به نام خدا

نام : پریا جبرئیلی شماره دانشجویی:993212018

در بخش اول با توجه به شماره دانشجویی 10 ستون به صورت رندوم از کل ستونهای دیتاست انتخاب شد. با کمک این ستونها یک دیتاست جدید ایجاد کردم.



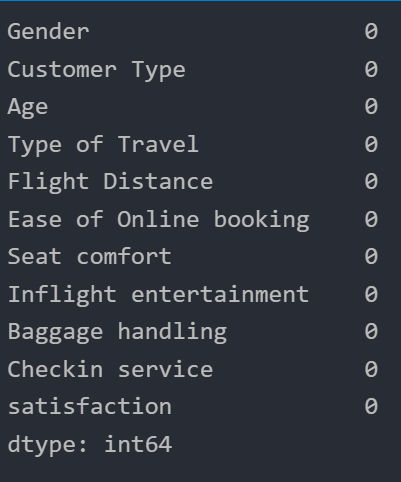
['Gender', 'Customer Type', 'Age', 'Type of Travel', 'Flight Distance', 'Ease of Online booking', 'Seat comfort', 'Inflight entertainment', 'Baggage handling', 'Checkin service', 'satisfaction']

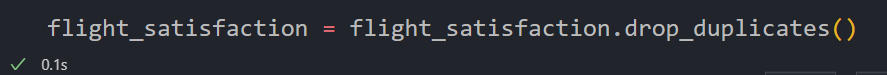
در بخش بعدی چک کردم اگر missing value ای وجود دارد یا خیر.

این مقادیر از دست رفته احتمالاً به این دلیل از دست رفته اند که ثبت نشده اند، نه به دلیل اینکه وجود ندارند.

اگر missing value وجود داشته باشد میتواند drop شود با اینکه می توان مقادیرش با مقدار منطقی دیگری پر کرد. این مقدار رو میتونیم خودمون بدیم یا مقدار سطر بعدی یا قبلی همون ستون رو بهش نسبت بدهیم.

که هیچ missing value ای نداشتم.



در بخش بعدی سعی کردم دیتا هایی که تکراری هستند را مدیریت کنم.

در بخش بعدی به شناسایی دیتاهای پرت یا outliers با visualization پرداختم.

Outlier چیست؟

Outlierبه نقاط داده‌ای گفته می‌شود که به طور قابل توجهی متفاوت از سایر داده‌ها هستند. در واقع، اوتلایرها نقاطی در مجموعه داده هستند که از الگوی کلی داده‌ها خارج بوده و ممکن است نشان‌دهنده تغییرات غیرعادی یا خطاهای اندازه‌گیری باشند.

وجود اوتلایرها می‌تواند بر تحلیل‌های آماری و مدل‌های پیش‌بینی تأثیر بگذارد، زیرا ممکن است باعث تحریف نتایج و برآوردهای نادرست شوند. به عنوان مثال، یک اوتلایر می‌تواند میانگین مجموعه داده‌ها را به شکل قابل توجهی تغییر دهد.

شناسایی و رسیدگی به اوتلایرها یک قسمت مهم در پردازش و تحلیل داده‌ها است. رویکردهای مختلفی برای شناسایی اوتلایرها وجود دارد، از جمله استفاده از نمودارهای آماری مانند جعبه‌ای (Boxplot) یا تحلیل فاصله‌های بین‌کوارتیلی (IQR).

یکی از راحت ترین روشها برای اینکه تشخیص بدهیم دیتا ما دارای outlier هست یا خیر plot کردن دیتا موردنظر هست.

با کمک روشهایی این outlier ها را تشخیص میدهیم و انها را از دیتا حذف میکنیم و با دیتاست جدید کار میکنیم.

یکی از این روشها z-score هست.

z-score چیست؟

z-score، یک معیار آماری است که نشان می‌دهد هر داده‌ای نسبت به میانگین مجموعه داده‌ها چقدر دور یا نزدیک است. به بیان دیگر، Z-اسکور مشخص می‌کند که یک نقطه داده خاص چند انحراف معیار بالاتر یا پایین‌تر از میانگین قرار دارد.

یکی از کاربردهای رایج Z-اسکور در شناسایی اوتلایرها (نقاط داده‌ای که به شدت از سایر داده‌ها متفاوت هستند) است.

بنابراین، Z-اسکور ابزاری قدرتمند در تحلیل آماری است که به ما اجازه می‌دهد داده‌ها را به شکل استاندارد بررسی کرده و نقاط غیرعادی را شناسایی کنیم.

یکی دیگر از روشها IQR هست.

IQR چیست؟

IQR اختلاف بین کوارتیل سوم (Q3) و کوارتیل اول (Q1) در یک مجموعه داده است. به بیان ساده‌تر، IQR نشان‌دهنده محدوده‌ای از داده‌ها است که نیمی از کل داده‌ها در آن قرار دارند.

IQR محاسبه می‌شود به صورت: IQR = Q3 – Q1

برای شناسایی اوتلایرها با استفاده از IQR، معمولاً از 1.5 برابر IQR فراتر از کوارتیل‌های اول و سوم استفاده می‌شود:

Upper Bound = Q3 + 1.5 \* IQR

Lower Bound = Q1 – 1.5 \* IQR

داده‌هایی که فراتر از این حدود قرار دارند، به عنوان اوتلایر در نظر گرفته می‌شوند.

روش بعدی percentile method هست .

Percentile method چیست؟

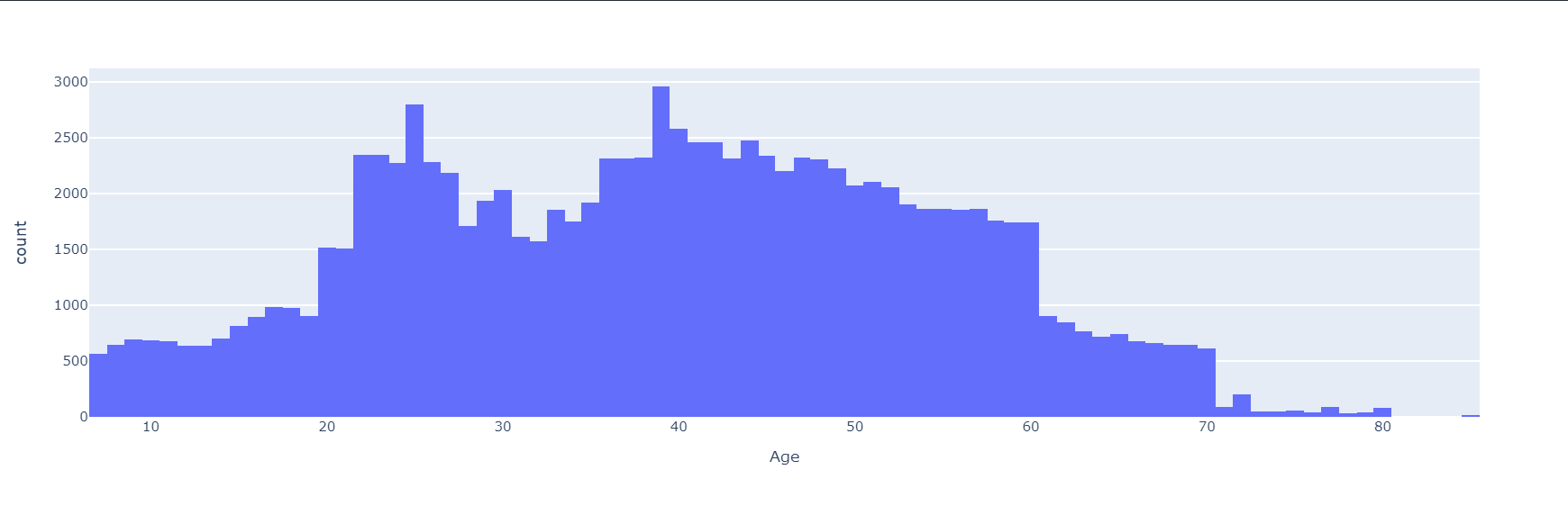
Percentile Methodیکی از روش‌های آماری است که برای شناسایی موقعیت یک نقطه داده در مقایسه با کل مجموعه داده‌ها استفاده می‌شود. Percentile Method مقادیری هستند که داده‌های یک مجموعه را به صد بخش مساوی تقسیم می‌کنند. به بیان دیگر، یک Percentile Method نشان‌دهنده مقداری است که یک درصد خاص از داده‌ها کمتر یا برابر با آن مقدار هستند.

برای ستونهای زیر ابتدا داده رو plot کردم و سپس outlier ها را تشخیص دادم واونها رو حذف کردم.

برای هر کدام از ستونها هر 3 روش را باری تشخیص و حذف outlierاستفاده کردم.

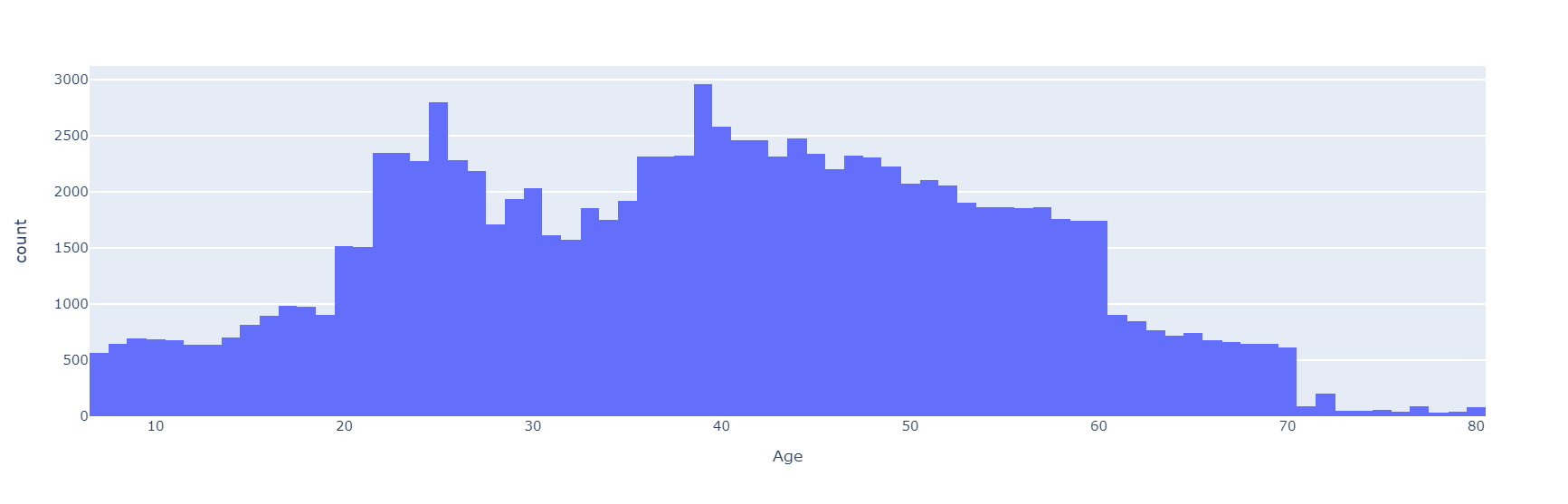
Age =

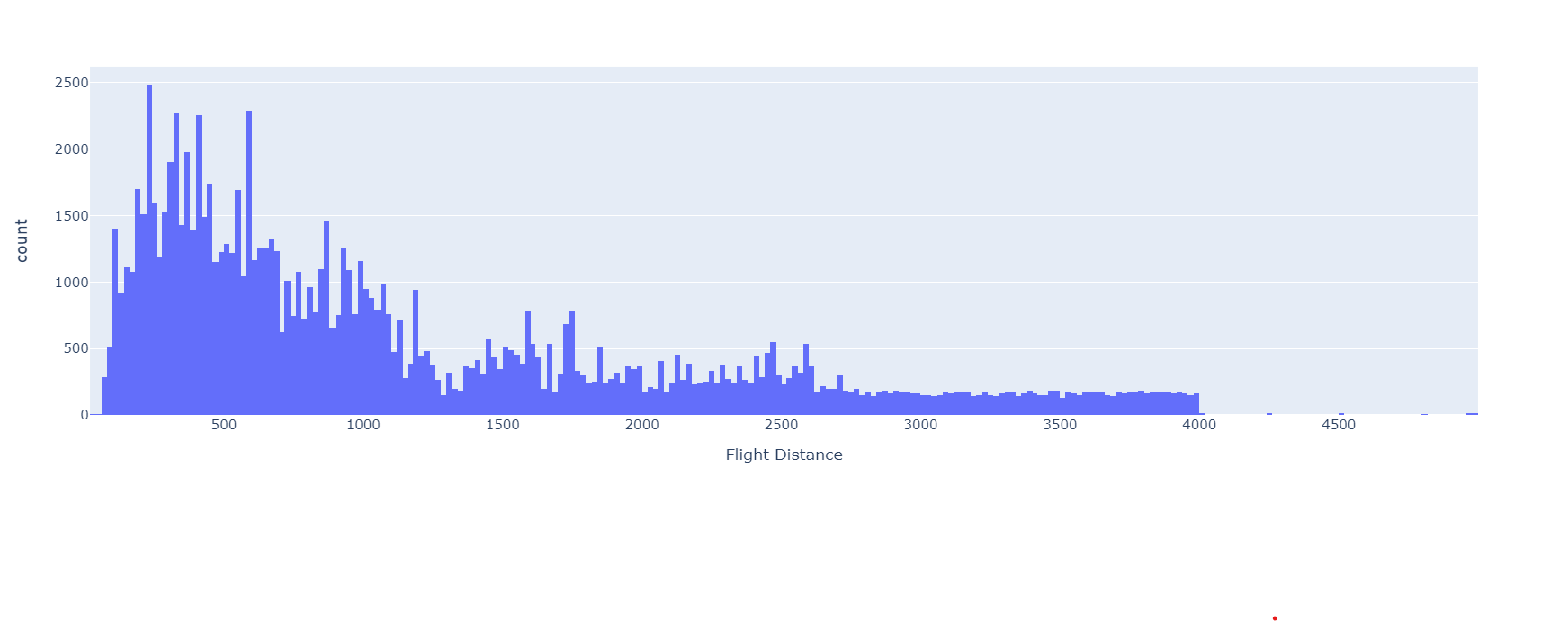
با توجه به نمودار متوجه میشویم یه سری بالاتر از 80 هستن و از داده های دیگر پرت هستند.



بعد حذف outlier ها :

داده های پرت حذف شده است.

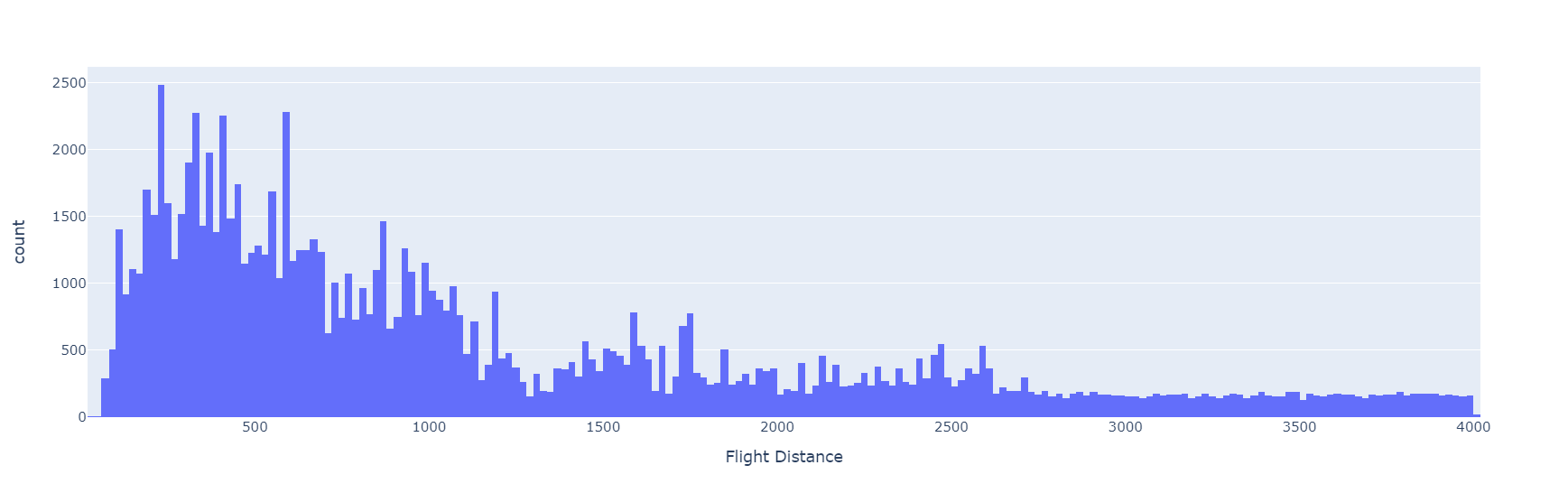


Flight Distance =

با توجه به نمودار به وضوح میتوان داده های پرت را دید که در فاصله بالاتر از 4000 هستند.

بعد حذف outlier ها :

داده های پرت حذف شده است.



چرا تبدیل نوع داده‌ها انجام می‌شود و توضیح آن:

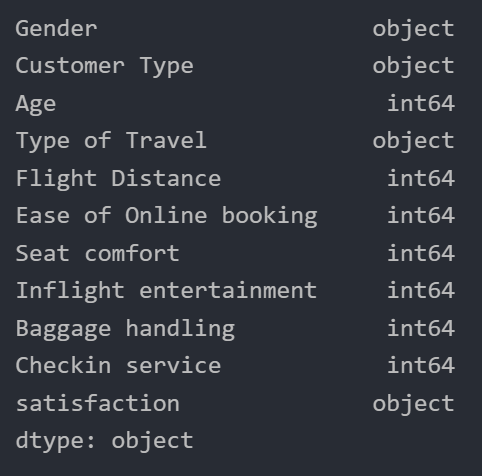
Data Type Conversionفرایندی است که در آن نوع داده‌ها در یک مجموعه داده تغییر می‌کند. این تغییر معمولاً برای اطمینان از سازگاری داده‌ها با نیازهای تحلیلی یا الزامات نرم‌افزاری خاص انجام می‌شود.دلایل و اهمیت این است که :

تغییر نوع داده‌ها می‌تواند به بهبود کارایی پردازش داده‌ها کمک کند. به عنوان مثال، تبدیل داده‌های متنی به دسته‌بندی‌های عددی می‌تواند در سرعت پردازش و تحلیل داده‌ها مؤثر باشد.

تبدیل نوع داده‌ها به استانداردسازی داده‌ها در میان مجموعه‌های مختلف کمک می‌کند.

تغییر نوع داده‌ها می‌تواند به سهولت بیشتر در تفسیر و تجزیه و تحلیل آن‌ها کمک کند، به ویژه در مواردی که نوع داده‌ها به صورت اولیه مناسب تحلیل نیست.

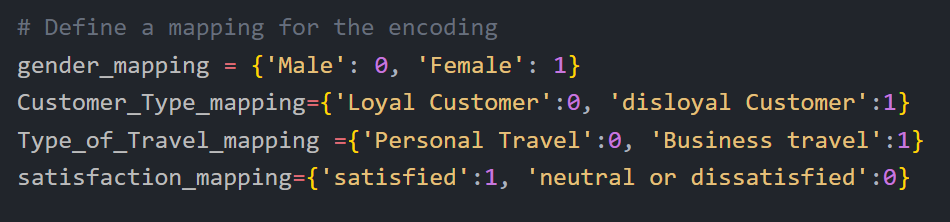
ابتدا نوع داده هایی که داشتم رو در خروجی چاپ کردم:



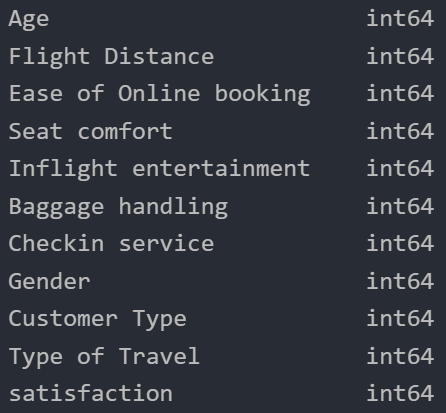
ستونهایی که نوع داده شون object بود رو به کمک numeric encoding به int تبدیل کردم.

دلیل اینکه از numeric encoding استفاده کردم این بود که شبکه عصبی ای که در بخش اخر استفاده شده نیاز به داده های عددی دارد و نمیتواند مستقیما با داده های متنی کار کند.

پس ستونهای gender و customer\_Type و Type\_of\_Travel و satisfaction را به کمک numeric encoding به نوع داده int تبدیل میکنیم.



خروجی نوع داده به این صورت میشود:



در بخش بعدی Min-Max Scaling را انجام دادم. این روش برای مقیاس‌بندی داده‌ها به گونه‌ای استفاده می‌شود که مقادیر هر ویژگی در بازه 0 تا 1 قرار گیرند.

تعیین مقادیر کمینه و بیشینه برای =Flight Distance

- min\_val = final\_dataset['Flight Distance'].min: این خط کمینه (کوچک‌ترین مقدار) ستون 'Flight Distance' در دیتاست `final\_dataset` را محاسبه و در `min\_val` ذخیره می‌کند.

- max\_val = final\_dataset['Flight Distance'].max: این خط بیشینه (بزرگ‌ترین مقدار) ستون 'Flight Distance' را محاسبه و در `max\_val` ذخیره می‌کند.

اعمال تبدیل مقیاس کمینه-بیشینه برای 'Flight Distance' =

- final\_dataset['Flight Distance'] = (final\_dataset['Flight Distance'] - min\_val) / (max\_val - min\_val):

این خط هر مقدار در ستون 'Flight Distance' را با استفاده از فرمول کمینه-بیشینه مقیاس‌بندی می‌کند. مقیاس‌بندی شده‌ها در همان ستون ذخیره می‌شوند.

تکرار روند برای ستون 'Age' =

همان مراحل برای ستون Age تکرار می‌شود. ابتدا کمینه و بیشینه محاسبه شده و سپس فرمول مقیاس‌بندی به کار گرفته می‌شود.

این روش مقیاس‌بندی به یکنواخت سازی مقیاس‌های مختلف داده‌ها کمک می‌کند، بدین معنی که تمام ویژگی‌ها در یک مقیاس استاندارد و یکسان قرار می‌گیرند.

شبکه عصبی نیاز به داده‌هایی با مقیاس‌های یکنواخت دارند تا بتوانند بهتر کار کنند. مقیاس‌بندی کمینه-بیشینه به تسهیل این فرایند کمک می‌کند.

در مرحله بعد شبکه عصبی را تشکیل میدهم .