این تکلیف در مورد مقایسه عملکرد چندین الگوریتم ماشین نظارت شده برای تشخیص spam است.
مجموعه داده را از https://archive.ics.uci.edu/m|/datasets/spambase : دانلود کنید.

۱. این داده ها چند ویژگی دارند؟ توضیحات مختصری در مورد ویژگی ها ارائه دهید. ارتباط بین ویژگی ها و متغیر هدف (برچسب) چیست؟ آیا همه ویژگی ها مفید هستند یا آموزنده؟

۲. داده ها را به مجموعه آموزشی و آزمایشی تقسیم کنید. از 30% داده برای آزمایش استفاده کنید.
چندین الگوریتم ML را آموزش دهید، از جمله: رگرسیون لجستیک، KNN، Naive Bayes، Decision trees، adaboost، جنگل تصادفی،) SVMهم خطی و هم غیرخطی) بر روی داده های آموزشی و استفاده از اعتبارسنجی متقابل برای انتخاب بهترین مقادیر برای پارامترهای هر الگوریتم بهترین پارامترها برای هر الگوریتم چیست؟ خطاهای آموزشی چیست؟(حداقل ۲ مورد را انتخاب کنید.)

۳. الگوریتم های آموزش دیده را روی داده های تست اعمال کنید. کدام الگوریتم بهترین عملکرد را دارد؟ نتایج را تفسیر کنید. آیا حذف ویژگی های غیر اطلاعاتی به بهبود نتایج کمک می کند؟ تفاوت ها را توضیح دهید و نشان دهید.

دقت، زمان آموزش (بر حسب ثانیه) و زمان تست (بر حسب ثانیه) را در جدولی برای هر یک از الگوریتم ها گزارش کنید.