- **مسئله اول- طبقه‌بندی چندکلاسه**

در این مسئله، شما یک طبقه‌بند برای تشخیص ارقام دست‌نویس موجود در دیتاست MNIST خواهید ساخت. همانطور که می‌دانید، الگوریتم KNN یک طبقه‌بند ذاتا چندکلاسه است. در مقابل اما الگوریتم SVM دوکلاسه است و نمی‌تواند در حالت اولیه برای حل مسائل چندکلاسه بکارگرفته شود. در این مسئله شما با پیاده‌سازی تکنیک one-vs-one این نقصان را در الگوریتم SVM برطرف خواهید نمود.

بخش ۱) دیتاست MNIST را با استفاده از دستور زیر بارگذاری کنید:

dataset = np.load('./mnist.npz')

در این دیتاست، فیلدهای X و y قرار گرفته است که به‌ترتیب تصاویر ارقام دست‌نویس و عدد متناظر هر تصویر است.

بخش ۲) با استفاده از مدل StandardScaler داده‌ها را نرمالیزه نمایید. دقت کنید که به‌این‌منظور باید این مدل را فقط روی دادگان آموزش fit کرده و سپس برروی دادگان آموزش و تست transform نمود.

بخش ۳) با استفاده از یک مدل KNN، هر ۱۰ کلاس این دیتاست را طبقه‌بندی نمایید. با تنظیم مقدار K در مدل KNN، دقت مدل را روی دادگان آموزش و تست محاسبه نموده و این دو مقدار را به ازای Kهای مختلف روی یک نمودار رسم نمایید. براساس نتایج، کدام مقدار K بهترین نتیجه را می‌دهد؟

بخش ۴) ابتدا دو کلاس از ۱۰ کلاس موجود در دیتاست را انتخاب نموده و یک مدل SVM را با پارامترهای دلخواه بر روی آن آموزش دهید. سپس دقت مدل را با روش K-Fold Cross-Validation محاسبه نمایید.

بخش ۵) با ساختن یک مدل SVM برای تمامی زوج کلاس‌های متمایز در دیتاست MNIST و مقایسه عملکرد آن‌ها، بنظر شما تشخیص کدام دو کلاس برای مدل از همه سخت‌تر است؟ چرا؟ (راهنمایی: در این بخش شما ۴۵ مدل SVM با ورودی‌های مختلف می‌سازید!)

بخش ۶) برای طبقه‌بندی چندکلاسه با استفاده از SVM، مشابه بخش قبل یک مدل SVM به ازای هر ترکیب متمایز از زوج کلاس‌ها بسازید (۴۵ مدل SVM) و آن را با استفاده از داده‌های متناظر fit کنید. این ۴۵ SVM درواقع مدل نهایی شما را تشکیل می‌دهند!

برای پیش‌بینی یک ورودی، باید آن را به تمامی ۴۵ مدل بدهید و با بررسی خروجی مدل‌ها، محاسبه نمایید که کدام کلاس بیشترین رای را کسب کرده است. بدین‌ترتیب خروجی مدل مشخص می‌گردد.

بخش ۷) عملکرد این مدل را از منظر دقت با بهترین الگوریتم KNN بخش ۳ مقایسه نمایید.

- **مسئله دوم- پیش‌بینی مرگ و میر**

در این مسئله، ما تلاش میکنیم که با استفاده از داده‌های پزشکی ثبت‌شده از بیماران قلبی، مردن یا زنده‌ماندن افراد را پیش‌بینی کنیم. همچنین تلاش میکنیم تا با تکنیک forward selection، مهم‌ترین عوامل پیش‌بینی کننده مرگ را بیابیم.

بخش ۱) دیتاست heart_failure.csv را در pandas بارگذاری نمایید و ستون‌های آن را به دقت بررسی کنید. بدین‌منظور می‌تواند از دستور df.describe() استفاده نمایید. سپس ۹۰ نمونه بصورت تصادفی از گروه افراد مرده و ۹۰ نمونه تصادفی از افراد زنده انتخاب نمایید. (نمونه‌برداری باید بدون جایگذاری باشد. برای این کار میتوانید از تابع np.random.choice استفاده کنید.)

بخش ۲) توزیع ستون‌های مختلف را در میان دو گروه زنده و مرده رسم کنید. با مقایسه این نمودارها، بنظر شما کدام متغیر بیش از همه در تعیین زنده یا مرده بودن بیمار اثرگذار است؟

بخش ۳) یک مدل SVM با پارامترهای دلخواه برای پیش‌بینی زنده یا مرده بودن هر نمونه بسازید. سپس با استفاده از GridSearchCV، بهترین مجموعه پارامترها را بیابید. دقت بهترین مدل چقدر است؟

بخش ۴) یکی از ستون‌های دیتاست را انتخاب نمایید و تنها با استفاده از آن، تلاش کنید که متغیر زنده یا مرده بودن را پیش‌بینی کنید. برای این پیش‌بینی از مدل SVM با پارامترهای یافت‌شده در بخش قبل استفاده کنید. دقت پیش‌بینی این مدل را با روش K-Fold Cross-Validation ارزیابی نمایید.

بخش ۵) با تکرار بخش قبل برای تمامی ستون‌ها، کدام متغیر بهترین عملکرد را در پیش‌بینی زنده یا مرده بیمار بودن دارد؟ چرا؟

بخش ۶) حال برای پیش‌بینی مرگ از دو متغیر استفاده کنید که یکی از آن‌ها متغیر بدست‌آمده در بخش قبل است. ترکیب کدام دو متغیر دقیق‌ترین پیش‌بینی را از مرده یا زنده بودن بیماران در بر دارد؟ آیا می‌توان گفت که این دو متغیر، مهمترین عوامل دخیل در مرگ‌و میر بیماران قلبی است؟

با آرزوی موفقیت برای شما