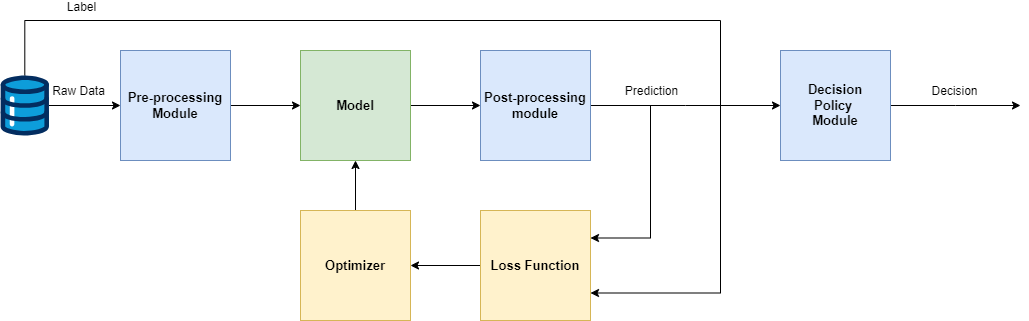
سوال 1

1.  فرآیند آموزش و ارزیابی یک مدل طبقه بند خطی را به صورت دیاگرامی بلوکی نمایش دهید و در مورد اجزای مختلف این دیاگرام بلوکی توضیحاتی بنویسید. تغییر نوع طبقه بندی از حالت دوکلاسه به چندکلاسه در کدام قسمت از این دیاگرام بلوکی تغییراتی ایجاد می کند؟ توضیح دهید.

Dataset:این قسمت شامل داده‌های خام و دسته‌بندی متناظر با آن می‌باشد. کیفیت و کمیت مجموعه داده‌ به صورت موثری روی عملکرد مدل یادگیری ماشین تاثیر خواهد گذاشت.

Pre-Processing Module: این بخش از ساختار یادگیری ماشین داده‌های خام را دریافت می‌کند و متناسب با وظیفه (Task) ساختار و نوع داده ورودی، با انجام انواع گوناگون تبدیل‌ها (Transforms) مانند حذف داده‌های پرت، آستخراج ویژگی و فیلترگذاری مناسب داده‌ها را برای پردازش و آموزش الگوریتم یادگیری ماشین بهبود می‌بخشد.  بخش پیش‌پردازش یک ساختار یادگیری ماشین می‌تواند روی کیفیت و کمیت یک مجموعه داده تاثیرگذار باشد، به عنوان نمونه با انجام تبدیل‌های ساخت داده مصنوعی (Augmentation) می‌توان تا حدی کمیت مجموعه داده را جبران کرد و با استفاده از این تبدیل‌ها به صورت وزن‌دار می‌توان برخی از مشکلات کیفی مجموعه داده مانند عدم تعادل (imbalance) را برطرف کرد.

Model: این بخش از ساختار یادگیری ماشین شامل الگوریتم یادگیری ماشین می‌باشد که با توجه به وظیفه (Regression, Classification, Segmentation) و ویژگی‌های متفاوت مجموعه داده مانند کیفیت و بار محاسباتی مطلوب، می‌تواند از میان انواع گوناگون انتخاب شود.

Post-Processing Module: این بخش از ساختار یادگیری ماشین شامل روش‌هایی است که با توجه به خروجی الگوریتم یادگیری ماشین می‌تواند انتخاب شود تا پیشبینی‌های اشتباهی که به صورت آماری از یک الگوی خاص پیروی می‌کنند اصلاح شوند.

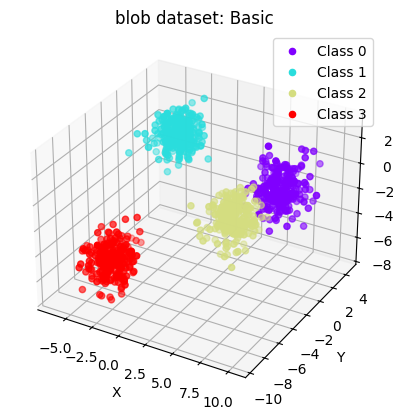
 Loss Function: این بخش از ساختار یادگیری ماشین شامل انواع گوناگونی از توابع هزینه مانند MSE، Cross-Entropy و یا انواع دیگر توابع هزینه می‌شود. تابع هزینه خصوصا برای الگوریتم‌هایی که شامل بهینه‌سازی از طریق مشتق هستند دارای اهمیت ویژه‌ای می‌باشد زیرا توابع هزینه‌ای که مشتق‌پذیر نیستند یا مشتق آنها در نواحی خاصی خیلی زیاد یا خیلی کم می‌باشد برای فرآیند آموزش مشکل‌ساز خواهند بود. تابع هزینه می‌تواند بعضی از مشکلات کیفی مجموعه داده مانند عدم تعادل را برطرف کند.

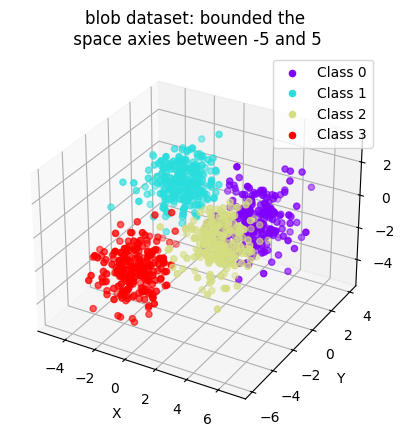
Optimizer: این قسمت از ساختار یادگیری ماشین شامل راه حل ریاضی برای بهینه‌سازی وزن‌های داخلی مدل یادگیری ماشین می‌باشد. با درنظر گرفتن ویژگی‌های مجموعه‌داده و انتخاب صحیح الگوریتم بهینه‌سازی می‌توانیم زمان آموزش و عملکرد ساختار یادگیری ماشین را بهبود ببخشیم. یک الگوریتم بهینه‌سازی مناسب باعث می‌شود تا مدل یادگیری ماشین از نقاط بهینه محلی عبور کند و به نقاطی برود که تابع هزینه برای آنها مقدار کمتری را نشان می‌دهد.

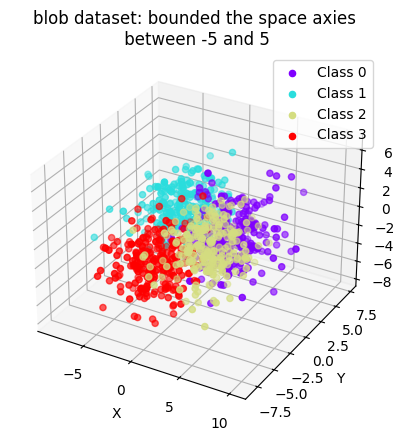
 Decision Policy Module: این قسمت از ساختار یادگیری ماشین باتوجه‌به وظیفه ساختار، یکی از انواع تبدیل‌های موجود را روی خروجی پیاده می‌کند. این تبدیل برای وظیفه طبقه‌بندی (Classification) می‌تواند تبدیل خروجی احتمالاتی به خروجی دو وضعیتی باشد که از آن با عنوان One-Hot Transform یاد می‌شود. همچنین در این بخش معیارهای ارزیابی ساختار مانند Recall، Precision و یا Accuracy محاسبه می‌شود.

 به‌منظور تغییر طبقه‌بندی از ۲ کلاس به چند کلاس، می‌تواند در تمام قسمت‌های ساختار تأثیر بگذارد بااین‌حال مهم‌ترین قسمت شامل Loss Function و Optimizer است که باید برای استفاده به‌عنوان یک طبقه‌بند چند کلاسه با شرایط خروجی مدل سازگاری داشته باشند. به‌عنوان نمونه BCELoss در کتابخانه PyTorch به‌منظور استفاده در یک مسئله طبقه‌بندی چند کلاسه مطلوب نیست و همچنین در مدل LogisticRegression از کتابخانه sklearn اگر برای وظیفه طبقه‌بندی چند کلاسه استفاده شود، نمی‌تواند از liblinear به‌عنوان solver استفاده کند. در انتها به‌منظور تبدیل مدل از طبقه‌بند دو کلاسه به چند کلاسه می‌توان دو رویکرد multinomial یا one-vs-rest استفاده کرد که هنگام استفاده از رویکرد one-vs-rest، وظیفه یک طبقه‌بندی چند کلاسه به چندوظیفه طبقه‌بندی دو کلاسه تبدیل می‌شود و محدودیت‌های فوق‌الذکر برای این رویکرد وجود ندارد.

1. با استفاده از datasets.sklearn، یک دیتاست با ۱۰۰۰ نمونه، ۴ کلاس و ۳ ویژگی تولید کنید و آن را به صورتی مناسب نمایش دهید. آیا دیتاستی که تولید کردید چالش برانگیز است؟ چرا؟ به چه طریقی می توانید دیتاست تولیدشده خود را چالش برانگیزتر و سخت تر کنید؟

به منظور ساخت یک مجموعه داده طبقه‌بندی از کتابخانه sklearn و متد make\_blobs استفاده شده است. نتیجه ساخت یک مجموعه داده متعادل با 4 کلاس و 3 ویژگی با تنظیمات پیش‌فرض متد make\_blobs در تصویر زیر نمایش داده شده است.

همانطور که از این تصویر مشخص است تفکیک‌پذیری داده‌ها در این مجموعه داده پالش برانگیز نمی‌باشد و تنها چالش این مجموعه داده جدا سازی نقاط بنفش (class0) از نقاط زرد (class2) می‌باشد. در ادامه به منظور چالش برانگیز کردن این مجموعه داده از میان پارامترهای قابل تنظیم در متد make\_blobs، آرگمان center\_box=[-5, 5] قرار داده می‌شود. نتیجه این تنظیمات مجموعه داده زیر می‌باشد.

در این مجموعه داده با کم شدن محدوده مجاز حضور نقاط در فضا، باعث نزدیک شدن مرکز دسته‌های هر کلاسه به یکدیگر شده است و در نتیجه آن چالش جداسازی نقاط از هم و تشخیص درجه تعلق آنها به هر دسته چالش برانگیزتر شده است اما همچنان جداسازی این داده‌ها از یکدیگر می‌تواند چالش برانگیزتر باشد. با تغییر انحراف از معیار مربوط به دستور ساخت مجموعه داده می‌توان چالش این مجموعه داده را افزایش داد. در ادامه احراف از معیار مجموعه داده برابر 2 قرار داده شده است و نتیجه آن در شکل زیر قابل مشاهده است.