People Gender Image Classification and Age Estimation using Ensemble Learning

Big Data Challenge 2021

OUTLINE

01 Problem Definition

02 Analytics Process & Algorithm

03 Result

04 Insight & Recommendation

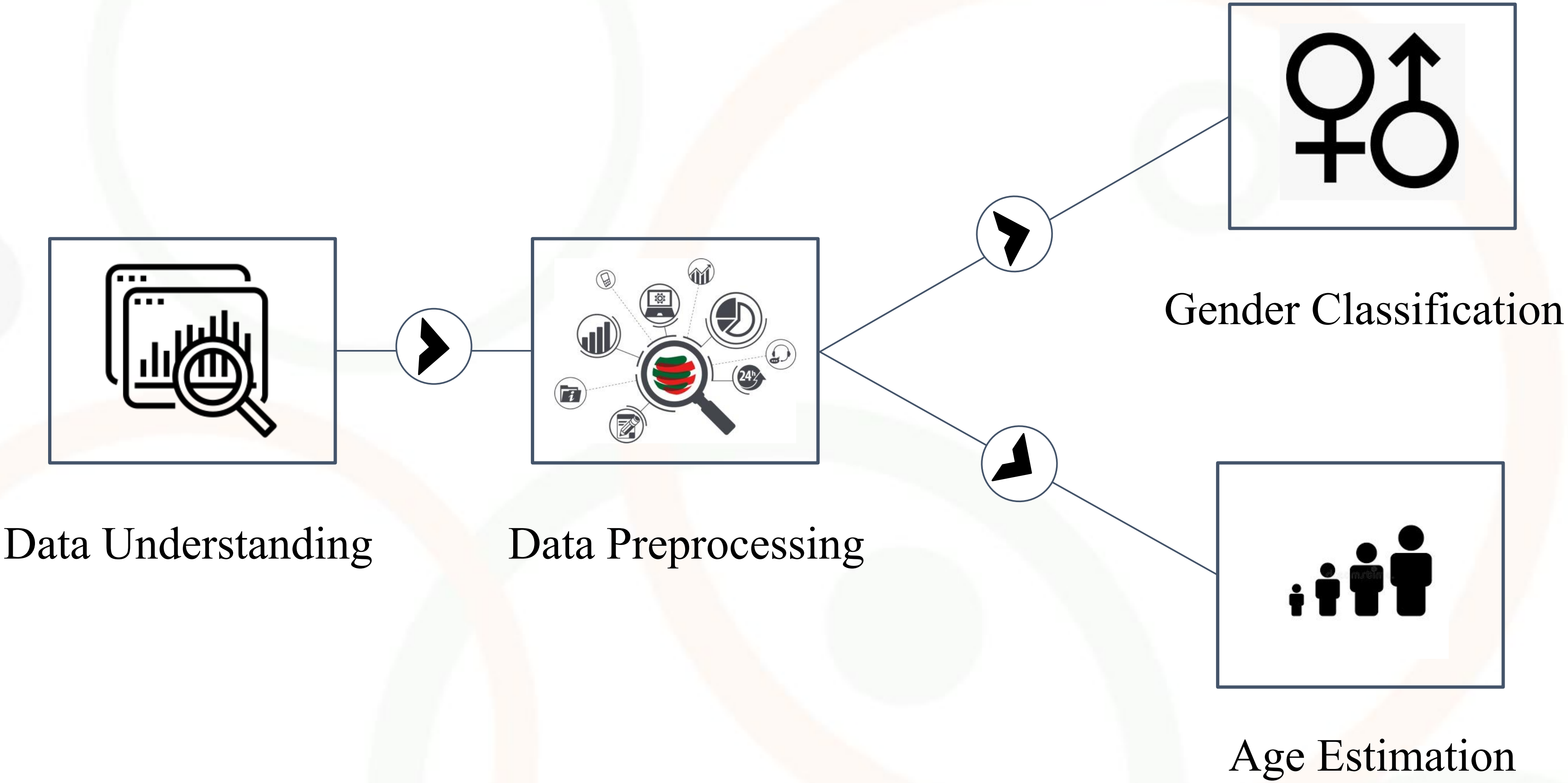
05 Conclusion

PROBLEM DEFINITION

Gender Classification

Age Estimation

ANALYTICS PROCESS & ALGORITHM



Data Understanding

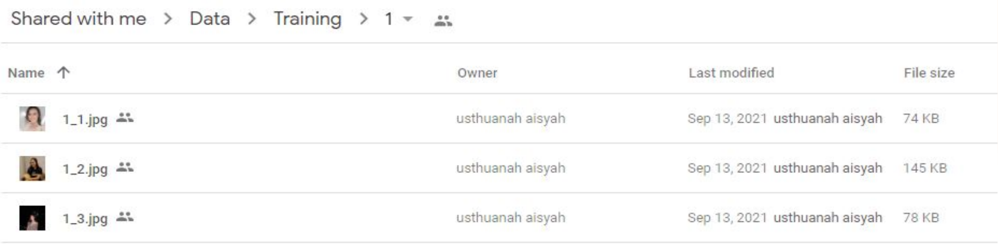
Data Preprocessing

Gender Classification

Age Estimation

ANALYTICS PROCESS & ALGORITHM

Data Understanding

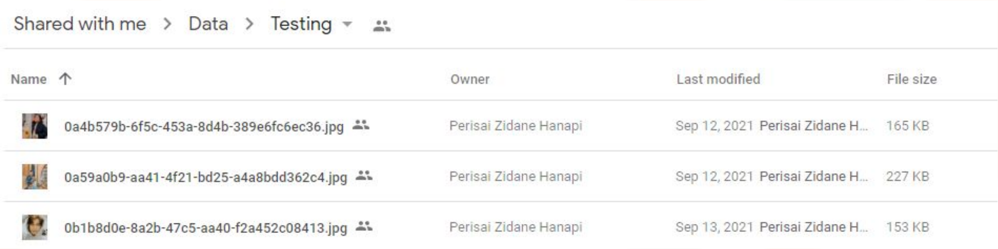


Shared with me > Data > Training > 1

Name	Owner	Last modified	File size
1_1.jpg	usthuanah aisyah	Sep 13, 2021 usthuanah aisyah	74 KB
1_2.jpg	usthuanah aisyah	Sep 13, 2021 usthuanah aisyah	145 KB
1_3.jpg	usthuanah aisyah	Sep 13, 2021 usthuanah aisyah	78 KB

Training Image

770 Folder, each folder consists of 3 image files (Foldername_n.jpg), each folder contains information about one person



Shared with me > Data > Testing

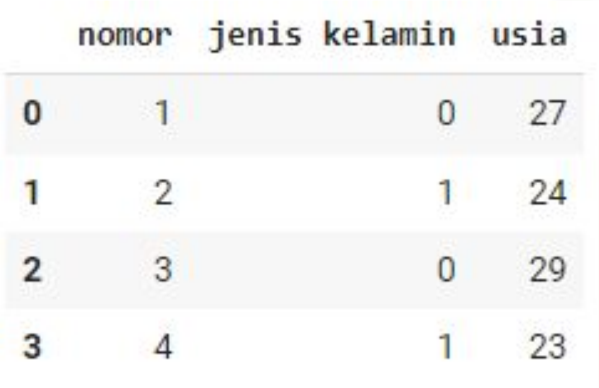
Name	Owner	Last modified	File size
0a4b579b-6f5c-453a-8d4b-389e6fc6ec36.jpg	Perisai Zidane Hanapi	Sep 12, 2021 Perisai Zidane H...	165 KB
0a59a0b9-aa41-4f21-bd25-a4a8bdd362c4.jpg	Perisai Zidane Hanapi	Sep 12, 2021 Perisai Zidane H...	227 KB
0b1b8d0e-8a2b-47c5-aa40-f2a452c08413.jpg	Perisai Zidane Hanapi	Sep 13, 2021 Perisai Zidane H...	153 KB

Test Image

990 image files that gender and age will be predict

ANALYTICS PROCESS & ALGORITHM

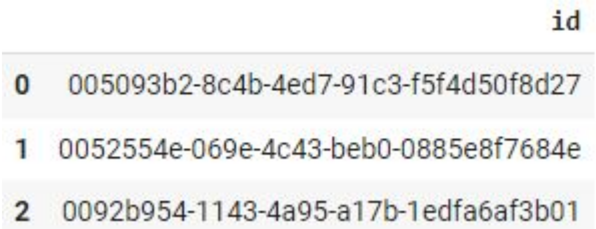
Data Understanding



	nomor	jenis kelamin	usia
0	1	0	27
1	2	1	24
2	3	0	29
3	4	1	23

Train.csv

770 rows of data containing information about gender and age



	id
0	005093b2-8c4b-4ed7-91c3-f5f4d50f8d27
1	0052554e-069e-4c43-beb0-0885e8f7684e
2	0092b954-1143-4a95-a17b-1edfa6af3b01

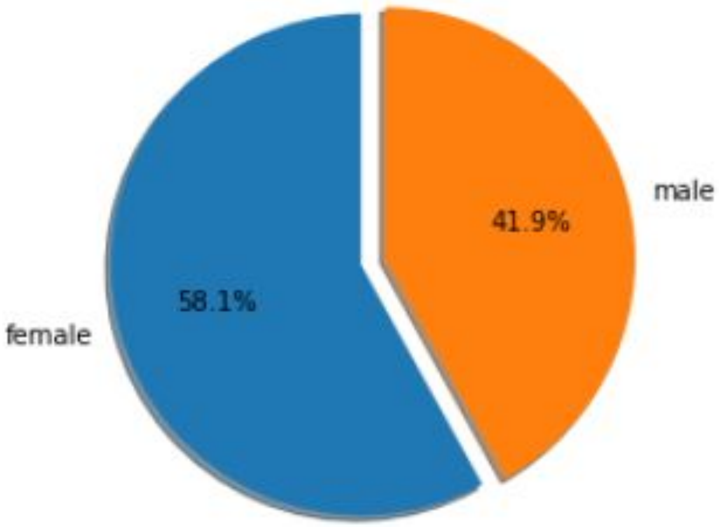
Submission.csv

990 rows of data contain file names in the testing folder, we will add gender and or age columns to make predictions

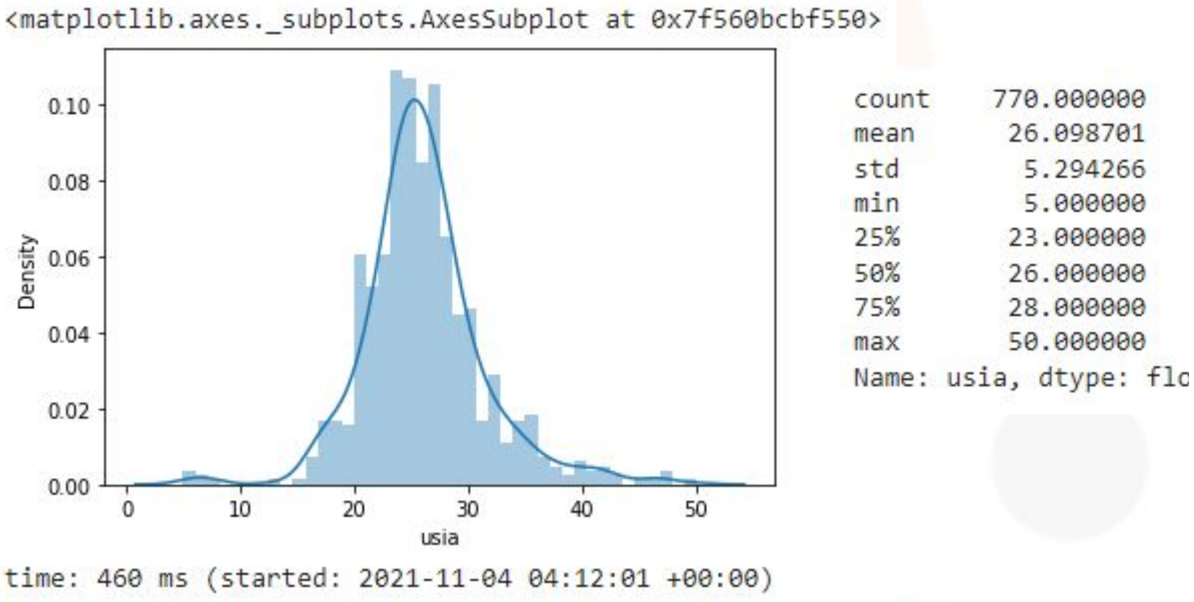
ANALYTICS PROCESS & ALGORITHM

Data Understanding

Gender Distribution



Age Distribution



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f560bcbf550>
```

```
count    770.000000
mean      26.098701
std        5.294266
min        5.000000
25%       23.000000
50%       26.000000
75%       28.000000
max       50.000000
Name: usia, dtype: flo
```

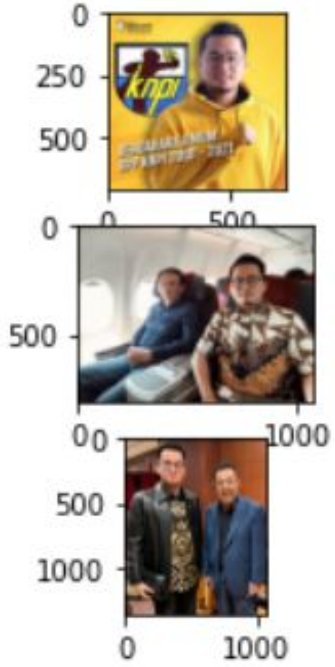
```
time: 460 ms (started: 2021-11-04 04:12:01 +00:00)
```

ANALYTICS PROCESS & ALGORITHM

Data Preprocessing

Training Image (1)

- Face Cropping (MTCNN Face Detector)



Face 1

Face 2

Face 3

Face 4

Face 5

ANALYTICS PROCESS & ALGORITHM

Data Preprocessing

Training Image (2)

- Face Similarity (DeepFace with threshold < 0.22)

Face 1

Face 2

Face 3

Face 4

Face 5

Distance	Face 1	Face 2	Face 3	Face 4	Face 5
Face 1			< 0.22	< 0.22	
Face 2					
Face 3	< 0.22			< 0.22	
Face 4	< 0.22		< 0.22		
Face 5					

Face 1

Face 3

Face 4

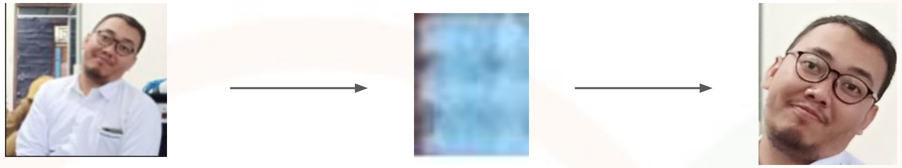
ANALYTICS PROCESS & ALGORITHM

Data Preprocessing

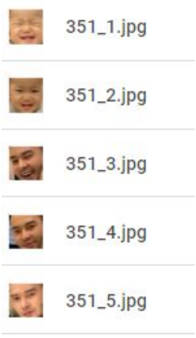
Training Image (3)

Manual Checking (1)

- Wrong when cropping face → Manual Cropping



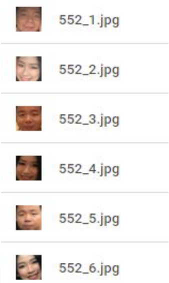
- There are 2 or more different faces extracted → Delete Irrelevant Face



351_1.jpg
351_2.jpg
351_3.jpg
351_4.jpg
351_5.jpg

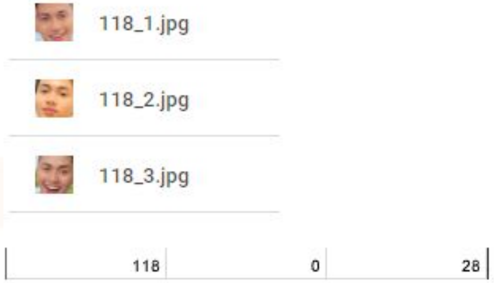
Manual Checking (2)

- Confusing Folder (ex : 552) → Delete Irrelevant Face



552_1.jpg
552_2.jpg
552_3.jpg
552_4.jpg
552_5.jpg
552_6.jpg

- Labeling Error by the Commitee → Relabeling



118_1.jpg
118_2.jpg
118_3.jpg

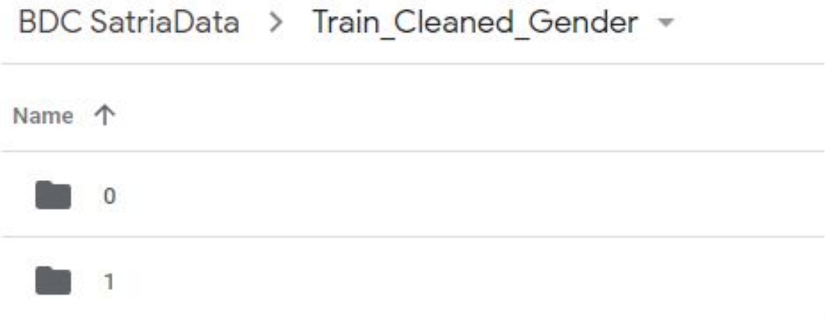
118 0 28

ANALYTICS PROCESS & ALGORITHM

Data Preprocessing

Training Image (Gender Classification)

- Splitting face image with mapping between filenames and gender column in train.csv (0 for female and 1 for male)



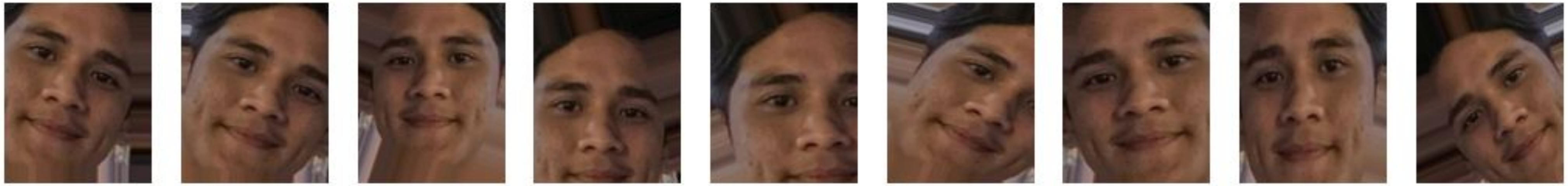
BDC SatriaData > Train_Cleaned_Gender

Name ↑

0

1

- Face Augmentation (Image Data Generator) → Rotate, flip, shear, zoom, shift

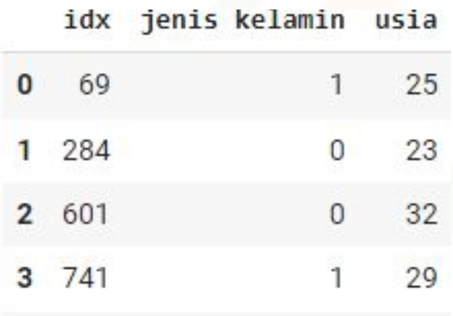


ANALYTICS PROCESS & ALGORITHM

Data Preprocessing

Training Image (Age Estimation)

- Mapping between filename and age column in train.csv then make filenames are indexes in mapping_age.csv



	idx	jenis kelamin	usia
0	69	1	25
1	284	0	23
2	601	0	32
3	741	1	29

- Convert age column to months (*add variety so it's not too discrete*)

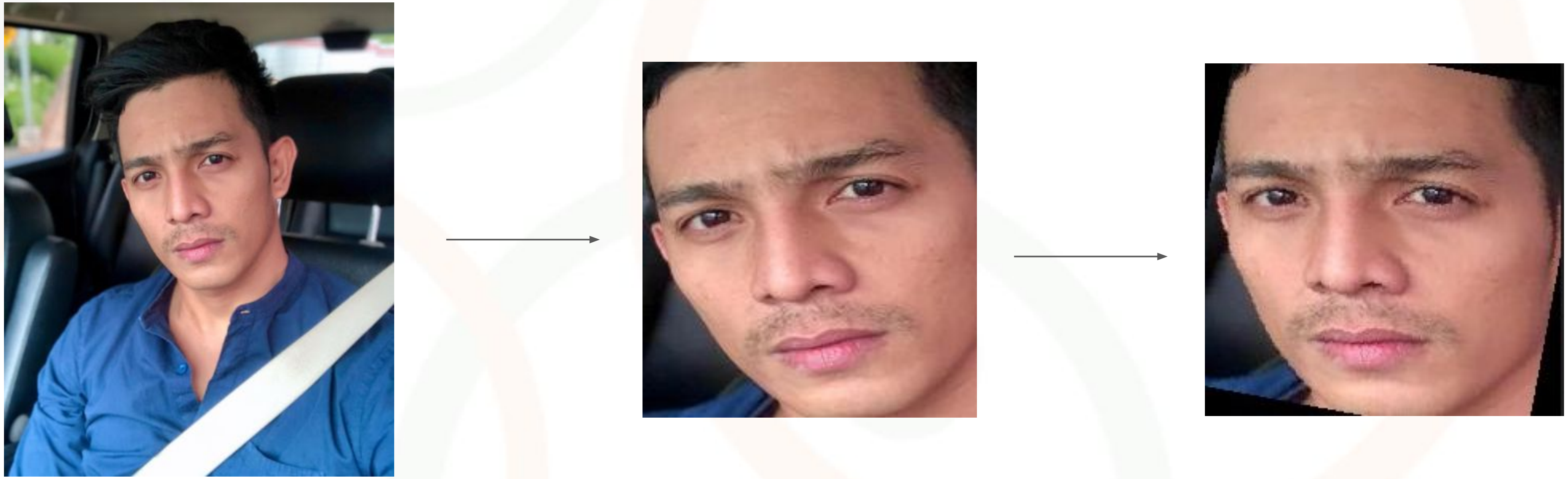
  Example age in year is 28. If we want convert it to months, it will be random value between 330 (27 years 6 months) to 341 (28 years 5 months)

ANALYTICS PROCESS & ALGORITHM

Data Preprocessing

Testing Image

- Face Cropping (MTCNN Face Detector) & Face Alignment



ANALYTICS PROCESS & ALGORITHM

Gender Classification

1 Deep Learning Modelling for Gender

2 Feature Extraction with **model_VGG_200_tl**

3 Handling Imbalance Training Data (SMOTETomek)

4 Support Vector Machine Classifier

5 Random Forest Classifier

6 Result Each Model

7 Ensemble Learning for Gender (Soft Voting)

8 Handling Two or More Face

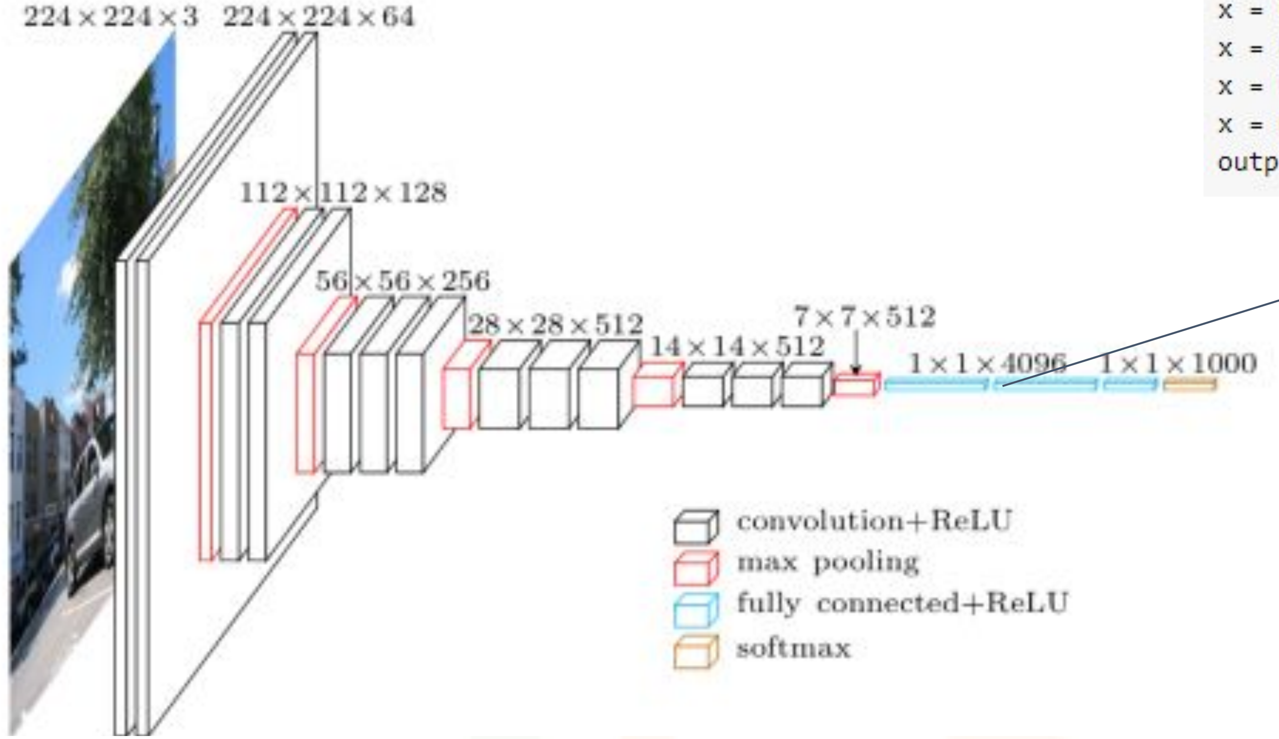
9 Submission History

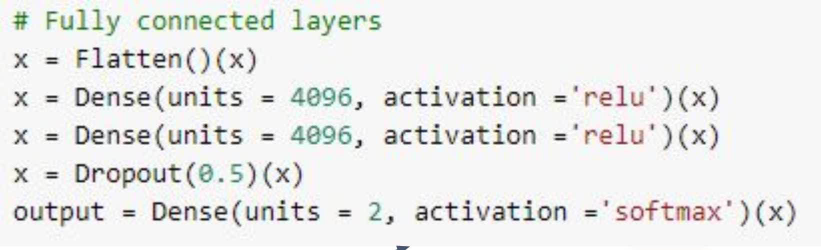
Puspresnas
Pusat Prestasi Nasional

ANALYTICS PROCESS & ALGORITHM

Gender Classification

1 Deep Learning Modelling for Gender





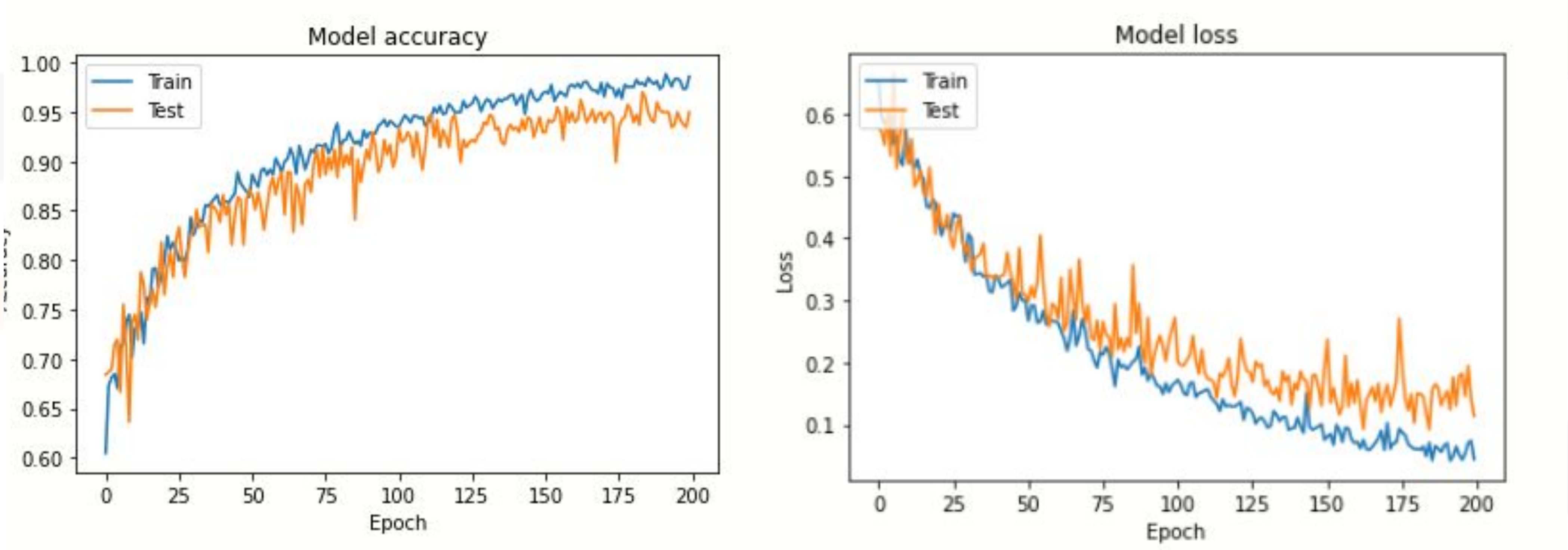
```
# Fully connected layers
x = Flatten()(x)
x = Dense(units = 4096, activation ='relu')(x)
x = Dense(units = 4096, activation ='relu')(x)
x = Dropout(0.5)(x)
output = Dense(units = 2, activation ='softmax')(x)
```

VGG 16 + SGD Optimizer (learning_rate = 0.001 [First 200 Epochs] and 0.0001 [After First 200 Epochs])

ANALYTICS PROCESS & ALGORITHM

Gender Classification

1 Deep Learning Modelling for Gender



VGG 16 Training Result (1)

ANALYTICS PROCESS & ALGORITHM

Gender Classification

1 Deep Learning Modelling for Gender

VGG 16 Training Result (2)

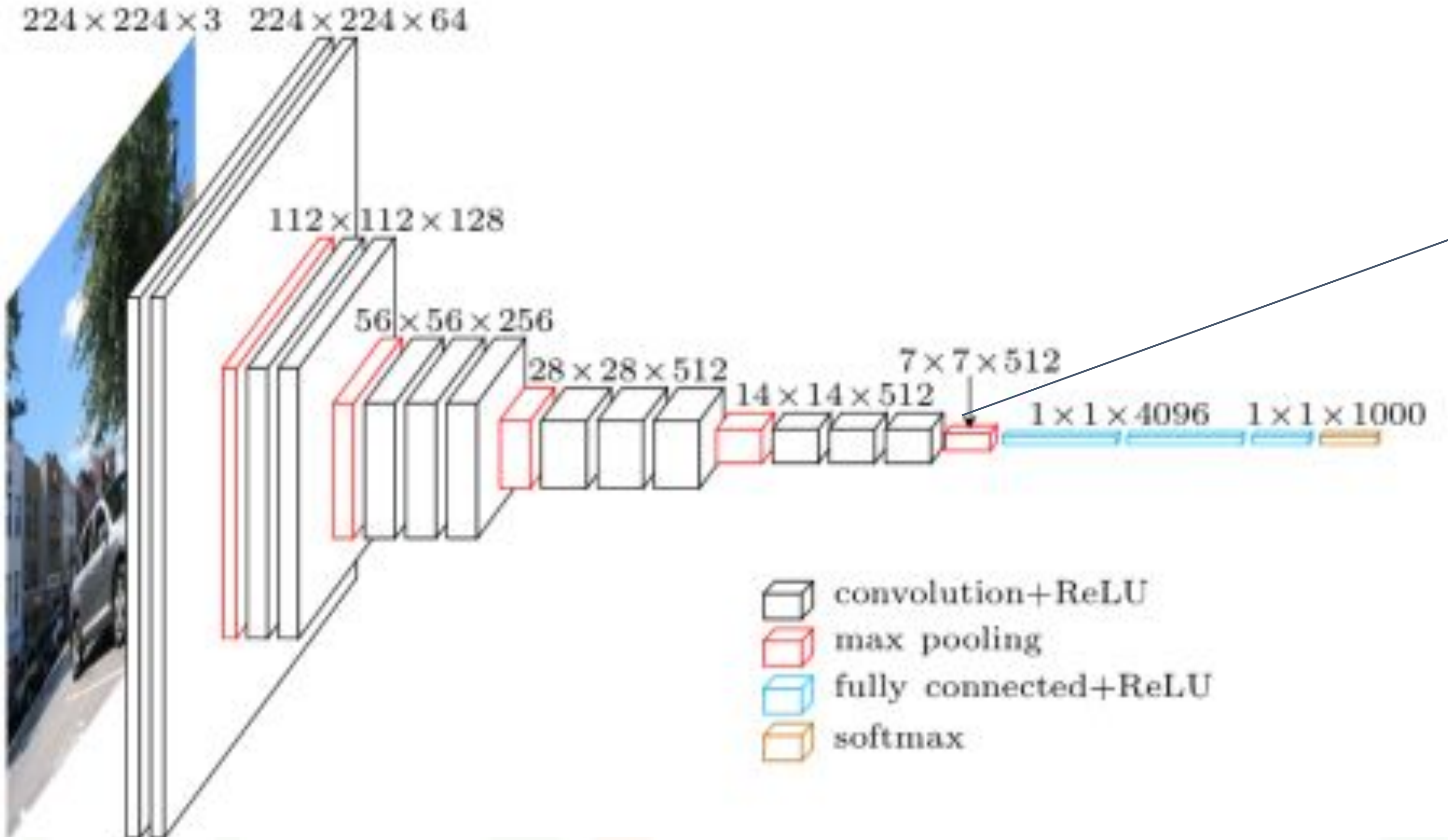
```
loss: 0.0292

accuracy: 0.9893

val_loss: 0.0839

val_accuracy: 0.9747
```

This model we save and give it name **model_VGG_200_tl**

ANALYTICS PROCESS & ALGORITHM

Gender Classification

2 Feature Extraction with **model_VGG_200_tl**



224 x 224 x 3 to 7 x 7 x 512

Each image have 25088 feature after extraction for ML modelling

ANALYTICS PROCESS & ALGORITHM

Gender Classification

3 Handling Imbalance Training Data (SMOTETomek)

Oversampling for handling imbalance training data

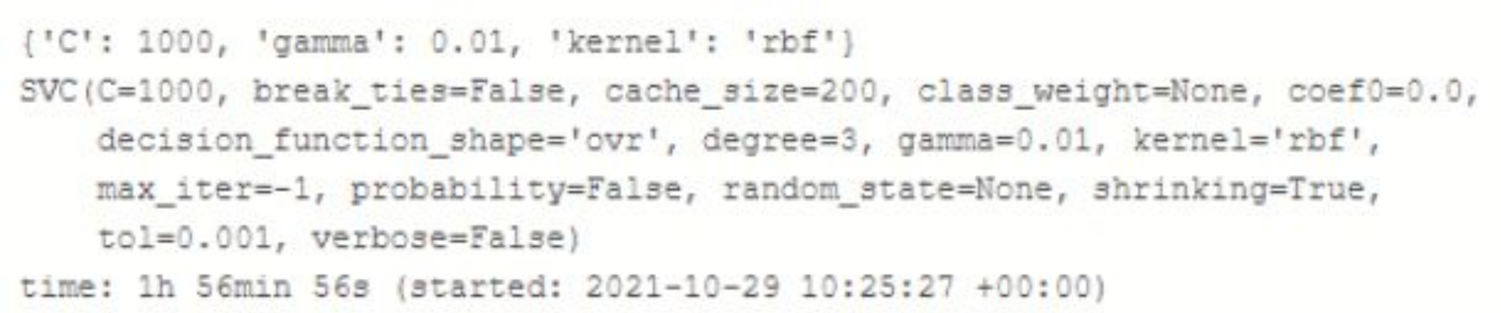
0	1187	→	1180
1	810		1180

ANALYTICS PROCESS & ALGORITHM

Gender Classification

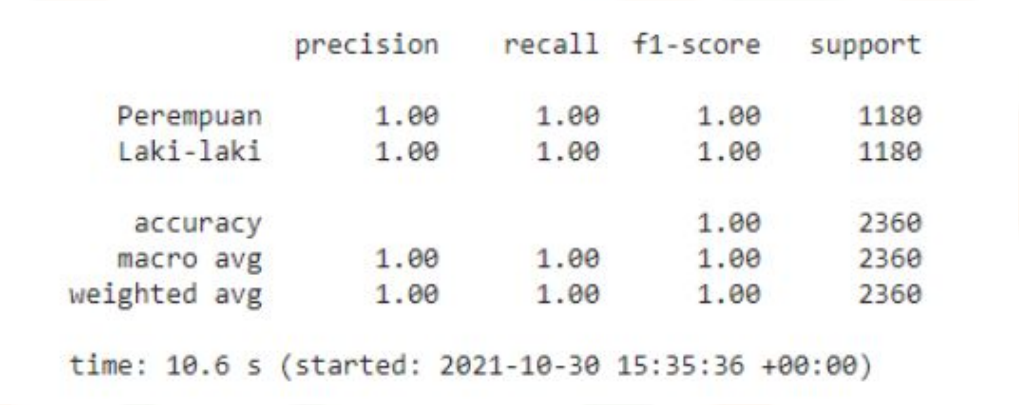
4 Support Vector Machine Classifier

Hyperparameter Tuning



```
{'C': 1000, 'gamma': 0.01, 'kernel': 'rbf'}
SVC(C=1000, break_ties=False, cache_size=200, class_weight=None, coef0=0.0,
    decision_function_shape='ovr', degree=3, gamma=0.01, kernel='rbf',
    max_iter=-1, probability=False, random_state=None, shrinking=True,
    tol=0.001, verbose=False)
time: 1h 56min 56s (started: 2021-10-29 10:25:27 +00:00)
```

SVMC Result (Training Data)



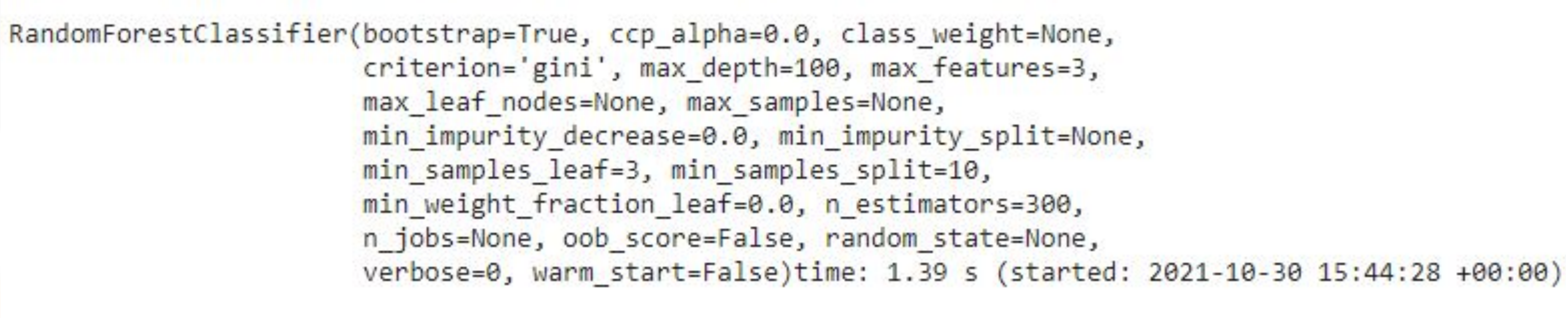
```
              precision    recall  f1-score   support

   Perempuan       1.00      1.00      1.00      1180
   Laki-laki       1.00      1.00      1.00      1180

    accuracy                           1.00      2360
   macro avg       1.00      1.00      1.00      2360
weighted avg       1.00      1.00      1.00      2360

time: 10.6 s (started: 2021-10-30 15:35:36 +00:00)
```

ANALYTICS PROCESS & ALGORITHM

Gender Classification

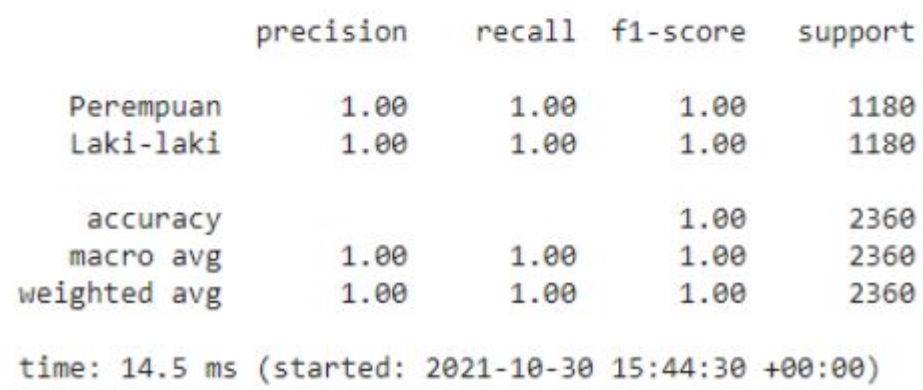
5 Random Forest Classifier

Hyperparameter Tuning



```
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=100, max_features=3,
                       max_leaf_nodes=None, max_samples=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=3, min_samples_split=10,
                       min_weight_fraction_leaf=0.0, n_estimators=300,
                       n_jobs=None, oob_score=False, random_state=None,
                       verbose=0, warm_start=False)time: 1.39 s (started: 2021-10-30 15:44:28 +00:00)
```

RFC Result (Training Data)



```
              precision    recall  f1-score   support

   Perempuan       1.00      1.00      1.00      1180
   Laki-laki       1.00      1.00      1.00      1180

    accuracy                           1.00      2360
   macro avg       1.00      1.00      1.00      2360
weighted avg       1.00      1.00      1.00      2360

time: 14.5 ms (started: 2021-10-30 15:44:30 +00:00)
```

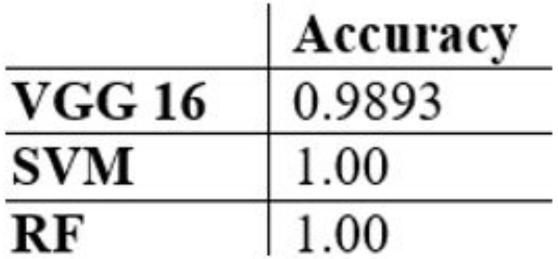
ANALYTICS PROCESS & ALGORITHM

Gender Classification

6 Result Each Model

The three models should not be overfitting because:

- For VGG 16 we can see that the visualization of the model does not show any severe overfitting
- For SVM and Random Forest it should not be overfitting because when performing Hyperparameter Tuning, it is already using cross-validation. Train some of the training data and predict some of the others, **but the validation score is not showing**

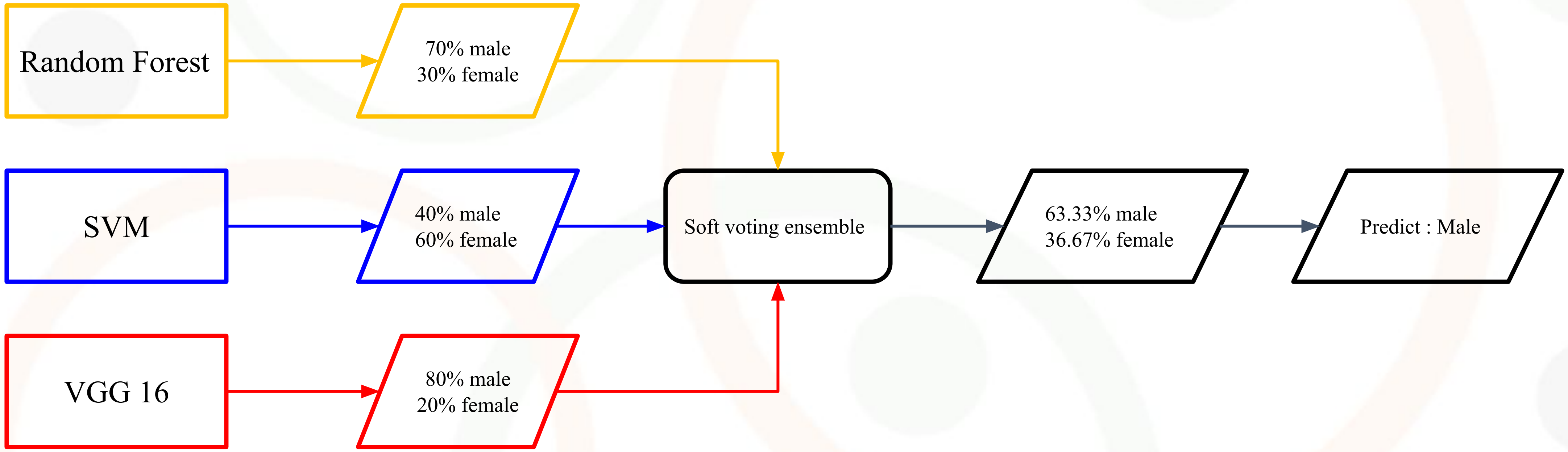


	Accuracy
VGG 16	0.9893
SVM	1.00
RF	1.00

ANALYTICS PROCESS & ALGORITHM

Gender Classification

7 Ensemble Learning for Gender (Soft Voting)



ANALYTICS PROCESS & ALGORITHM

Gender Classification

8 Handling Two or More Face

Ensemble Predict Male 97%

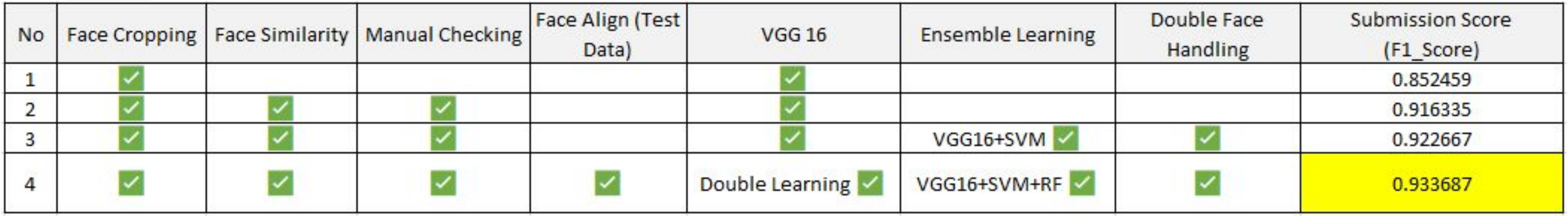
Ensemble Predict Female 93%

Win! Prediction
for 10.jpg is male

ANALYTICS PROCESS & ALGORITHM

Gender Classification

9 Submission History



No	Face Cropping	Face Similarity	Manual Checking	Face Align (Test Data)	VGG 16	Ensemble Learning	Double Face Handling	Submission Score (F1_Score)
1	☑				☑			0.852459
2	☑	☑	☑		☑			0.916335
3	☑	☑	☑		☑	VGG16+SVM ☑	☑	0.922667
4	☑	☑	☑	☑	Double Learning ☑	VGG16+SVM+RF ☑	☑	0.933687

ANALYTICS PROCESS & ALGORITHM

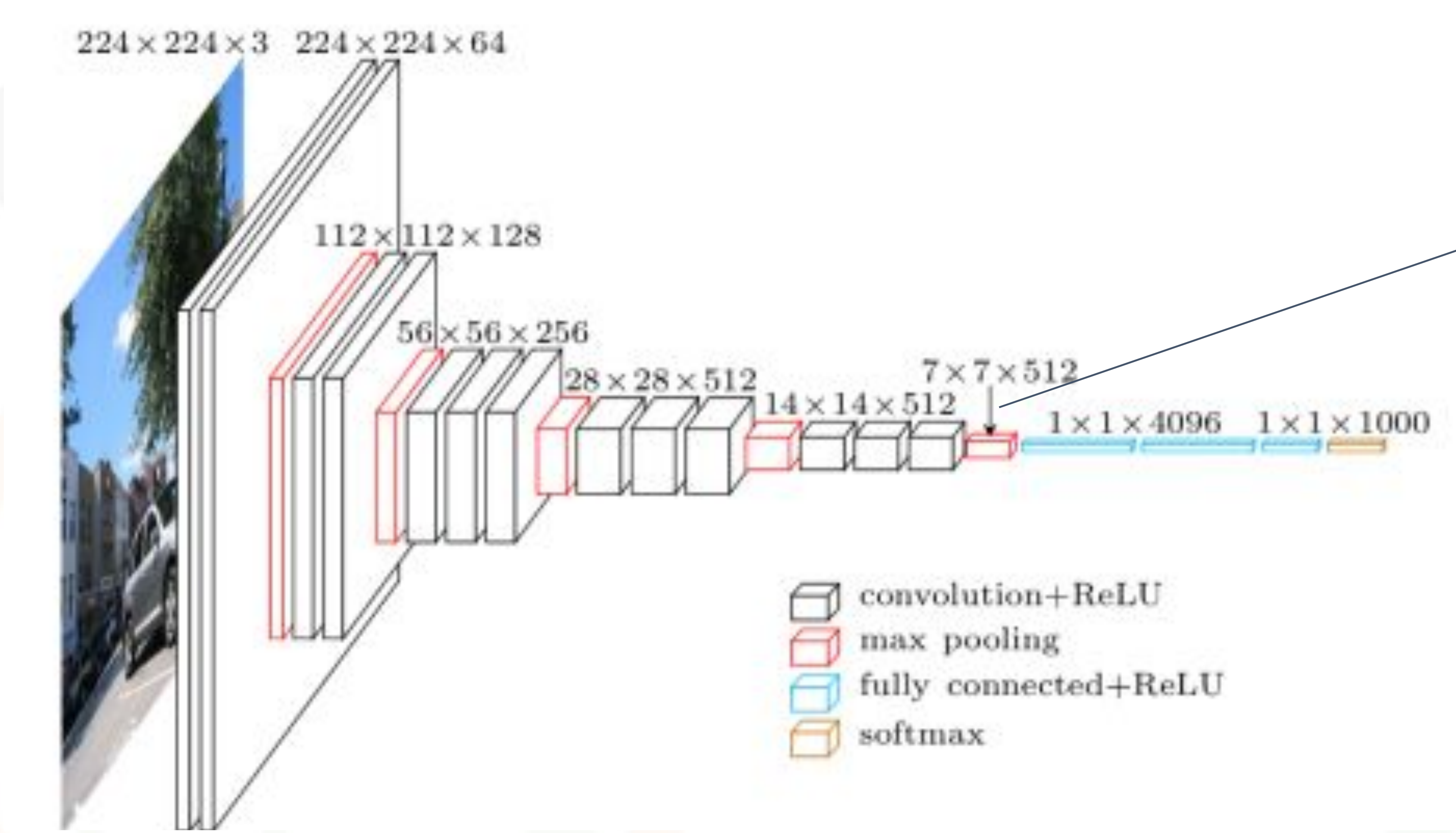
Age Estimation

1. Feature Extraction with VGG_Face
2. Support Vector Machine Regressor
3. Lasso Regressor
4. Ensemble Learning for Age (Stacking Lasso)

ANALYTICS PROCESS & ALGORITHM

Age Estimation

1 Feature Extraction with **VGG_Face**



224 x 224 x 3 to 7 x 7 x 512

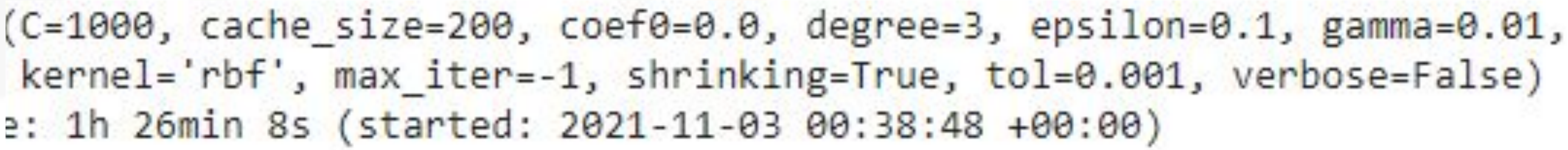
Each image have 25088 feature after extraction for ML modelling

ANALYTICS PROCESS & ALGORITHM

Age Estimation

2 Support Vector Machine Regressor

Hyperparameter Tuning



```
(C=1000, cache_size=200, coef0=0.0, degree=3, epsilon=0.1, gamma=0.01,
 kernel='rbf', max_iter=-1, shrinking=True, tol=0.001, verbose=False)
e: 1h 26min 8s (started: 2021-11-03 00:38:48 +00:00)
```

SVMR Result (Training Data)

MSE in months = 2624.544 ± 360.051

MSE in years (estimation) = 18.226 ± 2.500

ANALYTICS PROCESS & ALGORITHM

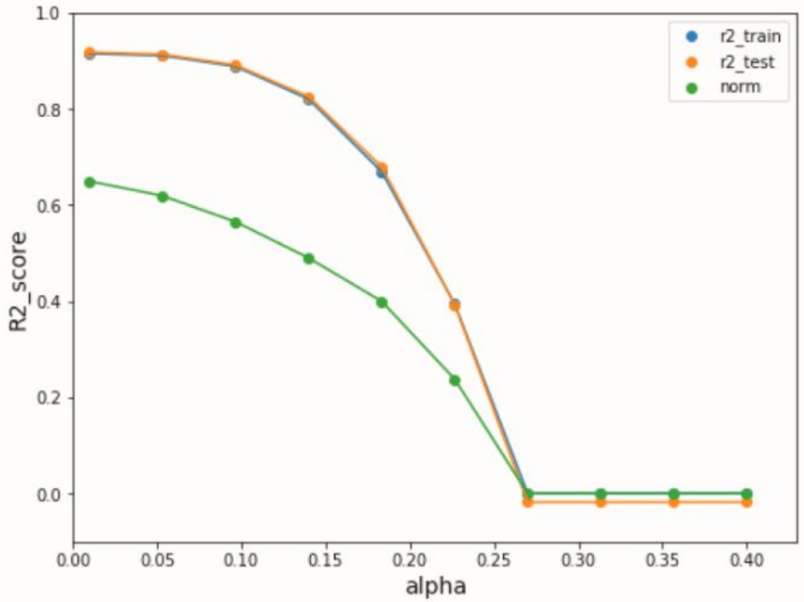
Age Estimation

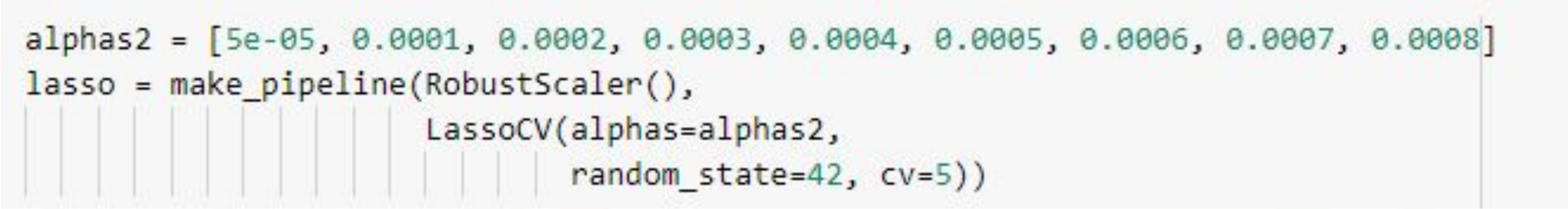
3 Lasso Regressor

$$\hat{y}_i = w_0 + \sum_{j=1}^{m} X_{ij} w_j$$

$$J(w) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{m} |w_j|$$

$$\|w\|^2 = \sum_{j=1}^{m} |w_j|^2$$



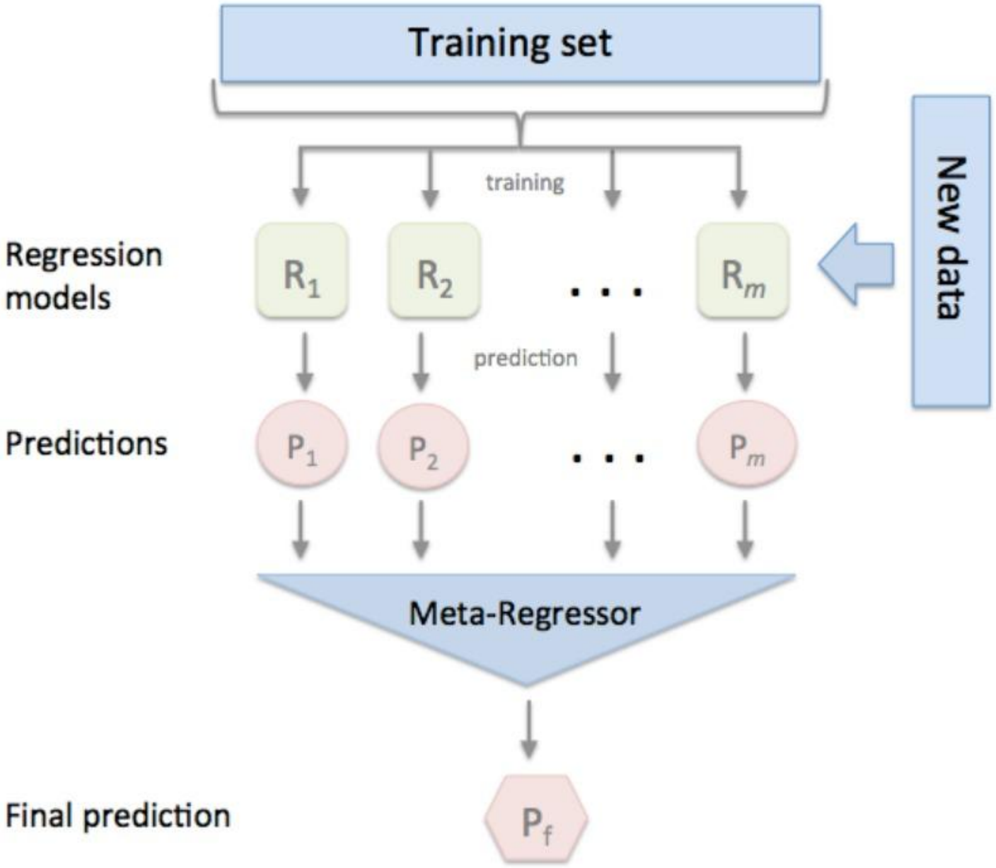


```
alphas2 = [5e-05, 0.0001, 0.0002, 0.0003, 0.0004, 0.0005, 0.0006, 0.0007, 0.0008]
lasso = make_pipeline(RobustScaler(),
                      LassoCV(alphas=alphas2,
                              random_state=42, cv=5))
```

ANALYTICS PROCESS & ALGORITHM

Age Estimation

4 Ensemble Learning for Age (Stacking Lasso)

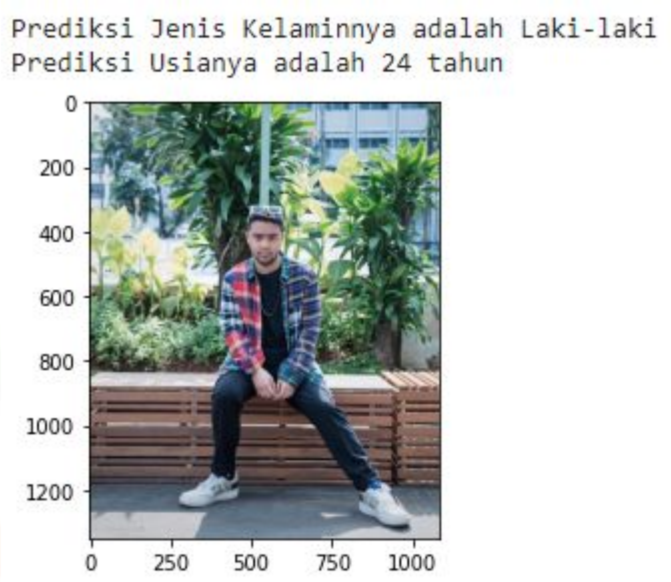


Base : SVMR
Meta : Lasso Regressor

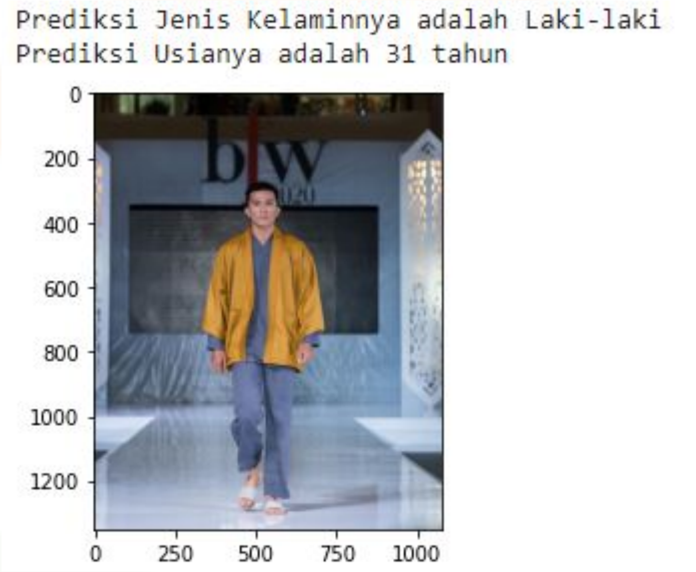
MSE in months = 2613.59 ± 357.74
MSE in years (est) = 18.15 ± 2.48

	MSE (months)
SVMR	2624.544
SVMR + Lasso	2613.59
Improvisation	10.954

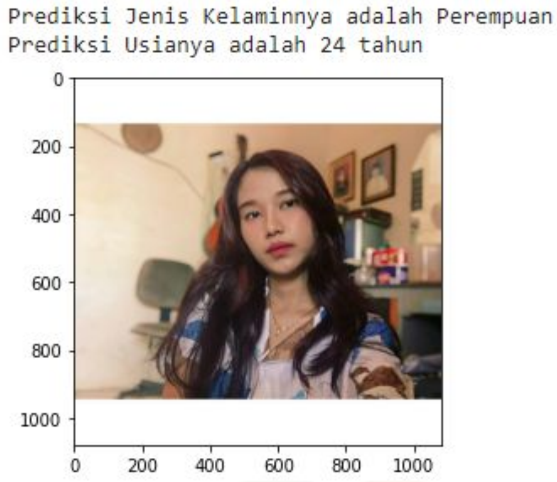
PREDICTION RESULT



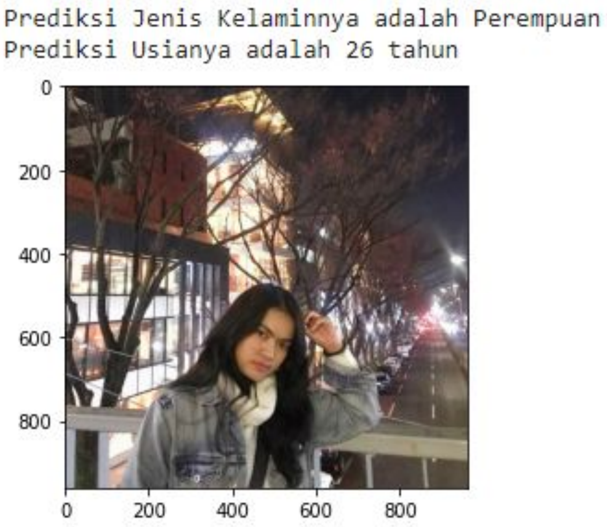
```
Prediksi Jenis Kelaminnya adalah Laki-laki
Prediksi Usianya adalah 24 tahun
```



```
Prediksi Jenis Kelaminnya adalah Laki-laki
Prediksi Usianya adalah 31 tahun
```



```
Prediksi Jenis Kelaminnya adalah Perempuan
Prediksi Usianya adalah 24 tahun
```



```
Prediksi Jenis Kelaminnya adalah Perempuan
Prediksi Usianya adalah 26 tahun
```

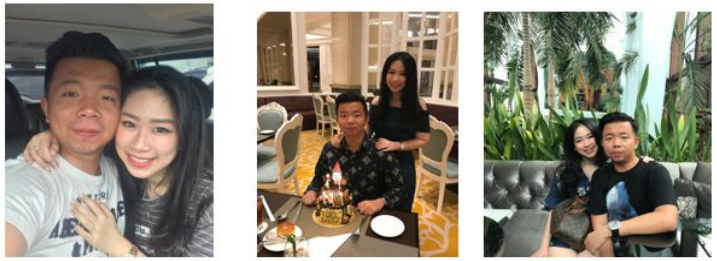
SUBMISSION RESULT

Gender Classification
F1 Score = 0.9337
(10th in Standings)

Age Estimation
MSE (years) = 24.3525
(1st in Standings)

INSIGHT & RECOMMENDATIONS

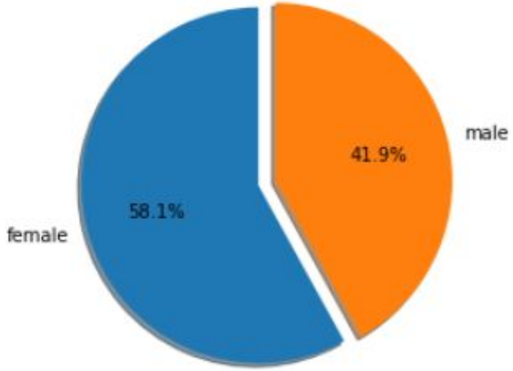
- Double anchor face in training data => When collecting data, it is better if the data train 1 folder contains only 1 anchor face so that there is no need to do manual checks



- Double Face in testing data => It's better if there is only one person in the face data testing because in the submission of the output, only one is requested => Double Face Handling

INSIGHT & RECOMMENDATIONS

- Imbalance Data (Gender) => SMOTETomek Oversampling

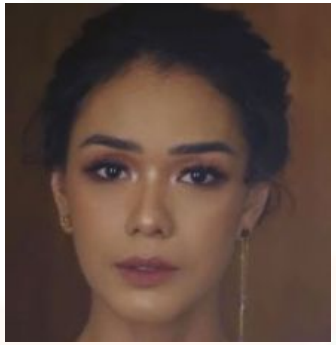


0	1187	→	1180
1	810		1180

Improving model to 0.9337 in F1_score

- Age Estimation is Subjective and based solely on appearance => Why we not trying turn into age brackets classification problem?

Folder 40 : 24 Years Old



Folder 37 : 31 Years Old

CONCLUSION

By conducting ensemble learning soft voting for gender classification and stacking for age estimation, we managed to get an F1 score of 0.9337 and an MSE of 24.3525 years (RMSE 4.934 years)

When compared with the DeepFace* library which got an F1 score of 0.9566 and an MAE of 4.65 years, it can be said that the ensemble learning model is quite effective for gender classification and age estimation.

*) DeepFace library is better because it uses more data [Uses IMDB (7 GB) and Wikipedia (1 GB) datasets]

TERIMA KASIH

Kampus Merdeka
INDONESIA JAYA

TUT WURI HANDAYANI

Puspresnas
Pusat Prestasi Nasional