

دانشکده مهندسي برق و کامپیوتر

نام و نام خانوادگی :

امیر اسماعیل زاده نوبری

شماره دانشجویی

40101924

درس یادگیری ماشین

امتحان میان ترم

استاد درس:

دکتر علیاری

لینک [GIT](https://github.com/CAmiren/Machine_learning4022)

لینک [Drive](https://drive.google.com/drive/folders/1p4VGGkg3_734OVDWwAc3shnKLMmd83sQ?usp=sharing)

بهار 1403



Contents

[سوال 1 4](#_Toc163947978)

[1 . **Error! Bookmark not defined.**](#_Toc163947979)

[2. **Error! Bookmark not defined.**](#_Toc163947980)

[3 . **Error! Bookmark not defined.**](#_Toc163947981)

[الف ) Logistic regression : **Error! Bookmark not defined.**](#_Toc163947982)

[ب ) Perceptron : **Error! Bookmark not defined.**](#_Toc163947983)

[4 . **Error! Bookmark not defined.**](#_Toc163947984)

[الف ) Logistic Regression **Error! Bookmark not defined.**](#_Toc163947985)

[ب ) perceptron : **Error! Bookmark not defined.**](#_Toc163947986)

[5 . **Error! Bookmark not defined.**](#_Toc163947987)

[سوال 2 5](#_Toc163947988)

[1 . **Error! Bookmark not defined.**](#_Toc163947989)

[2 . **Error! Bookmark not defined.**](#_Toc163947990)

[آ ) **Error! Bookmark not defined.**](#_Toc163947991)

[ب ) **Error! Bookmark not defined.**](#_Toc163947992)

[ج ) **Error! Bookmark not defined.**](#_Toc163947993)

[د ) **Error! Bookmark not defined.**](#_Toc163947994)

[3 . **Error! Bookmark not defined.**](#_Toc163947995)

[4 . **Error! Bookmark not defined.**](#_Toc163947996)

[سوال 3 6](#_Toc163947997)

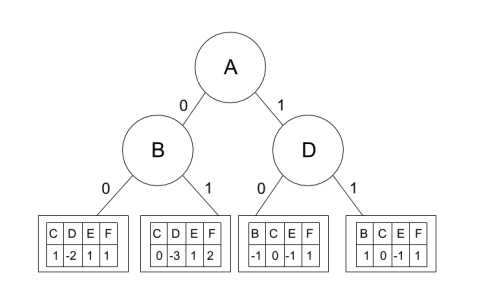
[2 . **Error! Bookmark not defined.**](#_Toc163947998)

[3 . **Error! Bookmark not defined.**](#_Toc163947999)

# سوال 1

1. خیر، این جمله به صورت کلی درست نیست. طبقه‌بند بیز یکی از روش‌های طبقه‌بندی است که در برخی موارد بسیار موثر و مفید است، اما نمی‌توان گفت که همیشه بهترین روش برای جداسازی یک مسأله دوکلاسه است. انتخاب بهترین طبقه‌بند به عوامل مختلفی از جمله:ماهیت داده‌ها ،توزیع داده‌ها ، پیچیدگی محاسباتی و بستگی دارد.
2. این جمله به طور کلی درست است، اما نیاز به توضیح و تحلیل بیشتری دارد. استفاده از رویکرد بیز برای تخمین پارامترهای توزیع می‌تواند به کاهش بیش‌برازش (Overfitting) کمک کند، به ویژه به دلیل ویژگی‌های ذاتی این رویکرد.
3. بله، این جمله به طور کلی درست است. استفاده از معیار Information Gain (کسب اطلاعات) برای ساخت درخت در شرایطی که بعضی از ویژگی‌ها حالات زیادی دارند، می‌تواند مناسب نباشد. از دلیل آن ، مشکل بایاس که Information Gain نسبت به ویژگی‌هایی با تعداد حالات بیشتر دارد و همچنین تقسیم بیش از حد داده‌ها.
4. بله ، هر شبکه عصبی چندلایه با توابع فعال‌ساز خطی قابل تقلیل به یک شبکه عصبی تک‌لایه با یک تابع خطی معادل است. این به این دلیل است که توابع خطی تجمعی می‌توانند به یک تابع خطی ساده‌تر تبدیل شوند.

# سوال 2



الف )



با توجه به نمونه داده شده A=1 و D=1 در leafnode سمت راست هستیم.

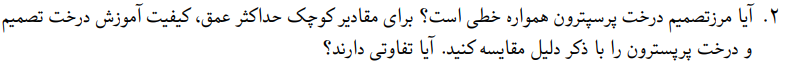
در نتیجه وزن های پرسپترون ما

و بایاس ما 1 و ورودی پرسپترون است.

داریم.

خروجی ما عدد 1 را نشان میدهد

ب)



خیر، مرز تصمیم درخت پرسپترون همواره خطی نیست. هر پرسپترون در گره‌های برگ مرزی خطی تعریف می‌کند، اما ساختار کلی درخت تصمیم می‌تواند مرزهای غیرخطی ایجاد کند. ساختار سلسله‌مراتبی درخت تصمیم فضای ویژگی‌ها را به مناطق مختلف تقسیم می‌کند و در هر منطقه پرسپترون عمل می‌کند. این ترکیب از تقسیمات درخت تصمیم و مرزهای خطی از پرسپترون‌ها، منجر به مرزهای تصمیم تکه‌ای-خطی می‌شود که به صورت کامل خطی نیستند.

برای مقادیر کوچک حداکثر عمق، درخت‌های پرسپترون می‌توانند کیفیت آموزش بهتری نسبت به درخت‌های تصمیم ساده ارائه دهند زیرا توانایی شناسایی الگوهای پیچیده‌تر از طریق مرزهای تکه‌ای-خطی را دارند. با این حال، این مسئله با افزایش پیچیدگی و زمان آموزش همراه است. انتخاب بین استفاده از درخت تصمیم یا درخت پرسپترون بستگی به مسئله خاص، طبیعت داده‌ها و موازنه بین قابلیت تفسیر و بیانگری دارد.

# سوال 3

سوال سه

بخش یک:

روابط لایه پنهان:

روابط لایه خروجی:

خروجی شبکه عصبی:

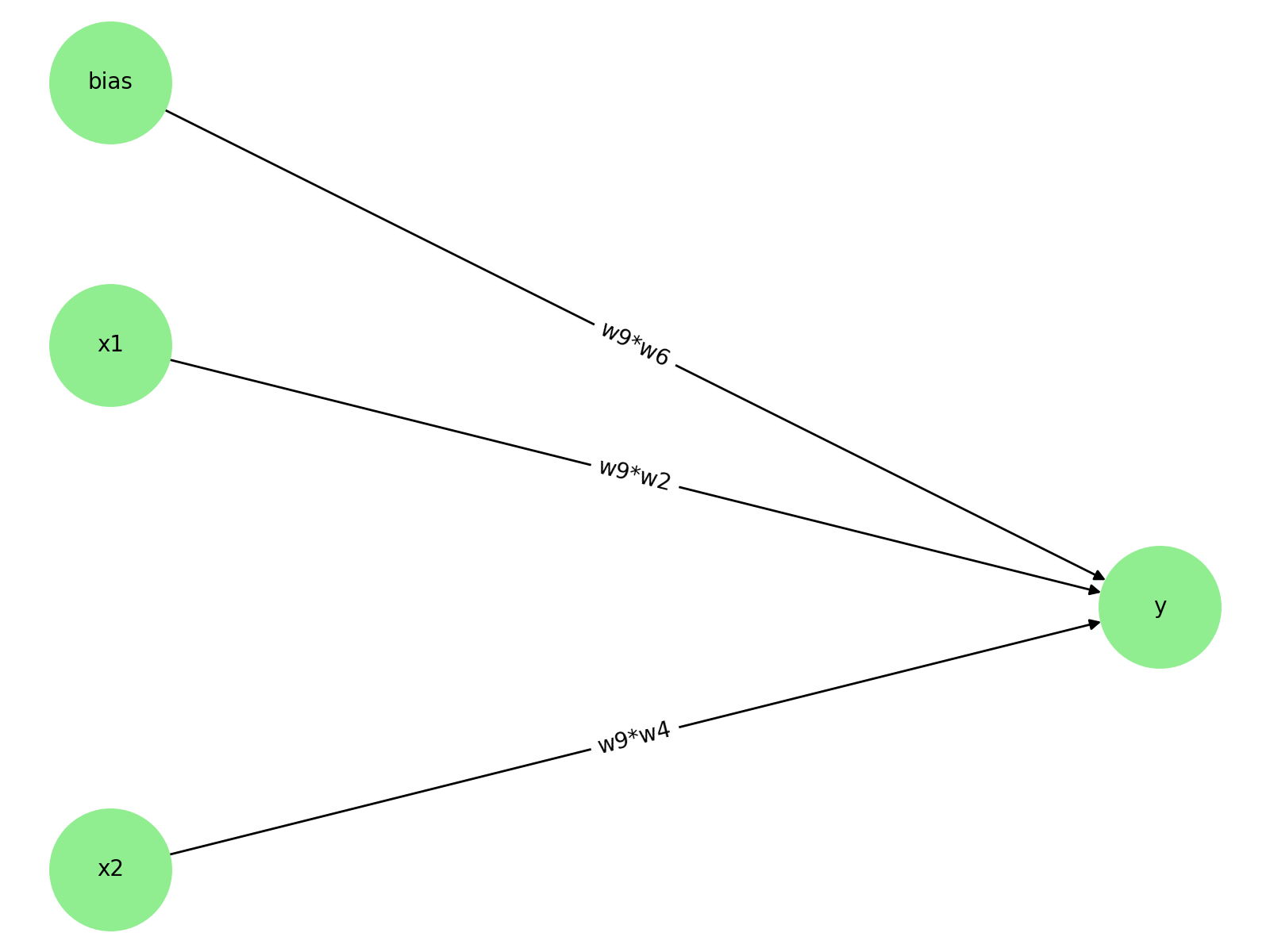
مقدار z را با توجه به روابط لایه پنهان، به شکل زیر به دست میاریم:

بنابراین:

بخش دوم:

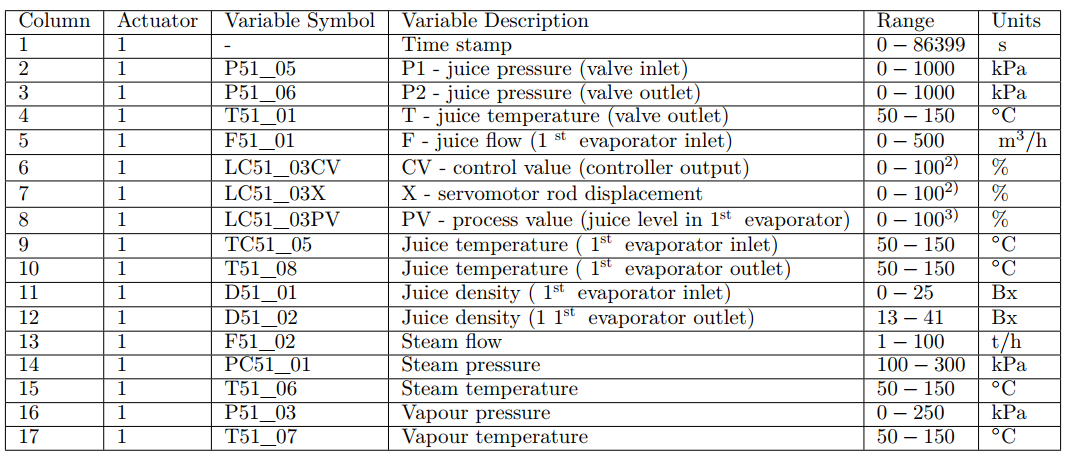
با توجه به این که ترکیب چندین تابع خطی همچنان یک تابع خطی است، هر شبکه عصبی چندلایه با توابع فعال‌ساز خطی در لایه‌های پنهان می‌تواند به عنوان یک شبکه عصبی بدون هیچ لایه پنهانی نمایش داده شود.

بنابراین شبکه عصبی مورد سوال، بدون لایه پنهان به شکل زیر در میاید:

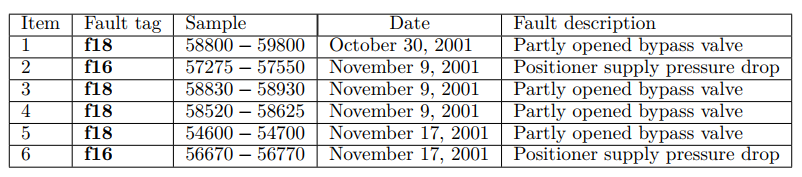


اگر روابط این شبکه را بنویسیم مشاهده میشود که با شبکه اصلی برابر است:

# سوال 4

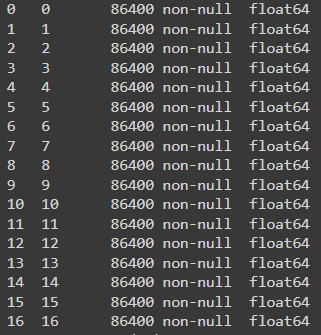


داده ها در 17 بعد داده شده اند که بعد اول آن ها شماره سمپل هاست که قابل حذف است.  
داده های 9 و 17 نوامبر با توجه به شکل زیر آورده شده اند:

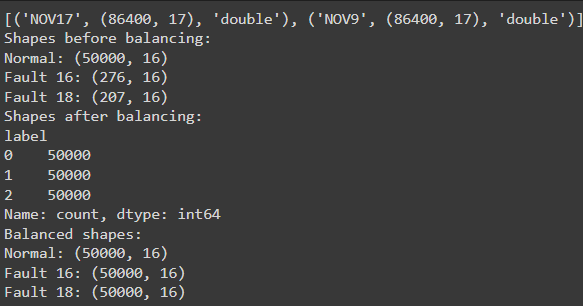


طبق خواسته ی سوال از داده های 9 نوامبر برای train و validation و از داده های 17 نوامبر برای تست استفاده می کنیم.

با دستور های مختلف python می توان اطلاعات زیادیی راجع به این دیتا دریافت کرد برای مثال با .info() داریم:



که نشان از این است که داده های null نداریم . dtype از جنس float است.

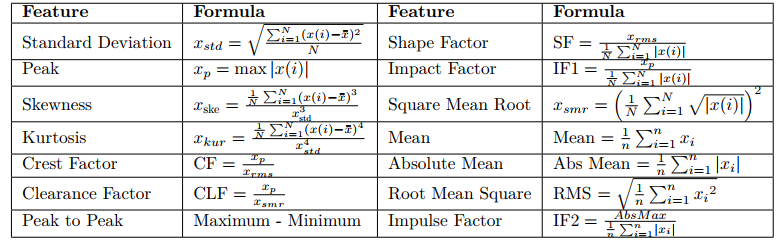
با توجه به رنج سمپل داده شده داده های نرمال(فقط 5000) و دو فالت f16 و f18 را استخراج می کنیم. میبینیم که تعداد داده ی فالت بسیار کم است پس برای رفع این unbalance با روش Smote داده ها balance می کنیم:  


با تعریف تابعی ماتریس M\*N با مقادیر 200\*100 برای استخراج ویژگی تشکیل می دهیم

برای داده های نرمال و دو فال گفته شده:



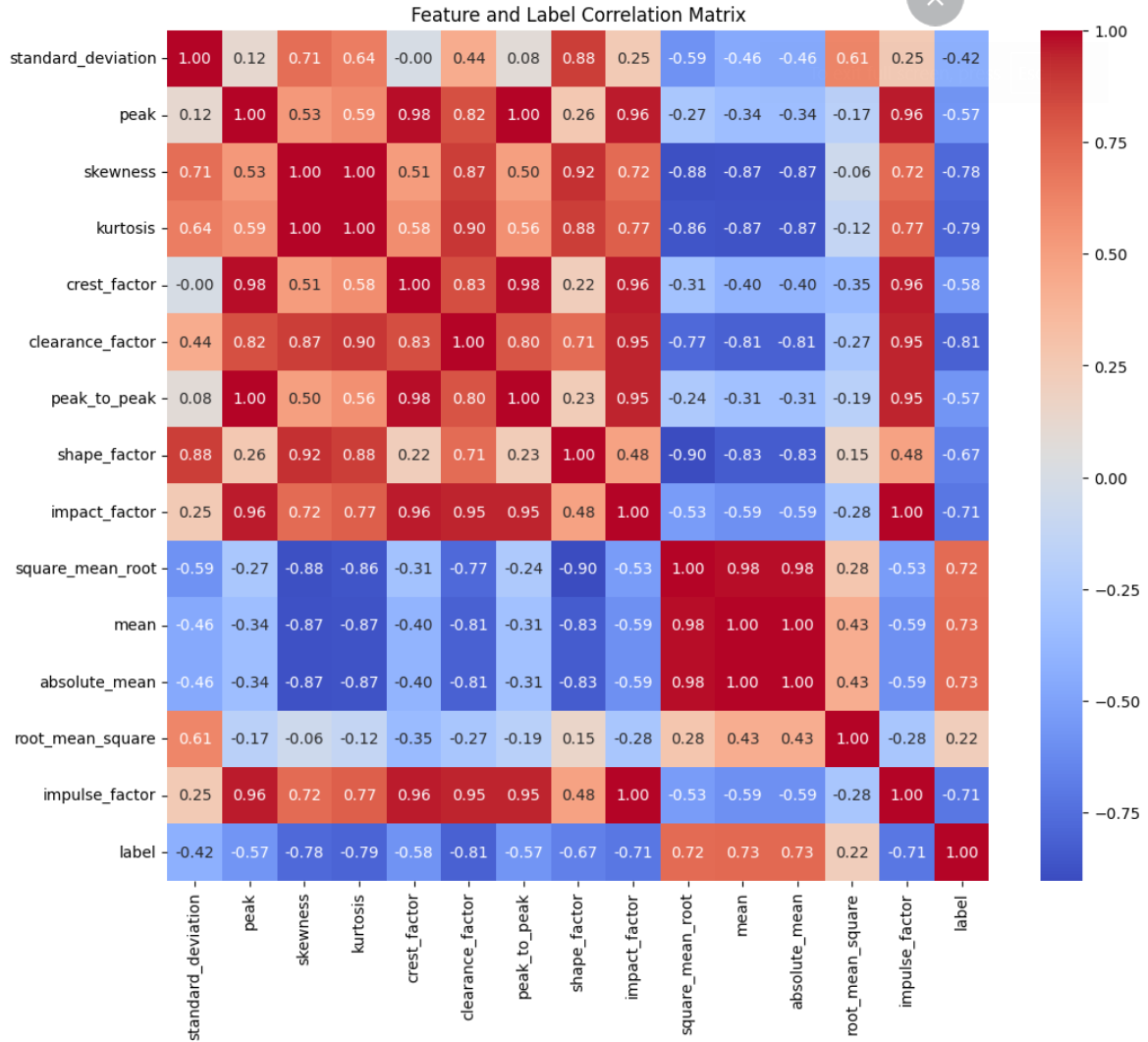
سپس برای هر سطر (هر 200 داده) 14 ویژگی زیر را که در تمرین هم زدیم با تعریف کلاسی در پایتون استخراج می کنیم:



و سپس آنها را به با اضافه کردن لیبل مربوطه به هم می چسبانیم:



نتیجه 100 سمپل برای هر کلاس و 15 ویژگی (با لیبل)



نمودار همبستگی بین ویژگی ها و لیبل را مشاهده می کنیم :

مطلوب است corr بین هدف و ویژگی زیاد و بین ویژگی ها با هم کم با شد.

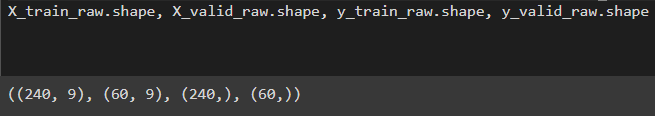
در بالا مشاهده می شود که ['standard\_deviation', 'peak', 'crest\_factor', 'peak\_to\_peak', 'root\_mean\_square'] با هدف همبستگی بالایی ندارند پس حذفشان می کنیم.



همچنان میتوانستیم ویژگی ها با CORR بالا باهم را (redundancy بالا)نیز حذف کنیم. که نکردیم.

داده ها را به train و validation با درصد تقسیم 80% برای 9 نوامبر تقسیم می کنیم.

همچنین آرگمانshuffle را نیز برای بر زدن فعال می کنیم:

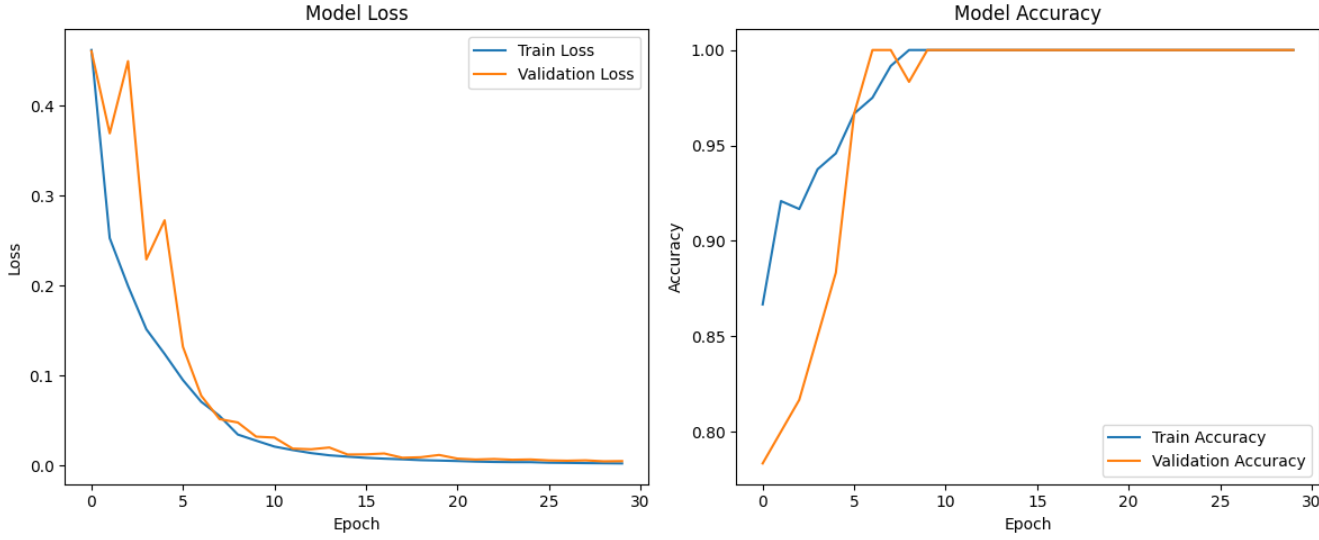


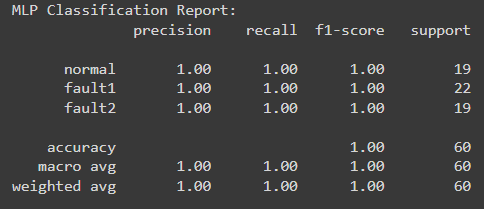
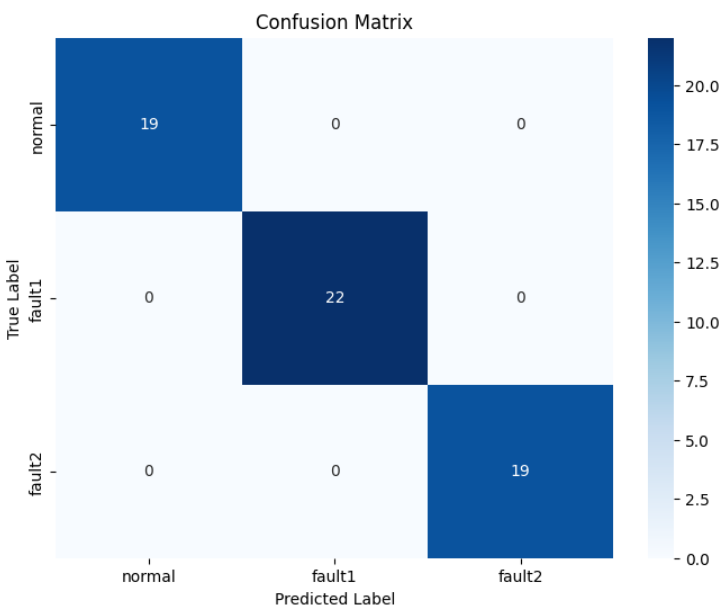
حال آنها را استاندارد سازی می کنیم.

برای استاندارد سازی آنها باید به داده های آموزش فیت و به تست اعمال بشه متاسفانه وقت نشد داده های تست را هم آماده کنم برای این کار.

(لیبل ها را one-hot کردیم)

MLP با 1 لایه پنهان Relu با 20 نرون و لایه نهایی softmax زدم. Learning rate را 0.01 و تعداد epoch را 30 با bachsize 10 گذاشتم. همچنین تابع اتلاف categorical crossentropy (برای one-hot ) و optimizer را adam گذاشتیم .

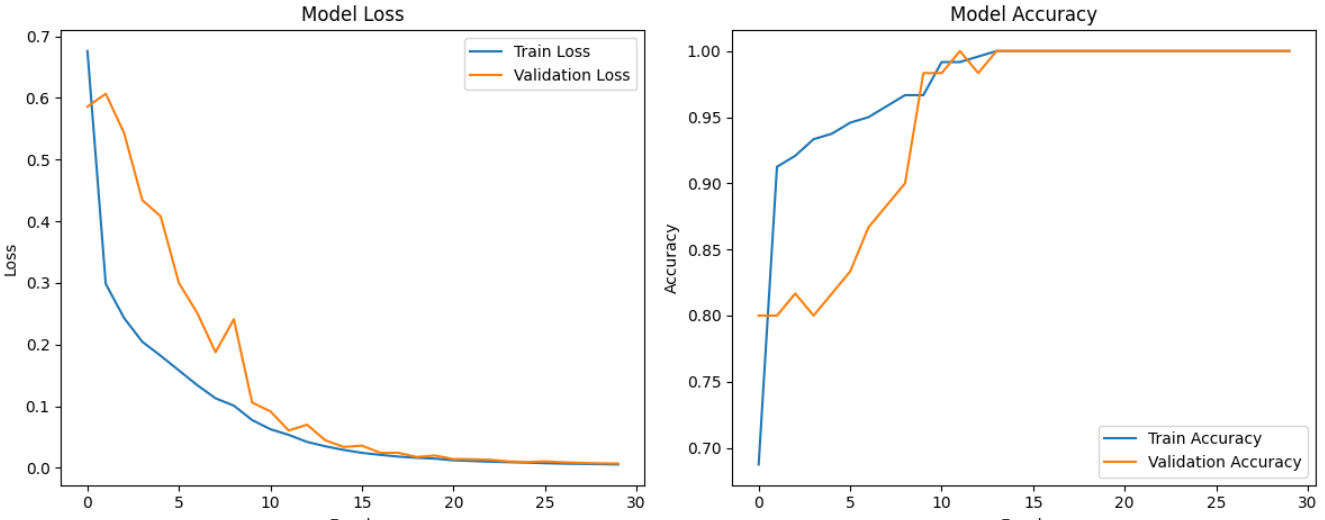


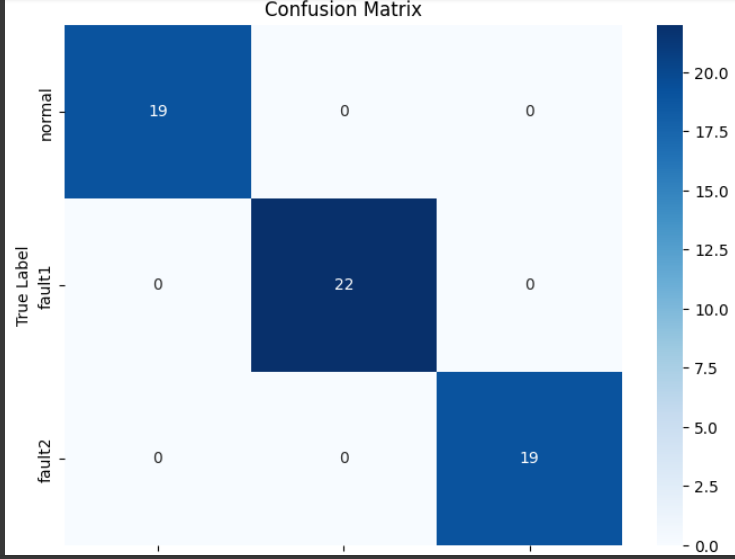
 

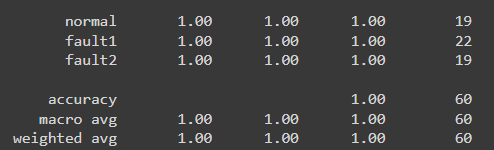
در بالا نمودار خطا accuracy ،classification report و confusion matrix آورده شده است

واضح است که دقت و خطا عالی بوده و داده missclasified نداریم

برای 10 نرون:



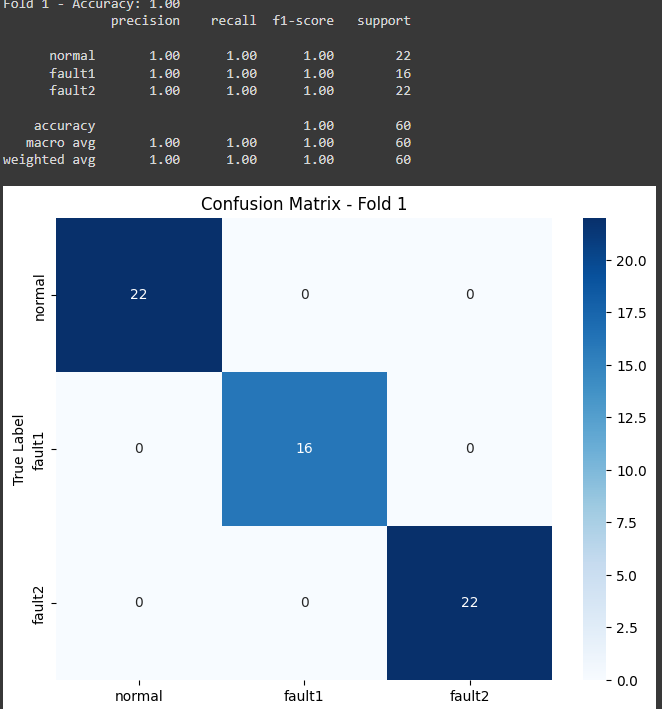




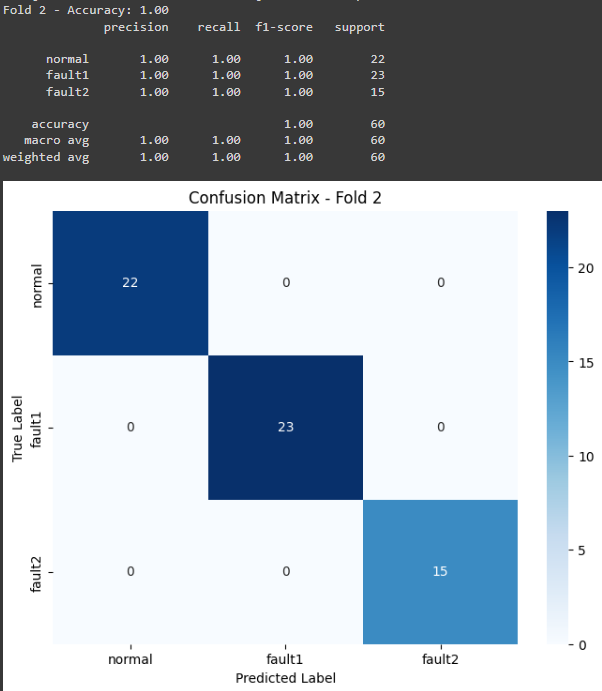
متاسفانه در preprocess داده های تست دچار مشکل شدم و نتوانستم این قسمت سوال را انجام بدم

K-fold

برای fold1:

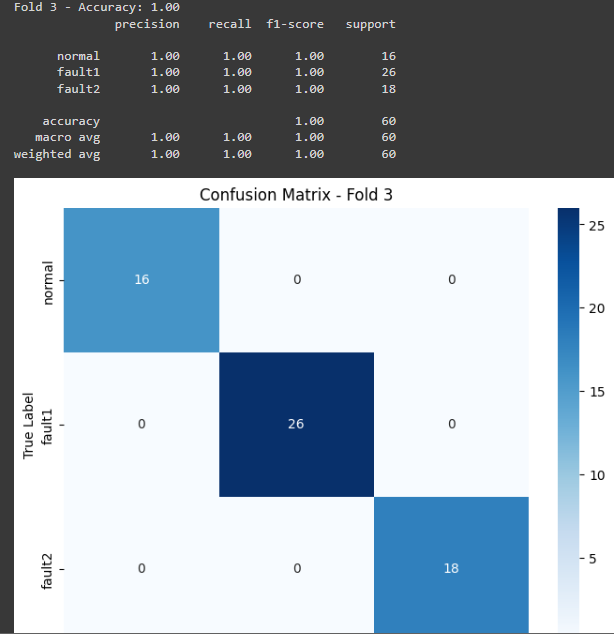


برای fold2:

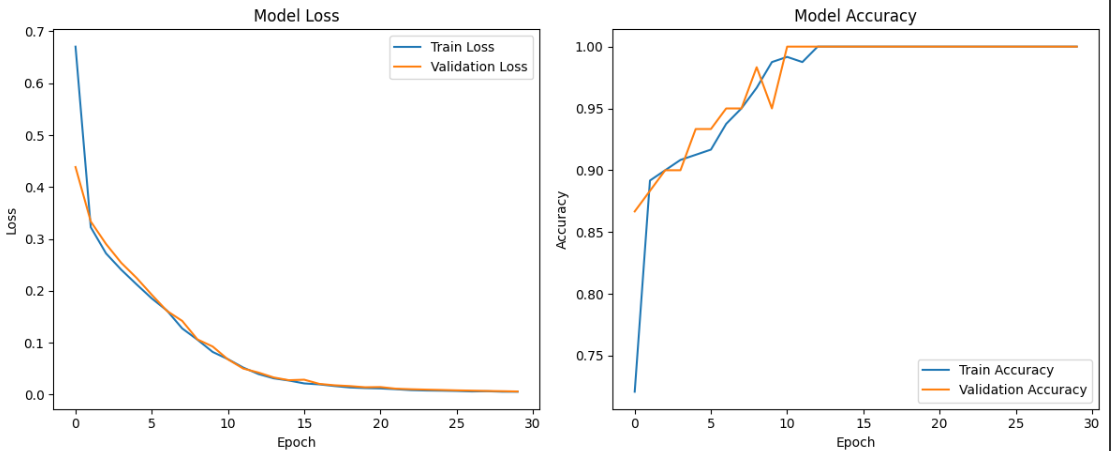


برای

fold3:



برای بقیه fold ها هم به همینطور بود.



Mean loss و mean accuracy را مشاهده می کنیم .