به نام پروردگار بخشنده و مهربان

درس یادگیری ماشین و بازشناسی الگو – تمرین شماره ۲

امین عباخواه درآباد **۴۰۲۱۱۵۰۶۴** **amin.abakhah@aut.ac.ir**

پاسخ سوال ۱:

الف) با استفاده از تابع CountVectorizer در کتابخانه sklearn، بردار ویژگی‌ها را می‌سازیم و تعداد ویژگی‌ها را محدود به ۱۰۰ عدد می‌کنیم تا محاسبات سنگین نشود.

ب) حال با استفاده از RandomForest Classifier برای روش Bagging و AdaBoost Classifier برای روش Boosting و ارزیابی بصورت 3-Fold Cross-Validation، ماتریس Confusion و همچنین مقادیر Error، Precision، Recall و F-measure را بدست می‌آوریم.

در اینجا ما از فایل HW2_AUT_MLPR_4021-2-Email-SPAM (2) استفاده کردیم. با مشاهده نتایج متوجه می‌شویم که مشکلی در دیتاست ما وجود دارد چرا که با توجه به اینکه مسئله ما تشخیص اسپم است، ما دو لیبل داریم و باید ماتریس ما ۲ در ۲ باشد اما می‌بینیم که ماتریس Confusion ما ۵ در ۵ هست. (تصویر زیر)

```
Bagging Metrics:
/usr/local/lib/python3.10/dist-packages/sklearn/model_selection/_split.py:700: UserWarning: The least populated class in y has only 1 members, which is less than n_splits=3.
  warnings.warn(
Confusion Matrix:
 [[   0    0    1    0    0]
 [   0    0    0    1    0]
 [   0    0 4289   69    0]
 [   0    0   92 1276    0]
 [   0    0    0    0    2]]
Precision: 97.11%
Recall: 97.16%
F-measure: 97.13%
Error: 2.84%

==============================

Boosting Metrics:
/usr/local/lib/python3.10/dist-packages/sklearn/model_selection/_split.py:700: UserWarning: The least populated class in y has only 1 members, which is less than n_splits=3.
  warnings.warn(
Confusion Matrix:
 [[   0    1    0    0    0]
 [   1    0    0    0    0]
 [   0    0 4358    0    0]
 [   0    0 1368    0    0]
 [   0    0    0    0    2]]
Precision: 57.92%
Recall: 76.09%
F-measure: 65.77%
Error: 23.91%
```

همچنین پیام warning داده شده به ما می‌گوید که برخی کلاس‌ها هستند که ۱ عضو دارند در حالیکه دیتای ما چند هزار عضو دارد که بخشی لیبل صفر دارد و بخشی لیبل ۱. از ماتریس هم می‌توان متوجه شد که ۳ کلاس اضافه وجود دارد که ۲ تای اول، فقط یک عضو داشته و آخری ۲ عضو دارد. لذا باید دیتاست را بررسی کنیم. بعد از بررسی دیتاست متوجه می‌شویم که ۴ سطر وجود دارد که مشکل دارد و طبق دیتاست لیبل درستی نداشته و بجای عدد صفر، رشته وجود دارد. در اینجا ما لیبل آن ۴ سطر را برابر صفر قرار دادیم

و دیتاست جدید را با نام HW2_AUT_MLPR_4021-2-Email-SPAM (3) ذخیره کرده و کد را اجرا کردیم و خروجی زیر را گرفتیم. (تصویر زیر)

```
Bagging Metrics:
Confusion Matrix:
 [[4295   67]
 [  93 1275]]
Precision: 97.19%
Recall: 97.21%
F-measure: 97.20%
Error: 2.79%

==============================

Boosting Metrics:
Confusion Matrix:
 [[4204  158]
 [  82 1286]]
Precision: 95.93%
Recall: 95.81%
F-measure: 95.85%
Error: 4.19%
```

همانطور که می‌بینیم، حال نتایج بدست آمده صحیح بوده و ماتریس Confusion ما ۲ در ۲ است. مقادیر محاسبه شده نیز آورده شده است که از دقت بالایی برخوردار هستند. با توجه به نتایج ارزیابی، می‌بینیم که در اینجا روش Bagging بهتر از روش Boosting عمل کرده است.

*توجه : لطفا برای اجرا، یا به تمام جزئیات ذکر شده توجه کنید و یا از فایلهای Excel ای که در اختیارتان قرار داده شده است استفاده نمایید تا اشتباهی رخ ندهد.

پاسخ سوال ۲:

الف) در ابتدا تاریخ فایل Stock-Index را به تاریخ میلادی تبدیل می‌کنیم. مابقی فایلها، تاریخشان میلادی است یا میلادی هم دارند. لذا داده‌ها را در ابتدا بر اساس تاریخ، بصورت صعودی (از گذشته به آینده) مرتب می‌کنیم. سپس از هر فایل، ستون Close (یا همان قیمت آخر روز یا همان بسته شدن) را جدا کرده و از فایل Stock-Index نیز، ستون Value را برمی‌داریم اما نام تمام این ستون‌ها را Close می‌گذاریم تا در هنگام merge آنها بتوانیم با یک حلقه به راحتی آنها را ادغام کنیم. ادغام را با روش outer انجام می‌دهیم تا تمام رکوردهای همه فایلها در فایلِ ادغام شده وجود داشته باشند. بعد از ادغام، فایل جدید را با نام Dataset ذخیره می‌کنیم. حال برای راحتی کار، بصورت دستی در برنامه Excel، نام هر کدام از ستون‌های Close را به نام دارایی خودش تغییر می‌دهیم. برای مثال، نام ستون Close مربوط به فایل EUR-USD را به EUR-USD تغییر می‌دهیم و الی آخر و فایل را با نام Dataset-Final ذخیره می‌کنیم.

حال با بررسی دیتا متوجه می‌شویم که بعد از ادغام، ترتیب رکوردها بهم خورده لذا مجددا فایل جدید را مرتب می‌کنیم و با نام جدید Dataset-Final-2 ذخیره می‌کنیم.

با مشاهده فایل فعلی، متوجه می‌شویم که تنها فایل مربوط به قیمت دلار از سال ۲۰۱۱ تا ۲۰۲۲ دیتا دارد و مابقی فایلها از سال ۲۰۱۵ تا ۲۰۲۲ دیتا دارند و از آنجایی که تارگت اصلی ما همان قیمت دلار است، در واقع ما باید با استفاده از ستون قیمت Close فایلهای دیگر بعنوان ویژگی، مقدار یا کلاس قیمت دلار را

پیش‌بینی کنیم و از آنجا که دیگر فایلها (در واقع ویژگیها) را در بازه زمانی ۲۰۱۱ تا ۲۰۱۵ نداریم، عملا آن دیتا را نمی‌توانیم استفاده کنیم مگر اینکه بخواهیم خود قیمت دلار را از روی خودش بدست بیاوریم و نه از روی ویژگیهای دیگر فایلها. در اینجا ما چون می‌خواهیم که با استفاده از فایلهای دیگر و در واقع ویژگیها، ستون قیمت دلار را از لحاظ مقدار یا کلاس مشخص کنیم، لذا بازه‌ای را باید انتخاب کنیم که تمام فایلها در آن بازه دیتا داشته باشند که این بازه برابر با ۲۵ مارس ۲۰۱۵ تا ۲۳ اکتبر ۲۰۲۲ است. لذا بصورت دستی و در Excel، صرفا رکوردهای این بازه را نگه می‌داریم و مابقی را حذف می‌کنیم و فایل جدید را با نام Dataset-Final-3 می‌گذاریم.

اما همچنان دیتای ما آماده نیست چون مقادیر Null بسیاری دارد که باید به نحو درستی مدیریت شود که برای رفع این مشکل، مقادیر گم شده را توسط درون‌یابی پر می‌کنیم. قبل از درون‌یابی، برخی مقادیر "–" را تبدیل به NaN می‌کنیم تا تابع درون‌یاب، آنها را بعنوان مقدار null شناسایی کند. سپس درون‌یابی را با روشی که بر اساس زمان است (method='time') که مناسب دیتاهای سری زمانی است انجام می‌دهیم و فایل بدست آمده را با نام Dataset-Final-4 ذخیره می‌کنیم.

حال با توجه به صورت سوال، ستون جدیدی با نام Next Day Date به دیتاست اضافه می‌کنیم و مقادیر آن را در هر سطر برابر با تاریخ رکورد بعدی قرار می‌دهیم و فایل جدید را با نام Dataset-Final-5 ذخیره می‌کنیم.

سپس ستون جدید دیگری با نام Gap به دیتاست اضافه می‌کنیم و مقادیر آن را در هر سطر برابر با فاصله بین تاریخ رکورد فعلی و تاریخ رکورد بعدی قرار می‌دهیم و فایل جدید را با نام Dataset-Final-6 ذخیره می‌کنیم.

در ادامه ستون جدیدی با نام Forecast به دیتاست اضافه می‌کنیم تا در قسمت Classification از آن بعنوان لیبل استفاده کنیم. مقادیر آن را در هر سطر با توجه به افزایش، عدم تغییر یا کاهش قیمت نسبت به روز بعد با مقادیر -۱، ۰، +۱ پر می‌کنیم و فایل جدید را با نام Dataset-Final-7 ذخیره می‌کنیم.

برای بررسی راحت‌ترِ مقادیر قیمت دلار (یعنی USD-IRR) با مقادیر لیبل (یعنی Forecast)، ستون USD-IRR را به جایگاه یکی مانده به آخر و در کنار لیبل منتقل می‌کنیم و فایل جدید را با نام Dataset-Final-8 ذخیره می‌کنیم.

ب) حال با استفاده از الگوریتم‌های SVM، Naïve Bayes، Decision Tree و KNN، کلاس تغییرات روز بعد را پیش‌بینی می‌کنیم و برای ارزیابی، دیتاست را بصورت ۲۰/۸۰ تقسیم می‌کنیم و مقادیر Precision، Recall، F-measure و Accuracy را گزارش می‌کنیم. (تصویر زیر)

```
SVM Metrics:
Precision: 48.64
Recall: 37.59
Accuracy: 37.59
F1 Score: 33.7

Decision Tree Metrics:
Precision: 42.17
Recall: 41.97
Accuracy: 41.97
F1 Score: 41.93

Naive Bayes Metrics:
Precision: 58.73
Recall: 37.04
Accuracy: 37.04
F1 Score: 28.04

KNN Metrics:
Precision: 46.64
Recall: 40.15
Accuracy: 40.15
F1 Score: 40.31
```

ج) حال با در نظر گرفتن ستون قیمت دلار (یعنی USD-IRR) بعنوان تارگت و دیگر ستون‌ها بعنوان ویژگی، می‌توان به کمک الگوریتم‌های رگرسیون، مقدار قیمت را برای روز بعد پیش‌بینی کرد و ارزیابی آن را با مقادیری همچون MSE، MAE، MAPE، RMSE و R2 Score سنجید. در اینجا ما از الگوریتم‌های Linear Regression، Ridge Regression، Decision Tree Regressor و KNN استفاده خواهیم کرد.

```
Linear Regression Metrics:
Mean Absolute Error: 23462.28
Mean Absolute Percentage Error: 30.4 %
Mean Squared Error: 833833695.86
Root Mean Squared Error: 28876.18
R2 Score: 90.7 %

Ridge Regression Metrics:
Mean Absolute Error: 23467.1
Mean Absolute Percentage Error: 30.42 %
Mean Squared Error: 833989808.59
Root Mean Squared Error: 28878.88
R2 Score: 90.7 %

Decision Tree Regressor Metrics:
Mean Absolute Error: 2041.64
Mean Absolute Percentage Error: 1.36 %
Mean Squared Error: 35544536.13
Root Mean Squared Error: 5961.92
R2 Score: 99.6 %

K-Nearest Neighbors Regressor Metrics:
Mean Absolute Error: 5377.18
Mean Absolute Percentage Error: 5.16 %
Mean Squared Error: 123328436.93
Root Mean Squared Error: 11105.33
R2 Score: 98.62 %
```

همانطور که از نتایج پیداست، رگرسیون با استفاده از درخت تصمیم، بهترین نمره و کمترین خطا را دارد.

*توجه: کدهای مربوط به این تمرین، بصورت فایل notebook است و در فضای GoogleColab قابل بارگذاری و اجراست.