یا خیر. با کمک دستور العمل های پایتون که در ادامه مطرح می شود، می توان داده های ML را با تجسم فهمید.



طرح های تک متغیره (univariate plot) :

درک ویژگی ها به طور مستقل: ساده ترین نوع تجسم، تجسم تک متغیره یا univariate است. با کمک تجسم تک متغیره، می توانیم هر ویژگی مجموعه داده خود را به طور مستقل درک کنیم. برخی از روش های موجود در پایتون برای پیاده سازی تجسم تک متغیره، در ادامه معرفی شده است.

۱- Histograms : هیستوگرام ها داده را در سطل ها (bins) گروه بندی می کنند. این سریع ترین راه برای درک توزیع هر ویژگی در مجموعه داده است.

۲- Density plots : یک روش ساده و سریع دیگر برای دریافت توزیع هر ویژگی Density plots است.

۳- Box and Whisker plots : این روش که به اختصار boxplots نیز نامیده می شود، یک روش مفید دیگر برای مرور توزیع هر ویژگی است.

طرح های چند متغیره (multivariate plots) :

تعامل بین چندین متغیر:

نوع دیگری از تجسم، تجسم چند متغیره یا multivariate نام دارد. به کمک تجسم چند متغیره، می توانیم تعامل بین چندین ویژگی از مجموعه داده خود را درک کنیم. برخی از روش های موجود در پایتون برای پیاده سازی تجسم چند متغیره، در ادامه معرفی شده است.

۱- Correlation matrix plot : همبستگی (Correlation) یک معیار برای نمایش تغییرات بین دو متغیر است.

۲- Scatter matrix plot: طرح های پراکندگی (Scatter) ، به کمک نقاط در دو بعد، نشان دهنده این است که یک متغیر و یا رابطه بین متغیر ها، چقدر تحت تاثیر دیگری قرار گرفته است.