لطفا نکات زیر را رعایت کنید:

۱) فایل گزارش به همراه تمامی کدها در یک فایل فشرده و با عنوان HW#3_StudentNumber در سایت بارگذاری کنید.

۲) بخش‌های پیاده سازی مربوط به هر سوال را در فایل مربوطه با شماره‌ی آن سوال و در پوشه‌ای برای آن سوال قرار دهید.

۳) از زبان برنامه‌نویسی پایتون در یکی از محیط‌های Jupyter notebook و یا google colab برای کد نویسی و تست کدهای خود استفاده کرده و فایل کدها را با فرمت IPYNB ارسال کنید.

۴) گزارش نهایی باید شامل توضیح پیاده‌سازی و نتایج و تحلیل‌های خواسته‌شده در متن تمرین باشد. توجه کنید که در گزارش نهایی خود به تمامی سوال‌های پرسیده شده در متن تمرین -به‌خصوص در بخش عملی تمرین- پاسخ دهید. (می‌توانید توضیحات سوالات را در همان فایل IPYNB و توضیحات هر بخش را در ذیل کد مربوطه بنویسید)

۱. **دیتاست *"fruit data with colors"* را در نظر گرفته و خواسته‌های ذکر شده را بر روی آن پیاده کنید:**

الف) داده‌ها را خوانده و میزان یکنواختی آن‌ها را در کلاس‌های موجود بررسی نمایید. با توجه به نتیجه، برای آموزش مدل چه چالش‌ها و راه حل‌هایی وجود دارد؟

ب) هیستوگرام مربوط به ویژگی‌های این داده‌ها را رسم نمایید. چه نتیجه‌ای می‌توان از این نمودارها گرفت؟

پ) از الگوریتم KNN برای آموزش داده‌ها و با انتخاب معیار مناسب نتیجه را ارزیابی کنید. مقدار k بهینه را در یک حالت به دست آورده و در طول انجام مراحل بخش (پ) آن را تغییر ندهید. مراحل زیر را انجام دهید:

۱- از تمامی ویژگی‌ها برای آموزش استفاده نمایید.

۲- سه ویژگی را برای آموزش انتخاب نمایید.

۳- از یک ویژگی برای آموزش استفاده کنید.

دلیل انتخاب ویژگی‌های مورد نظر را در هر مرحله توضیح داده و نتایج به دست آمده را تفسیر کنید.

ت) یکی دیگر از الگوریتم‌های طبقه‌بندی که در کلاس مطرح شده است را انتخاب کرده و برای بهترین حالت مرحله قبل آن را آموزش دهید. دلیل انتخاب خود را توضیح داده و نتایج را مقایسه کنید.

۲. الف) با استفاده از الگوریتم Naive Bayes و دیتاست spam مدلی برای تشخیص ایمیل‌های spam بسازید. با نسبت ۸۰-۲۰ داده‌های تست و آموزش را جدا کنید.

ب) بخش قبل را با KNN انجام داده و از بین k=5, 10, 20 بهینه آن‌ها را انتخاب نمایید.

پ) برای داده‌های تست هر مدل ماتریس درهم‌ریختگی را نمایش داده و precision و recall را گزارش کنید.

ت) بهترین مدل خود را برای متن زیر استفاده نموده و نتیجه را گزارش کنید. (ویژگی‌ها را باید مطابق فایل‌های راهنمای دیتاست بسازید)

"Your email address was selected to receive a prize money of $500,000.00 in the 2021 European lottery! To claim your prize for free please contact our staff at Logas, Nigeria. The prize has a transfer fee of $500 that must be paid upfront."

۳. می‌خواهیم برای دیتاست Fashion MNIST یک مدل طبقه‌بندی ایجاد کنیم.

الف) با استفاده از الگوریتم‌های زیر و داده‌های آموزش این دیتاست، مدل‌های طبقه‌بندی مناسبی به دست آورده و بر روی داده‌های test ارزیابی نمایید. (پیش‌پردازش مناسب ممکن است به بهبود نتایج کمک کند.)

- Decision Tree
- Linear Classifier
- Logistic Regression
- SVM

ب) با استفاده از روش‌های Ensemble Learning **همچون** Ada-boost، مدل دیگری بسازید و با بهترین نتیجه بخش قبل مقایسه کنید. (دلیل انتخاب مدل‌های پایه خود را توضیح دهید)

۴. داده‌های مربوط به ساعات مطالعه و استراحت تعدادی دانش‌آموز در جدول زیر آورده شده‌اند. می‌خواهیم با استفاده از الگوریتم Logistic Regression یک طبقه بندی دو کلاسه بر روی آن‌ها انجام دهیم. با استفاده از تابع هزینه اتلاف کراس آنتروپی و روش گرادیان نزولی، **به صورت دستی (بدون استفاده از کدنویسی)** این کار را انجام داده و وزن‌های مناسب را به دست آورید.

نتیجه امتحان	ساعات استراحت	ساعات مطالعه
قبول	۸	۸
قبول	۸	۶
مردود	۱۰	۲
قبول	۶	۱۲
مردود	۱	۳
مردود	۱۰	۰/۵

راهنمایی: معادلات تابع هزینه و تابع سیگموید به شرح زیر می‌باشند:

$$L = -\frac{1}{N}\sum_{i=1}^{N} y_t Log y_p + (1 - y_t) Log(1 - y_p)$$

$$y_p = \frac{1}{1 + e^{-(Wx+b)}}$$