Result Table

**– Section 2:** Report the number of subgroups and their unique characteristics.

There are 3 distinct subgroups. Subgroup 0 has 1161 data points, subgroup 1 has 2666, and subgroup 2 has 1980.

The mean values of subgroup 0 indicate that companies in this subgroup tend to have relatively high values for some features (e.g., 'Liability\_Composite', 'Net\_Value\_Per\_Share\_BC'), and moderate to low values for other features. This subgroup has the highest standard deviation amongst all the groups.

On the the other hand, subgroup 1 has moderate mean values for most features, yet features like 'Liability\_Composite' and 'Interest\_Rate\_Composite' have relatively lower mean values compared to other subgroups.

Subgroup 2 shows lower mean values for many features compared to the other subgroups, and features like 'Interest\_Rate\_Composite' and 'Gross\_Margin\_Composite' have relatively low mean values in particular.

**– Section 3 and 4:** Construct a table as shown in Table 2. In the notebook file, present the length of the train data column and the confusion matrix.

| Subgroup ID | Name of Student | Average accuracy score  base models [TT(TF)] | accuracy score  Meta model [TT(TF)] | accuracy  score  k-fold CV [TT(TF)] | N*features* |
| --- | --- | --- | --- | --- | --- |
| 0  1  2 | Justine Wang  Justine Wang  Mann Amin | 0.81[119(28)]  0.98[50(1)]  1.00[147(0)] | 0.92[182(16)]  0.92[182(16)]  0.957[401(18)] | 0.64[126(72)]  0.78[154(44)] | 40  40 |

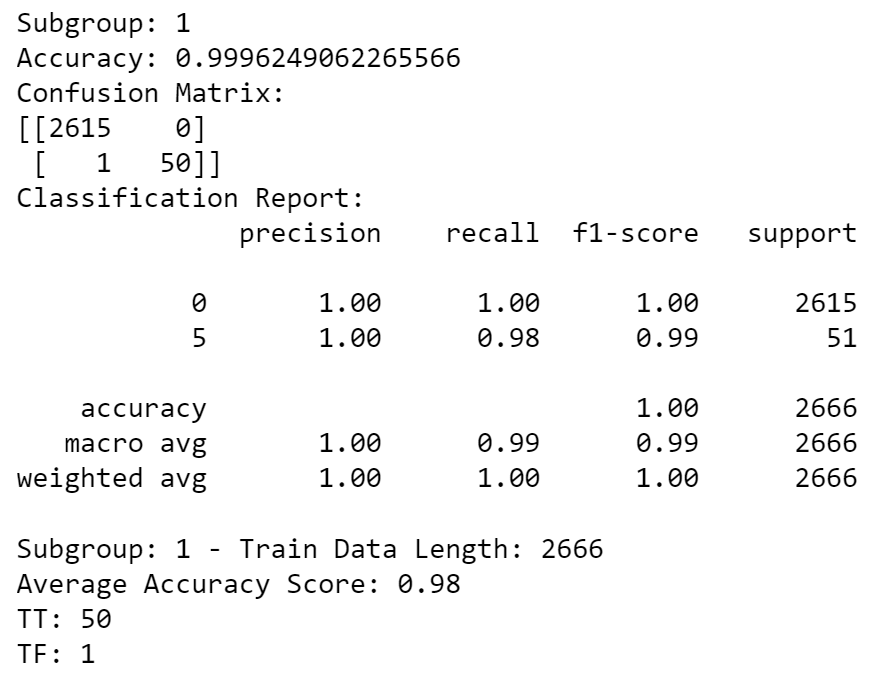
**– Section 5:** Explain the best model from Sections 3 and 4.

The best model from section 3 Stacking Method was the GradientBoostingClassifer() for the base model, as it performed with an acc value of 0.98. The best model from Section 4 was AdaBoostClassifier() for the k-fold cross validation with an accuracy score of 0.78. We can see that the GradientBoostingClassifer() as base with logisticRegression meta-model stacked onto it was the best performing model across the attempts with stacking method and k-fold cross verification.

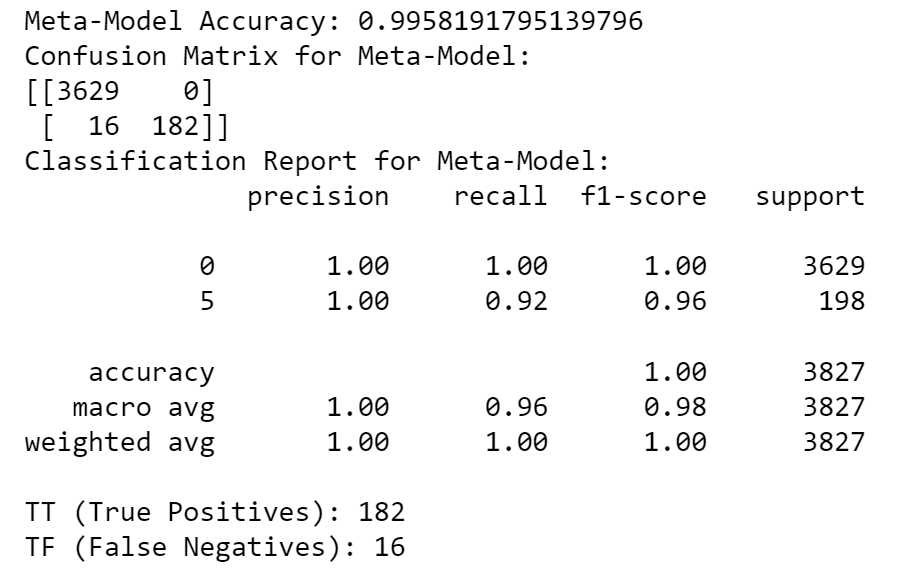
**– Section 6:** Report the accuracy scores, acctrain, the result of the best model selected in Section 5, and the number of features Nfeatures, used in the best k-fold cross validation model in Section 5. These values will be used for the ranking score.

* acctrain is the accuracy score of the train model selected above:  
  *Accuracy score of stacking method with GradientBoosting found by equation 1: 0.92*
* Nfeatures is the number of features for the best model in Section 4: *40*

Base model



Meta model:



**– Video Presentation:** Record the video presentation to describe the workflow, models, and results. You can either include the video file in the zip file or provide the link.

\*\* All reported results must be consistent with notebook files. Make sure to use random.seed() or random state= whenever needed so the same results can be reproduced. Each notebook file must display the result. Any non-consistent results will be marked 0. – Name the file as GroupNumber CourseSection Results