

LAPORAN PRAKTIKUM
IF3270 PEMBELAJARAN MESIN



Dosen Pengampu: Dr. Nur Ulfa Maulidevi, S.T, M.Sc.

Dibuat Oleh:

Bryan Bernigen / 13520034

Ng Kyle / 13520040

PROGRAM STUDI TEKNIK INFORMATIKA
SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA
INSTITUT TEKNOLOGI BANDUNG

I. Hasil Analisis Data

Menggunakan data train hasil Handout, didapat statistik data sebagai berikut:

1. Total Data Train : 169551
2. Duplicate Value : 3309
3. Missing Value and Outlier

Column (Attribute)	Missing Value Count	Outlier Count
hour	0	0
temp	0	838
temp_min	0	1061
temp_max	0	322
pressure	0	682
humidity	0	150
wind_speed	0	2219
wind_deg	0	0
raining	0	-

Terlihat bahwa tidak terdapat missing value pada dataset ini.

Untuk outlier terbanyak adalah untuk temp_min sedangkan hour dan wind_deg tidak terdapat outlier.

4. Target Data (Checking Imbalance Data)
 - False: 147238
 - True : 22313

Terlihat data imbalance dengan mayoritas data berkelas False.

II. Penanganan Hasil Analisis Data

Terhadap analisis yang didapat pada bagian I, dilakukan penanganan sebagai berikut:

1. Duplicate Value: Jumlah data duplikat sangat sedikit dibandingkan dengan jumlah data, maka data duplikat dihapus
2. Missing Value: Tidak ada missing value pada data, maka tidak ada penanganan yang dilakukan
3. Outlier: Jumlah outlier pada data sangat sedikit dibandingkan dengan jumlah data, maka data outlier akan dihapus, dengan hasil penghapusan menjadi:

Column (Attribute)	Outlier Count
hour	0
temp	0
temp_min	436
temp_max	2
pressure	152
humidity	0
wind_speed	407
wind_deg	0

4. Imbalance Data: Karena data rain True sangat sedikit, dilakukan Oversampling dengan teknik SMOTE (Synthetic Minority Oversampling Technique) untuk menghasilkan data baru terhadap data rain True.

Hasil akhir

- True 147238
- False 147238

III. Justifikasi Teknik yang Dipilih

1. Duplicate Value: kami memutuskan untuk menghapus duplicate value pada dataset ini karena kami tidak memiliki data pendukung yang menyatakan bahwa data duplicate tersebut merupakan 2 kejadian berbeda sehingga kami mengasumsikan bahwa data duplicate merupakan data yang salah dimasukkan (terinput 2 kali) sehingga kami memutuskan untuk menghapus data tersebut.
2. Missing Value: Tidak ada data missing sehingga tidak perlu dilakukan apa-apa
3. Outlier: Kami memutuskan untuk menghapus outlier karena data outlier biasanya merupakan data yang salah dimasukkan atau memang kasus khusus sehingga kami memutuskan untuk menghapusnya agar data tersebut tidak merusak model general kami
4. Imbalance Data: kami memutuskan untuk melakukan oversampling dengan SMOTE karena berdasarkan riset kami, semakin banyak data, akan semakin baik model yang dihasilkan namun akan semakin lama juga waktu yang dibutuhkan untuk melakukan training data

IV. Perubahan Desain Eksperimen

1. Terdapat perubahan pada level dan ranges pada hyperparameter pengujian SVM yakni mengubah nilai C dari [0.1, 1, 10, 100, 1000]

menjadi [0.1, 1, 10] dan mengubah nilai gamma dari [1, 0.1, 0.01, 0.001, 0.0001] menjadi [0.1, 0.01, 0.001] karena berdasarkan hasil pengujian kami, SVM pada data eksperimen kami membutuhkan waktu yang cukup lama, yakni sekitar 3 menit untuk 1x SVM (C=0.1, gamma=0.1) sehingga kami memutuskan untuk mengurangi jumlah parameter yang diuji agar jumlah pengujian berkurang.

Untuk menentukan nilai mana yang akan dihapus, kami mencoba menggunakan data dengan teknik undersampling dan tanpa mengubah jam menggunakan one hot encoding. Dari percobaan tersebut, diperoleh waktu untuk 1 x SVM dengan C=0.1 dan gamma=1 adalah 5 detik dengan score sekitar 0.5. Sedangkan untuk C=1000 dan gamma=0.01, diperoleh waktu selama 12 menit dengan score 0.77

Berdasarkan hasil percobaan tersebut, kami memutuskan untuk menghapus nilai gamma 1 dan 0.0001 karena berdasarkan hasil pra praktikum, gamma tersebut cenderung memberikan hasil yang kurang baik. Kami juga menghapus nilai C=100 dan 1000 karena berdasarkan hasil pra praktikum, nilai C tersebut akan memakan waktu yang jauh lebih lama dibandingkan dengan ketika C=0.1. Bahkan ketika nilai C=100, waktu SVM dapat membengkak hingga 100x lipat sehingga kami memutuskan untuk menghapus nilai tersebut karena keterbatasan waktu

2. Terdapat perubahan pada skema validasi dari k-fold cross validation menjadi hold out karena masalah keterbatasan waktu dimana k-fold memakan waktu yang sangat lama sehingga kami menggantinya dengan holdout

V. Desain Eksperimen

1. Tujuan Eksperimen
 - a. Problem Statement
Melakukan analisis kebenaran kondisi hujan pada kota Denpasar.
 - b. Eksperimen
Mencari model eksperimen terbaik beserta hyper parameternya untuk melakukan prediksi keterjadian hujan dari data dari
2. Variabel Dependen
Kolom target: "raining"
3. Variabel Independen
 - Feature Selection: Kolom atribut pada data "weather_main" terkecuali kolom "raining" sebagai target, yaitu "hour", "temp", "temp_min", "temp_max", "pressure", "humidity", "wind_speed", "wind_deg".
 - Mode Tune: Hyperparameter model
 - Model Comparison: Model (Logreg, SVM, serta gabungannya)
4. Strategi Eksperimen
 - a. Metric Penilaian model terhadap response variable:
 - Accuracy
 - Precision
 - Recall

- F1-Score
- AUROC

b. Model yang akan diuji

- Logistic Regression
- Support Vector Machine (SVM) dengan kernel 'rbf'

Alasan pemilihan model: Logistic Regression sebagai baseline, SVM mampu melakukan klasifikasi terhadap non linearly separable dataset dengan meningkatkan dimensi memungkinkan model yang baik.

c. Factor hyperparameter pengujian (SVM)

- C (Regularization parameter)
- Gamma (Kernel Coefficient)

d. Levels and Ranges

Ranges

- C: [0.1, 1, 10, ~~100, 1000~~]
- Gamma: [~~1~~, 0.1, 0.01, 0.001, ~~0.0001~~]

Levels

Hasil terbaik dari Grid Search hyperparameter SVM digabung dengan Logistic Regression

- Soft Voting
- Hard Voting
- Stacking

e. Strategi Eksperimen

Grid Search

5. Skema Validasi

~~K-Fold Cross Validation dengan nilai k=5~~

Holdout Validation, seperti halnya dari data yang diberikan:

64% Training Set, 16% Test Set, dan 20% Validation Set

VI. Hasil Eksperimen

1. Model Logistic Regression (Baseline)

Accuracy: 0.7106964235160894

AUROC: 0.7288166514921892

	Precision	Recall	F1-Score	Support
False	0.95	0.70	0.81	36791
True	0.28	0.73	0.41	5597
Accuracy			0.71	43288

Macro Avg	0.61	0.73	0.61	42388
Weighted Avg	0.86	0.71	0.76	42388

2. Model SVM (**C=10, gamma=0.1**)

Accuracy: 0.8308955364725866

AUROC: 0.7006330118609502

	Precision	Recall	F1-Score	Support
False	0.92	0.88	0.90	45994
True	0.39	0.52	0.45	6991
Accuracy			0.83	52985
Macro Avg	0.66	0.70	0.67	52985
Weighted Avg	0.85	0.83	0.84	52985

3. Model Soft Voting*

Accuracy: 0.711923185807304

AUROC:0.7404303800149755

	Precision	Recall	F1-Score	Support
False	0.95	0.70	0.81	36791
True	0.28	0.78	0.42	5597
Accuracy			0.71	42388
Macro Avg	0.61	0.74	0.61	42388
Weighted Avg	0.87	0.71	0.76	42388

4. Model Hard Voting*

Accuracy: 0.7308672265735585

AUROC: 0.7333922149969998

	Precision	Recall	F1-Score	Support
--	-----------	--------	----------	---------

False	0.95	0.73	0.82	36791
True	0.29	0.74	0.42	5597
Accuracy			0.73	42388
Macro Avg	0.62	0.73	0.62	42388
Weighted Avg	0.86	0.73	0.77	42388

5. Stacking*

Accuracy: 0.7308672265735585

AUROC: 0.7333922149969998

	Precision	Recall	F1-Score	Support
False	0.95	0.73	0.82	36791
True	0.29	0.74	0.42	5597
Accuracy			0.73	42388
Macro Avg	0.62	0.73	0.62	42388
Weighted Avg	0.86	0.73	0.77	42388

*Perhitungan menggunakan data undersampling pada notebook lain (terlampir)

VII. Analisis Hasil Eksperimen

Berdasarkan hasil eksperimen, model logistic regression belum berhasil menghasilkan model yang memuaskan. Model logistic regression hanya memiliki akurasi sebesar 71.1% dan nilai AUROC sebesar 72.2%. Berdasarkan kedua nilai tersebut, model ini tergolong kategori buruk. Berdasarkan analisis kami, model tersebut belum berhasil melakukan klasifikasi dengan baik karena karena data yang digunakan tidak dapat dipisahkan secara linear sehingga tidak cocok dengan model logistic regression yang pada dasarnya merupakan model linear.

Analisis tersebut didukung oleh hasil SVM yang memperoleh hasil yang lebih baik pada angka akurasi sebesar 83.1% dan nilai AUROC 70,1%.

Berdasarkan nilai tersebut, model SVM tersebut dapat dikategorikan cukup. Sebenarnya hasil tersebut masih dibawah ekspektasi kami, namun karena keterbatasan waktu, kami hanya dapat menerima dan melakukan analisis apa saja yang dapat di improve.

Berdasarkan analisis kami, ada beberapa hal yang dapat dilakukan untuk mengimprove hasil model yakni dengan melakukan grid search yang lebih banyak seperti menambah parameter C dan gamma yang dicek serta menambah jenis kernel seperti linear, polynomial, dan sigmoid ke grid search. Kami rasa kernel rbf belum berhasil memisahkan data dengan baik sehingga ada baiknya untuk dicoba menggunakan kernel lain. Kami belum sempat menambah parameter-parameter tersebut karena keterbatasan waktu. Salah satu kelemahan SVM adalah waktu train yang lama sehingga untuk melakukan grid search dengan parameter yang banyak dengan berbagai macam kernel akan menghabiskan banyak waktu. Sebenarnya hal tersebut dapat diakali dengan menjalankan grid search secara paralel di beberapa perangkat. Namun karena keterbatasan resource, kami tidak dapat melakukan hal tersebut. Jika memungkinkan, kami juga merekomendasikan untuk mengganti library SVM yang digunakan menjadi ThunderSVM karena library tersebut dapat melakukan komputasi di GPU sehingga perhitungan akan jauh lebih cepat. Kami tidak melakukan hal tersebut karena resource GPU yang tidak kami miliki.

Sebenarnya untuk Softvoting, Hardvoting, dan Stacking, kami tidak berhasil mendapatkan hasil pada data yang di oversampling karena 1 perhitungan memakan waktu lebih dari 1 jam sehingga tidak keburu untuk mendapatkan hasil. Oleh karena itu, kami melakukan sedikit modifikasi pada data yakni kami ganti dengan undersampling pada notebook lain. Berdasarkan hasil tersebut, diperoleh model yang tetap tergolong kurang karena model menggunakan data yang tidak terlalu banyak sehingga hasil tersebut kalah dengan model SVM saja namun menggunakan data yang banyak.

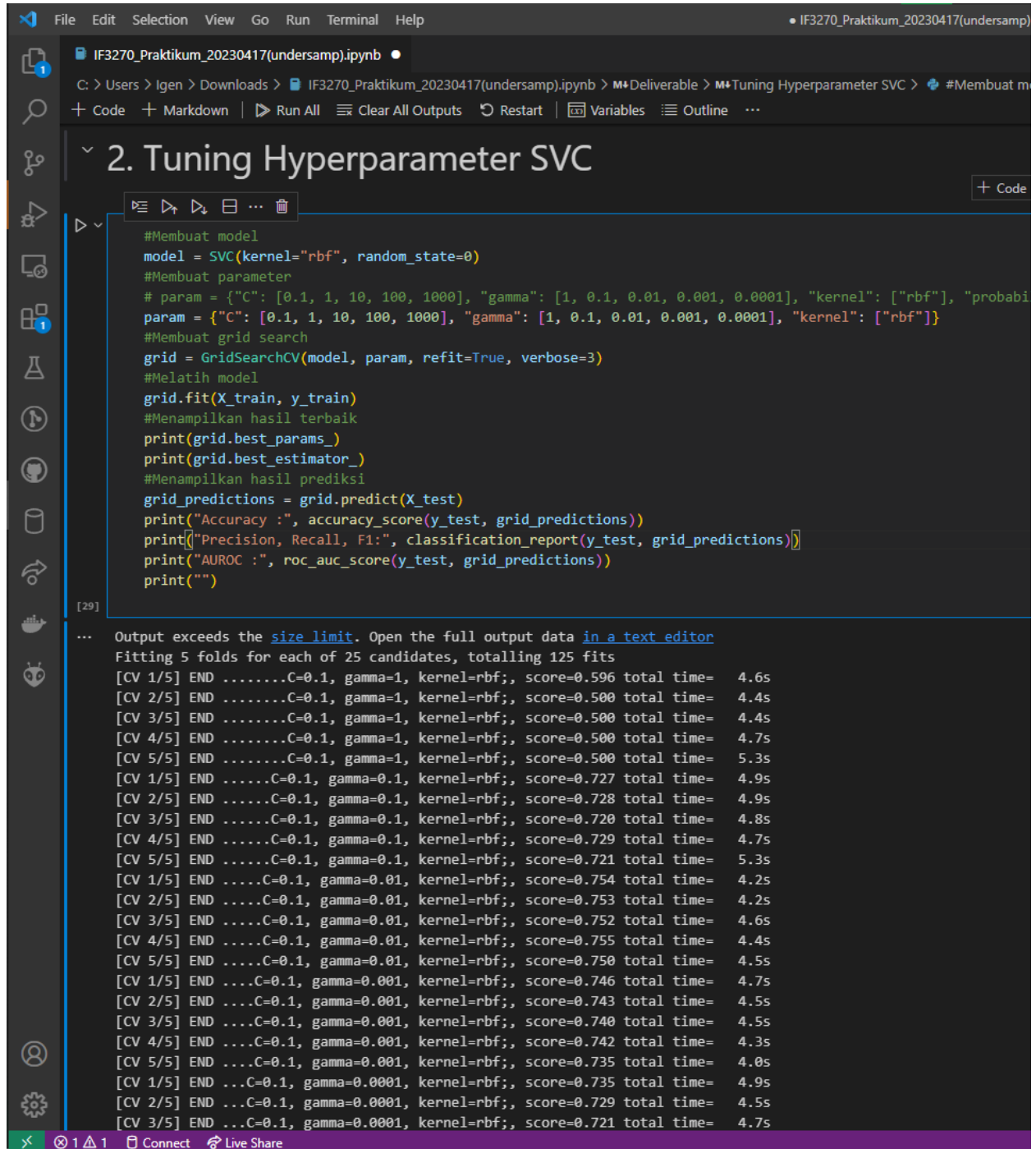
VIII. Kesimpulan

Semakin banyak data, maka akan semakin baik model yang dibuat, namun waktu yang digunakan untuk melakukan training akan semakin banyak sehingga akan semakin sedikit kesempatan untuk hyperparameter tuning. Oleh karena itu dibutuhkan balance antara jumlah data dan hyperparameter tuning.

IX. Pembagian Tugas

13520034 - Bryan Bernigen	Semua bersama
13520040 - Ng Kyle	Semua bersama

Lampiran



```
File Edit Selection View Go Run Terminal Help • IF3270_Praktikum_20230417(undersamp)

IF3270_Praktikum_20230417(undersamp).ipynb •
C: > Users > Igen > Downloads > IF3270_Praktikum_20230417(undersamp).ipynb > M+ Deliverable > M+ Tuning Hyperparameter SVC > # Membuat m
+ Code + Markdown ▶ Run All ≡ Clear All Outputs ↺ Restart | Variables ≡ Outline ...

2. Tuning Hyperparameter SVC + Code

#Membuat model
model = SVC(kernel="rbf", random_state=0)
#Membuat parameter
# param = {"C": [0.1, 1, 10, 100, 1000], "gamma": [1, 0.1, 0.01, 0.001, 0.0001], "kernel": ["rbf"], "probabi
param = {"C": [0.1, 1, 10, 100, 1000], "gamma": [1, 0.1, 0.01, 0.001, 0.0001], "kernel": ["rbf"]}
#Membuat grid search
grid = GridSearchCV(model, param, refit=True, verbose=3)
#Melatih model
grid.fit(X_train, y_train)
#Menampilkan hasil terbaik
print(grid.best_params_)
print(grid.best_estimator_)
#Menampilkan hasil prediksi
grid_predictions = grid.predict(X_test)
print("Accuracy :", accuracy_score(y_test, grid_predictions))
print("Precision, Recall, F1:", classification_report(y_test, grid_predictions))
print("AUROC :", roc_auc_score(y_test, grid_predictions))
print("")

[29]

... Output exceeds the size limit. Open the full output data in a text editor
Fitting 5 folds for each of 25 candidates, totalling 125 fits
[CV 1/5] END .....C=0.1, gamma=1, kernel=rbf;; score=0.596 total time= 4.6s
[CV 2/5] END .....C=0.1, gamma=1, kernel=rbf;; score=0.500 total time= 4.4s
[CV 3/5] END .....C=0.1, gamma=1, kernel=rbf;; score=0.500 total time= 4.4s
[CV 4/5] END .....C=0.1, gamma=1, kernel=rbf;; score=0.500 total time= 4.7s
[CV 5/5] END .....C=0.1, gamma=1, kernel=rbf;; score=0.500 total time= 5.3s
[CV 1/5] END .....C=0.1, gamma=0.1, kernel=rbf;; score=0.727 total time= 4.9s
[CV 2/5] END .....C=0.1, gamma=0.1, kernel=rbf;; score=0.728 total time= 4.9s
[CV 3/5] END .....C=0.1, gamma=0.1, kernel=rbf;; score=0.720 total time= 4.8s
[CV 4/5] END .....C=0.1, gamma=0.1, kernel=rbf;; score=0.729 total time= 4.7s
[CV 5/5] END .....C=0.1, gamma=0.1, kernel=rbf;; score=0.721 total time= 5.3s
[CV 1/5] END .....C=0.1, gamma=0.01, kernel=rbf;; score=0.754 total time= 4.2s
[CV 2/5] END .....C=0.1, gamma=0.01, kernel=rbf;; score=0.753 total time= 4.2s
[CV 3/5] END .....C=0.1, gamma=0.01, kernel=rbf;; score=0.752 total time= 4.6s
[CV 4/5] END .....C=0.1, gamma=0.01, kernel=rbf;; score=0.755 total time= 4.4s
[CV 5/5] END .....C=0.1, gamma=0.01, kernel=rbf;; score=0.750 total time= 4.5s
[CV 1/5] END .....C=0.1, gamma=0.001, kernel=rbf;; score=0.746 total time= 4.7s
[CV 2/5] END .....C=0.1, gamma=0.001, kernel=rbf;; score=0.743 total time= 4.5s
[CV 3/5] END .....C=0.1, gamma=0.001, kernel=rbf;; score=0.740 total time= 4.5s
[CV 4/5] END .....C=0.1, gamma=0.001, kernel=rbf;; score=0.742 total time= 4.3s
[CV 5/5] END .....C=0.1, gamma=0.001, kernel=rbf;; score=0.735 total time= 4.0s
[CV 1/5] END .....C=0.1, gamma=0.0001, kernel=rbf;; score=0.735 total time= 4.9s
[CV 2/5] END .....C=0.1, gamma=0.0001, kernel=rbf;; score=0.729 total time= 4.5s
[CV 3/5] END .....C=0.1, gamma=0.0001, kernel=rbf;; score=0.721 total time= 4.7s
```

FileEditSelectionViewGoRunTerminalHelp

IF3270_Praktikum_20230417(undersamp) •

IF3270_Praktikum_20230417(undersamp).ipynb •

C: > Users > Igen > Downloads > IF3270_Praktikum_20230417(undersamp).ipynb > M+Deliverable > M+Improvement (Soft V

+ Code + Markdown | ▶ Run All ≡ Clear All Outputs ↺ Restart | [x] Variables ≡ Outline ...

Soft Voting

```
softVoting = VotingClassifier(estimators=models, voting='soft')
model_fit(softVoting)
score(softVoting)
```

[25] ✓ 34.3s

... Accuracy: 0.711923185807304
Precision, Recall, F1:

	precision	recall	f1-score	support
False	0.95	0.70	0.81	36791
True	0.28	0.78	0.42	5597
accuracy			0.71	42388
macro avg	0.62	0.74	0.61	42388
weighted avg	0.87	0.71	0.76	42388

AUROC : 0.7404303800149755
Confusion Matrix:

</>

Confusion Matrix

True labels

Predicted labels

25816 10975 1236 4361

0 1 0 1

5000 10000 15000 20000 25000

⌕ 1 ⚠ 1

Connect Live Share

FileEditSelectionViewGoRunTerminalHelp

IF3270_Praktikum_20230417(undersamp).ipynb

C: > Users > Igen > Downloads > IF3270_Praktikum_20230417(undersamp).ipynb > M+Deliverable > M+Improvement (Soft V

+ Code + Markdown | ▶ Run All ≡ Clear All Outputs ↺ Restart | [G3] Variables ≡ Outline ...

Hard Voting

```
hardVoting = VotingClassifier(estimators=models, voting='hard')
model_fit(hardVoting)
score(hardVoting)
```

[23] ✓ 33.0s

... Accuracy: 0.7308672265735585
Precision, Recall, F1:

	precision	recall	f1-score	support
False	0.95	0.73	0.82	36791
True	0.29	0.74	0.42	5597
accuracy			0.73	42388
macro avg	0.62	0.73	0.62	42388
weighted avg	0.86	0.73	0.77	42388

AUROC : 0.7333922149969998
Confusion Matrix:

</>

Confusion Matrix

	0	1
0	26856	9935
1	1473	4124

< 1 1

Connect Live Share

FileEditSelectionViewGoRunTerminalHelp

IF3270_Praktikum_20230417(undersamp)

IF3270_Praktikum_20230417(undersamp).ipynb

C: > Users > Igen > Downloads > IF3270_Praktikum_20230417(undersamp).ipynb > M+Deliverable > M+Improvement (Soft V

+ Code + Markdown | ▶ Run All ≡ Clear All Outputs ↺ Restart | Variables Outline ...

Stacking

```
stacking = StackingClassifier(estimators=models, final_estimator=LogisticRegression())
model_fit(hardVoting)
score(hardVoting)
```

[24] ✓ 34.4s

... Accuracy: 0.7308672265735585
Precision, Recall, F1:

	precision	recall	f1-score	support
False	0.95	0.73	0.82	36791
True	0.29	0.74	0.42	5597
accuracy			0.73	42388
macro avg	0.62	0.73	0.62	42388
weighted avg	0.86	0.73	0.77	42388

AUROC : 0.7333922149969998
Confusion Matrix:

Confusion Matrix

	0	1
0	26856	9935
1	1473	4124

1 1 1 Connect Live Share