تمرین اول - یادگیری ماشین

در این تمرین شما باید با دانلود یک دیتاست ساده، مسئله مطرح شده را به کمک یادگیری ماشین حل کنید. دیتاست مورد نظر، با دریافت تعدادی ویژگی از بیماران، وجود یا عدم وجود سرطان سینه در آنها را بررسی کرده است. این دیتاست را می‌توانید از لینک زیر دریافت کنید:
https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra#
توضیحات مربوط به این دیتاست و همچنین لینک دانلود دادگان در این سایت قرار دارد.
برای انجام این تمرین ابتدا توضیحات مربوط به دیتاست و ستون‌های مختلف را مطالعه کنید. سپس با دریافت داده، برای هر فیچر نمودارهایی رسم کنید. هدف از رسم این نمودارها درک بهتر دادگان است. توصیه می‌شود برای درک بهتر این دیتاست، نمودار تک تک ویژگی ها بر مبنای کلاس را رسم کنید (محور افقی انیس، محور عمودی یکی از ویژگی ها و هر کلاس را با یک رنگ مشخص کنید.) همچنین می‌توانید نمودار هایی برای هر زوج ویژگی نیز رسم کنید (محور افقی ویژگی اول، محور عمودی ویژگی دوم و هر کلاس با یک رنگ مشخص می‌شود).

پس از رسم این نمودارها، تحلیل خود در مورد این نمودارها را بنویسید. به نظر شما از این نمودارها چه چیزی را در مورد دادگان می‌توان فهمید؟ آیا دادگان ما به صورت خطی تفکیک پذیر هستند؟

سپس با استفاده از مدل‌های مختلف یادگیری ماشین، اقدام به حل این مسئله کنید. در این تمرین نیازی به استفاده از شبکه‌های عصبی نیست.
برای حل این دیتاست، با کمک کتابخانه scikit-learn مدل‌ها زیر را بر روی این مجموعه داده امتحان کنید و دقت نهایی خود را گزارش کنید:

1. Logistic Regression
2. SVM with linear kernel
3. SVM with rbf kernel
4. Decision Tree
5. KNN (مقدار k را تنظیم کنید طوری که دقت بهتری به دست بیاورید)

همچنین برای ارزیابی این مدل‌ها، از روش kfold cross validation استفاده کنید (مقدار k=5) و در نهایت برای مقایسه مدل‌ها، میانگین دقت تست بر روی این ۵ فولد را در نظر بگیرید.

زمان تحویل: **جمعه ۷ مرداد ماه تا ساعت ۱۲ بامداد.**
ارسال پاسخ به:
mailto:aimedic.internship@gmail.com?subject=project1
حتما موضوع ایمیل را project1 قرار دهید و در متن ایمیل حتما نام خود را به فارسی ذکر کنید تا قابل جست و جو باشد.