pipeline_dataset_v2

August 24, 2025

1 PIPELINE: model 4, dataset v2(https://app.roboflow.com/sidsproject/bac used-pehn2-8iy1j/generate/preprocessing),different types of embeddings

1.1 PIPELINE

1.1.1 1. load model face_detection

1.1.2 2. extract embeddings from dataset

1.1.3 3. retrieval to evaluate embeddings goodness

1.1.4 4. train model classification

1.2 1. load the model face_detection

Step 1 is the same for each pipeline, so we do it once at the beginning.

Import dependencies

[1]:
```
%load_ext autoreload
%autoreload 2

from embeddings import EmbeddingBuilder
import os
import ipynbname

from retrieval import ImageRetrieval
from classifier import Classifier
import file_manager

project_dir = f"{os.getcwd().
 ↪split('SIDS_revelation_project')[0]}SIDS_revelation_project"
image_dataset_path = f"{project_dir}/datasets/onback_onstomach_v2"

model_path = f"{project_dir}/models/4.weights/best.pt"
```

Choose how to process the dataset: 1. “extract_features”: extracting features and labels 2. “extract_features_imageswithinference”: extracting features and labels and saving predicted images with bboxes 3. “load”: loading features and labels

[2]:
```
emb_builder = EmbeddingBuilder(model_path, image_dataset_path, "load")
```

```
Extracting dataset info from .coco.json
file:-----------------------------------------
Dataset contains 1506 valid samples, and labels are {'baby_on_back': 1,
'baby_on_stomach': 2}
-------------------------------------------------------------------------------
----------

Loading features from
.csv-------------------------------------------------------------
Features loaded succesfully, in particular there are 1506 files in the dataset
-------------------------------------------------------------------------------
----------

Embedding builder initialized
successfully-----------------------------------------------
Face detection model: 4 (YOLOv8)
Dataset: /home/terra/Documents/AI_engineering/SIDS-
project/python_project/SIDS_revelation_project/datasets/onback_onstomach_v2
Dataset dimension: 1506
Dataset labels: {'baby_safe': 0, 'baby_unsafe': 1}
-------------------------------------------------------------------------------
----------
```

1.3 2. Extract embeddings from dataset

Create embeddings

[3]:
```
embeddings = emb_builder.embedding_flags()
```

```
Creation of flags features
embedding-------------------------------------------------------
1506 embedding created
-------------------------------------------------------------------------------
----------
```

[4]:
```
embeddings.head()
```

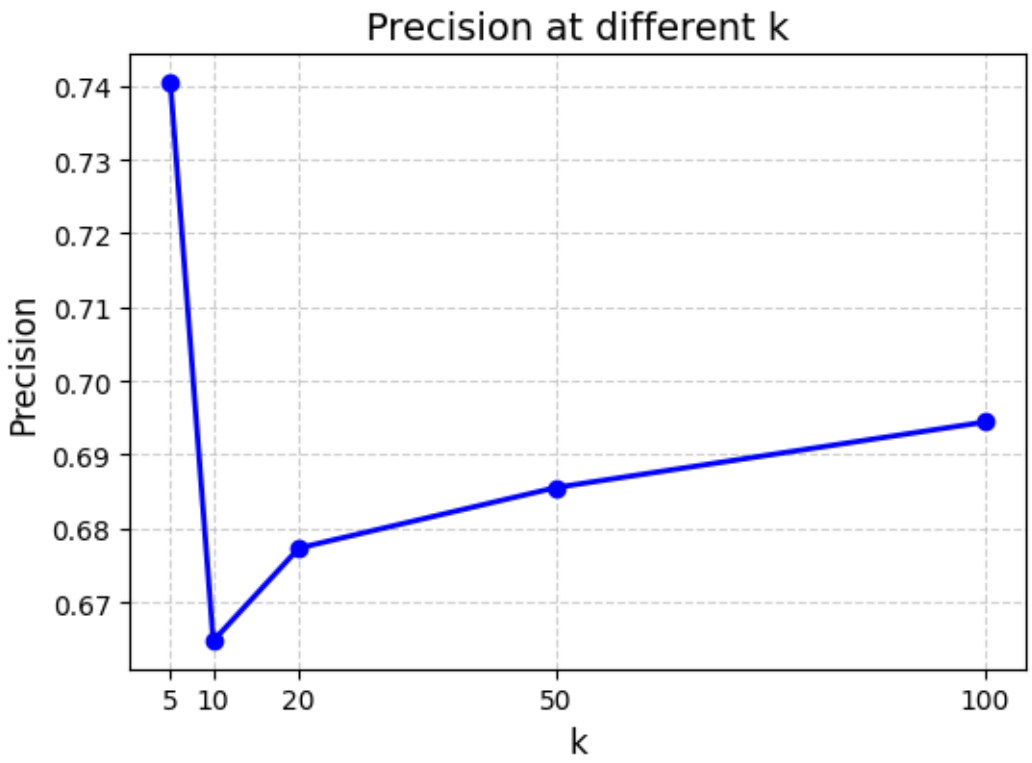
[4]:

	flag_eye1	flag_eye2	flag_nose	flag_mouth
0	0	0	0	0
1	1	1	1	1
2	0	0	0	0
3	1	1	1	1
4	1	1	1	1

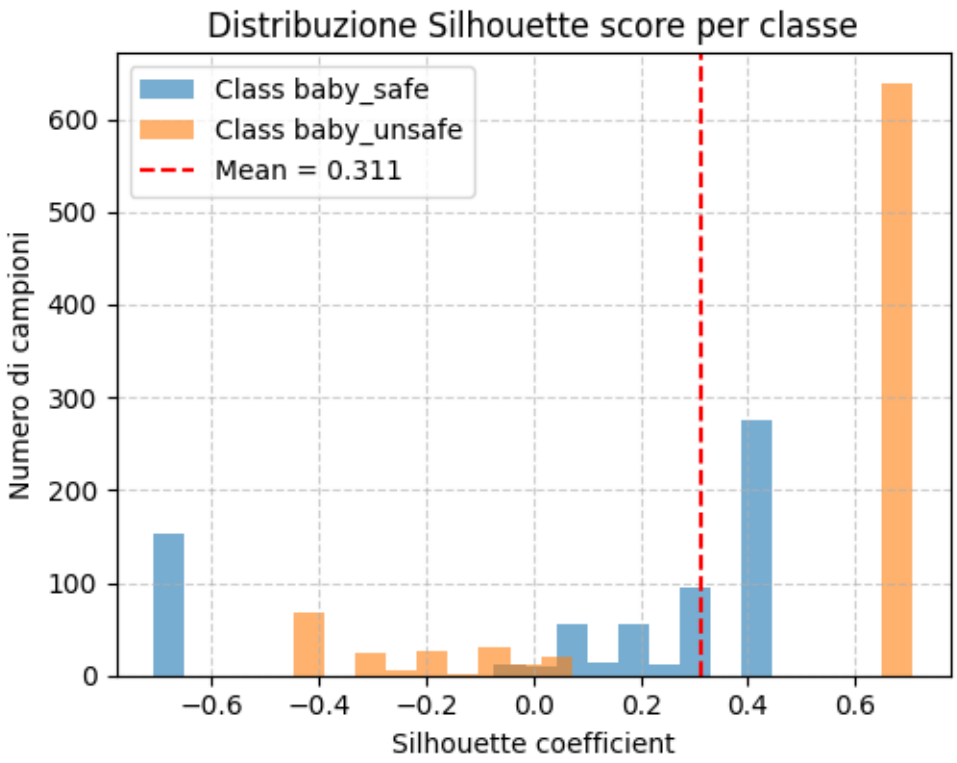
1.4 3. Retrieval to evaluate embedding goodness

[5]:
```
ret = ImageRetrieval(embeddings, emb_builder.y, emb_builder.image_paths,␣
 ↪image_dataset_path, emb_builder.classes_bs)
ret.report("euclidean")
```

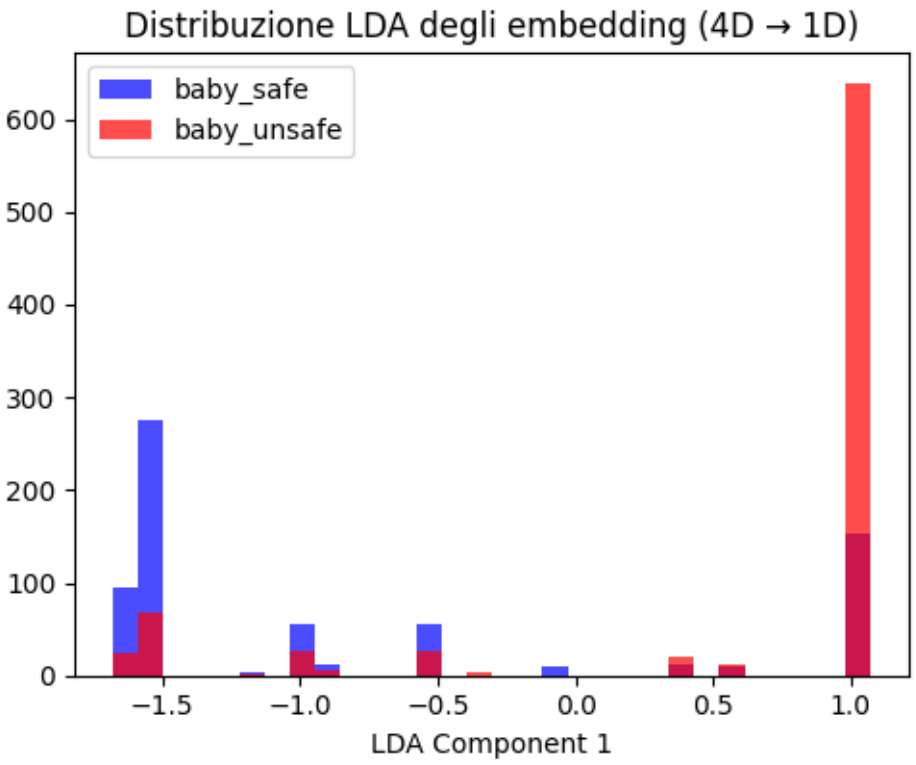
```
Precision at different
k:----------------------------------------------------------------------
```

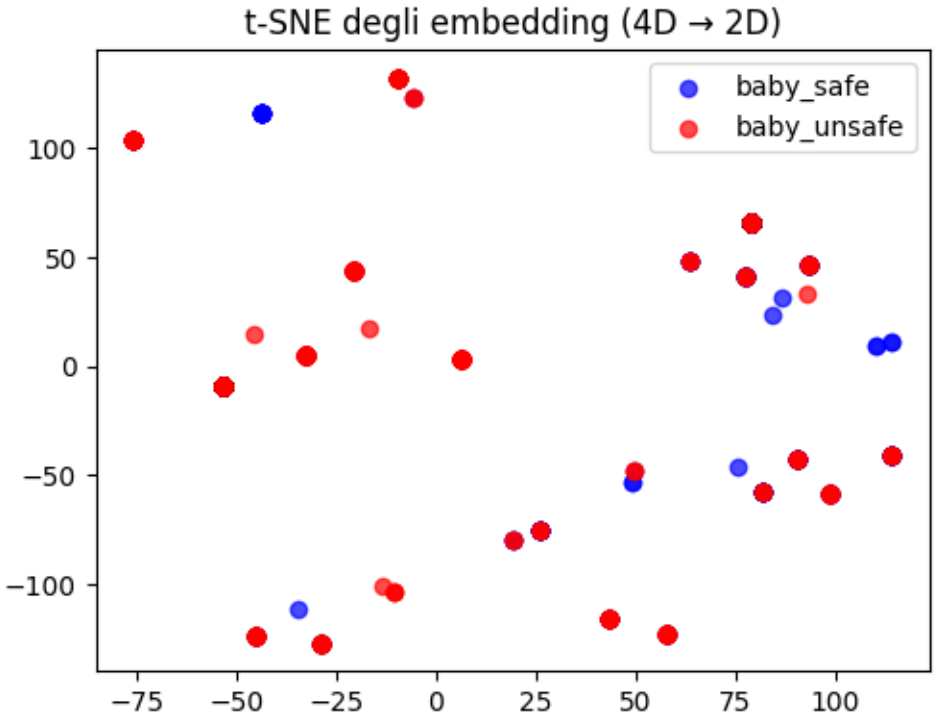



```
Recall at
R-----------------------------------------------------------------------------
0.6558580010669358

Silhouette
score---------------------------------------------------------------------------
```

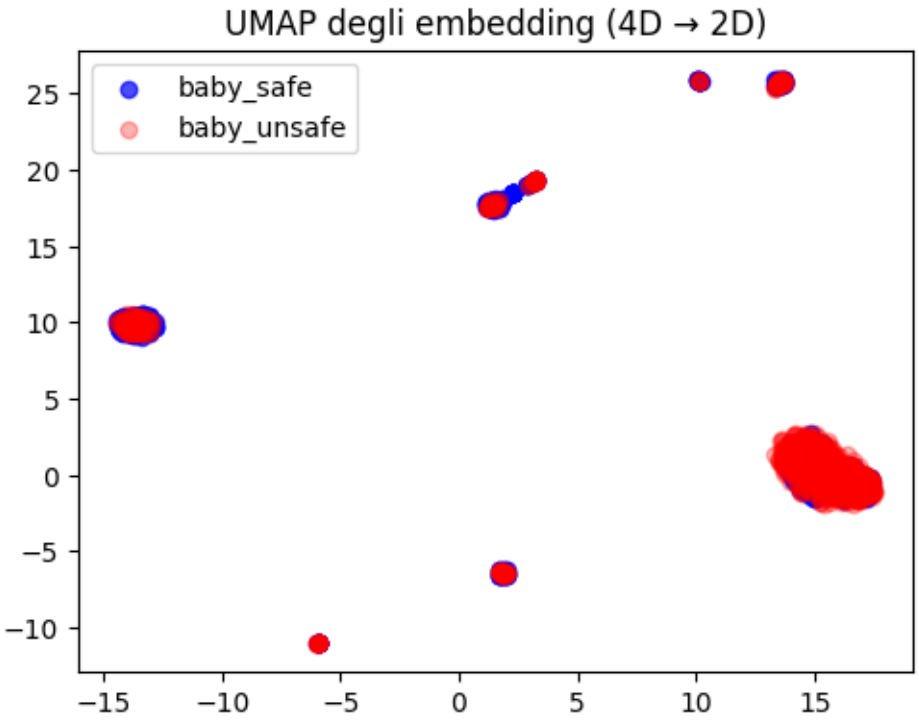


Embeddings
distributions----------


Distribuzione LDA degli embedding (4D → 1D)
baby_safe
baby_unsafe
600
500
400
300
200
100
0
−1.5
−1.0
−0.5
0.0
0.5
1.0
LDA Component 1

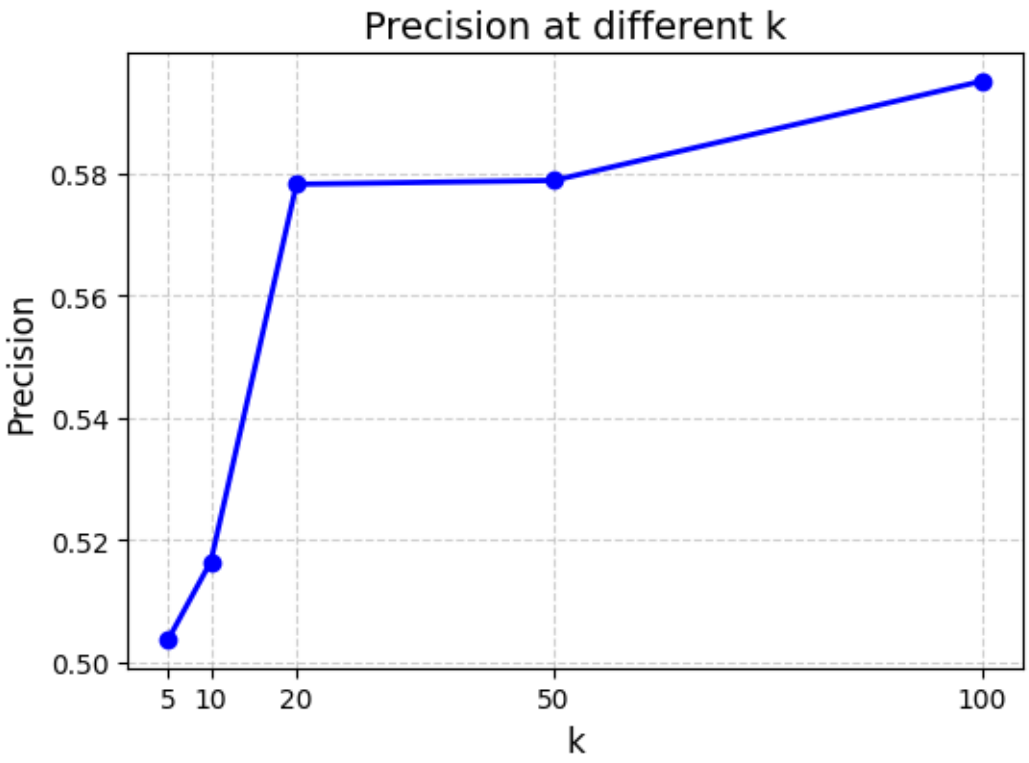

t-SNE degli embedding (4D → 2D)
baby_safe
baby_unsafe
100
50
0
−50
−100
−75
−50
−25
0
25
50
75
100

```
/home/terra/anaconda3/envs/SIDS_revelation_project/lib/python3.10/site-
packages/umap/umap_.py:1952: UserWarning: n_jobs value 1 overridden to 1 by
setting random_state. Use no seed for parallelism.
  warn(
```



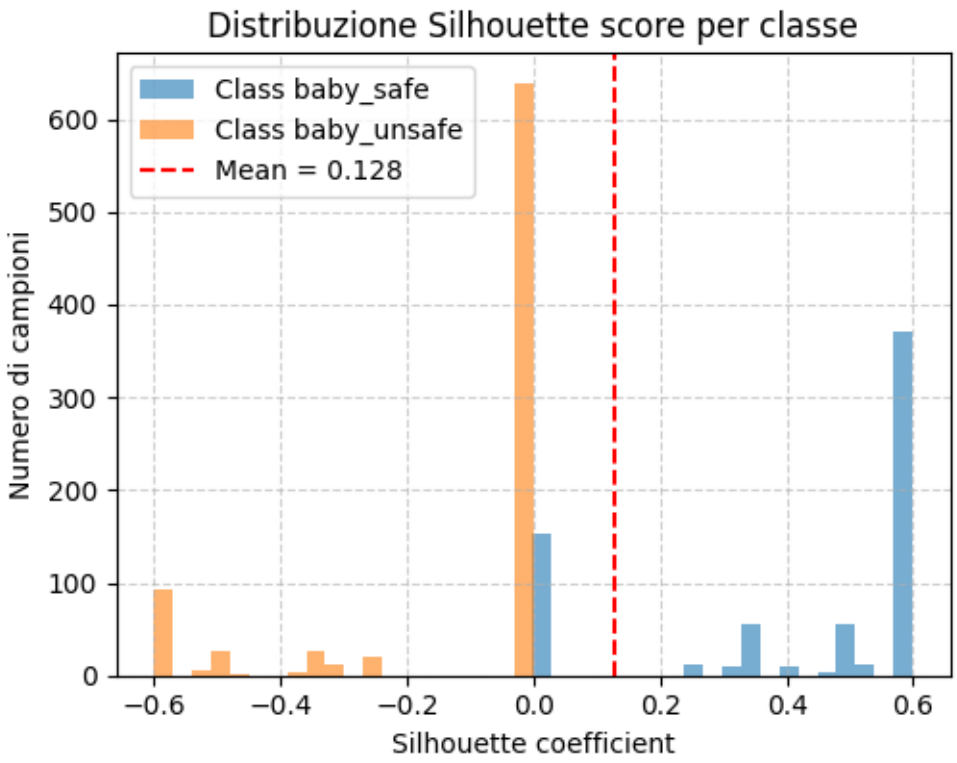
[6]:
```
ret.report("cosine")
```

```
Precision at different
k:----------------------------------------------------------------------
```

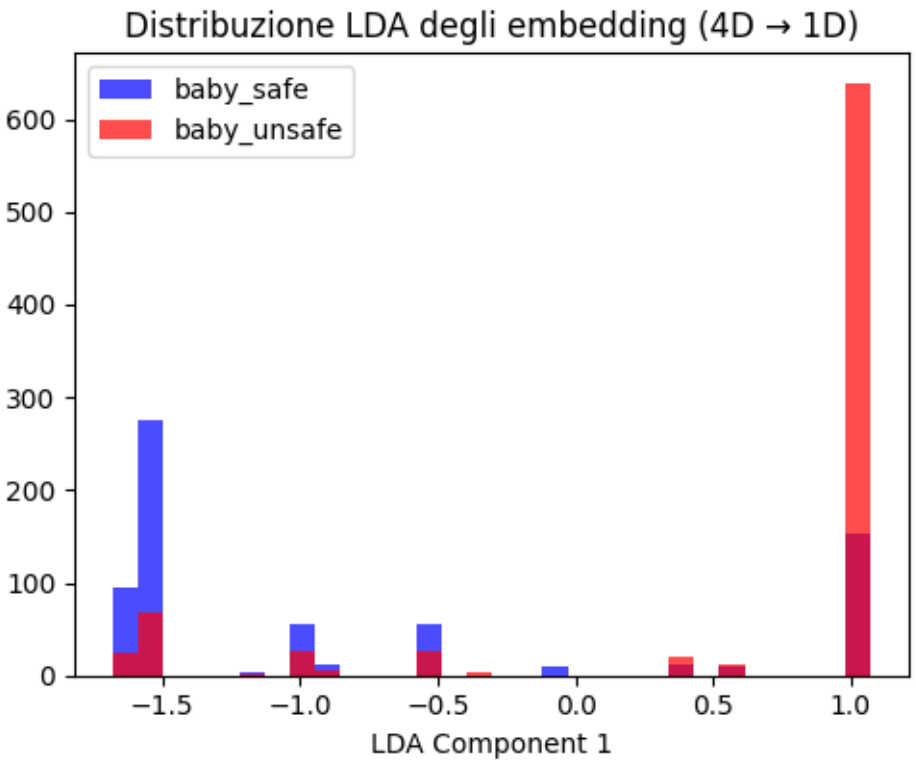


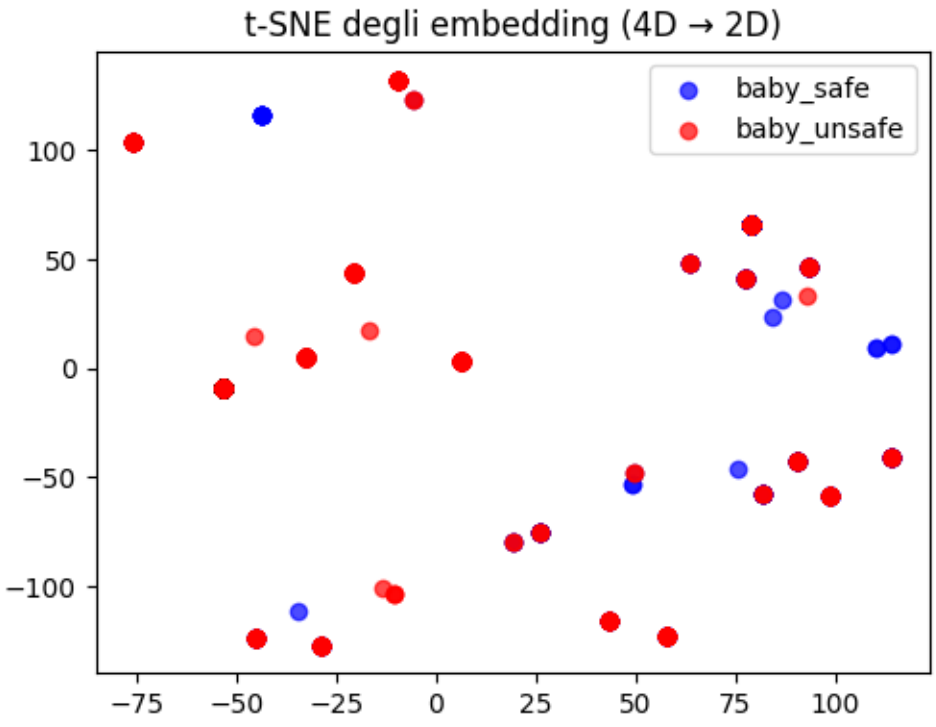
Recall at
R----------
0.5754122273017442

Silhouette
score----------

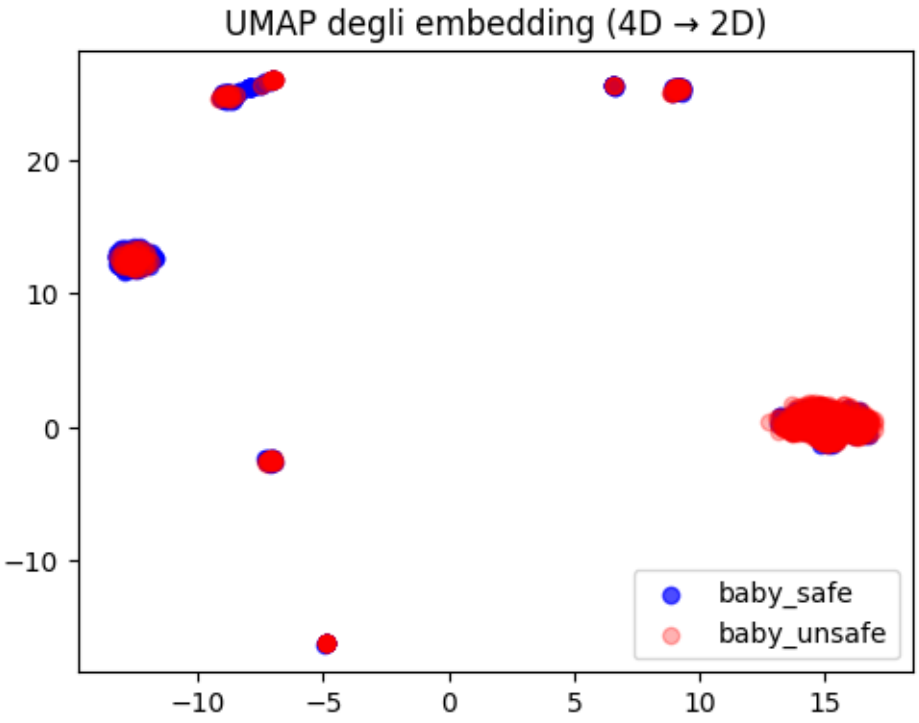


Embeddings
distributions----------


Distribuzione LDA degli embedding (4D → 1D)
baby_safe
baby_unsafe
600
500
400
300
200
100
0
−1.5
−1.0
−0.5
0.0
0.5
1.0
LDA Component 1

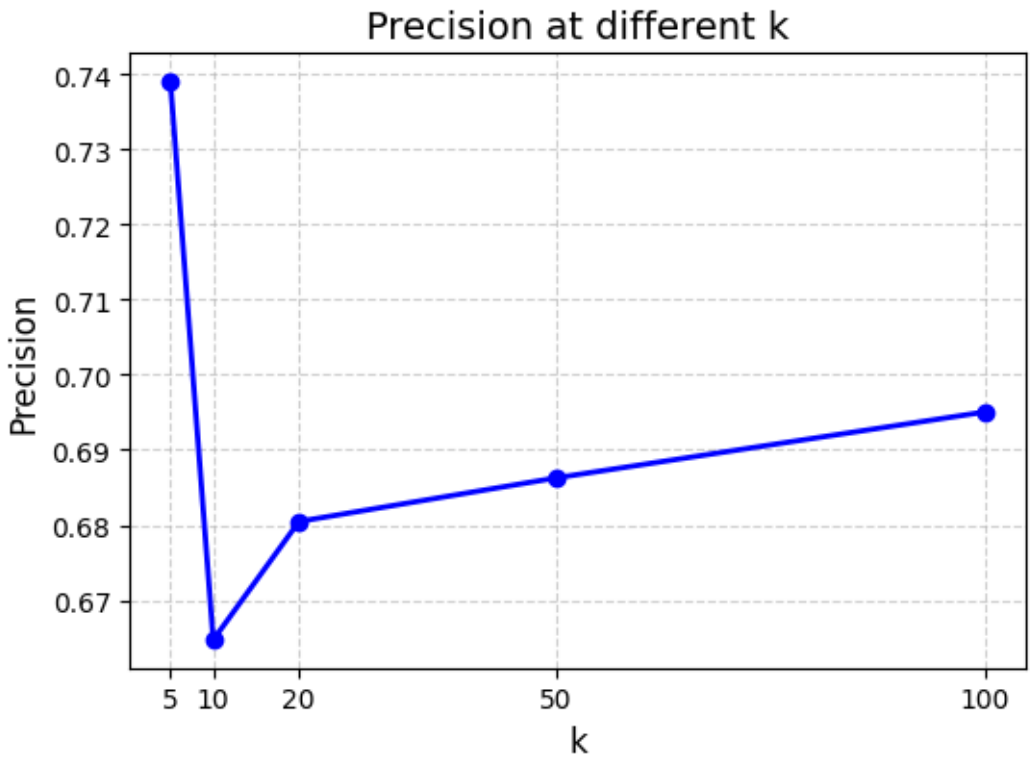

t-SNE degli embedding (4D → 2D)
baby_safe
baby_unsafe
100
50
0
−50
−100
−75
−50
−25
0
25
50
75
100

```
/home/terra/anaconda3/envs/SIDS_revelation_project/lib/python3.10/site-
packages/umap/umap_.py:1952: UserWarning: n_jobs value 1 overridden to 1 by
setting random_state. Use no seed for parallelism.
  warn(
```

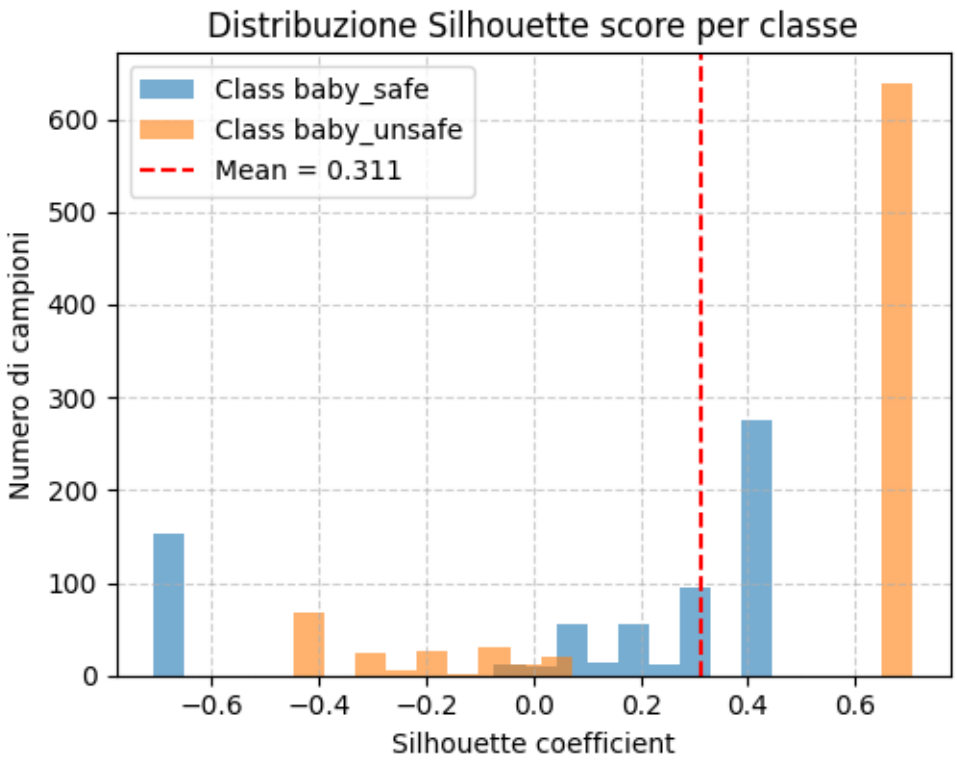


[7]:
```
ret.report("minkowski")
```

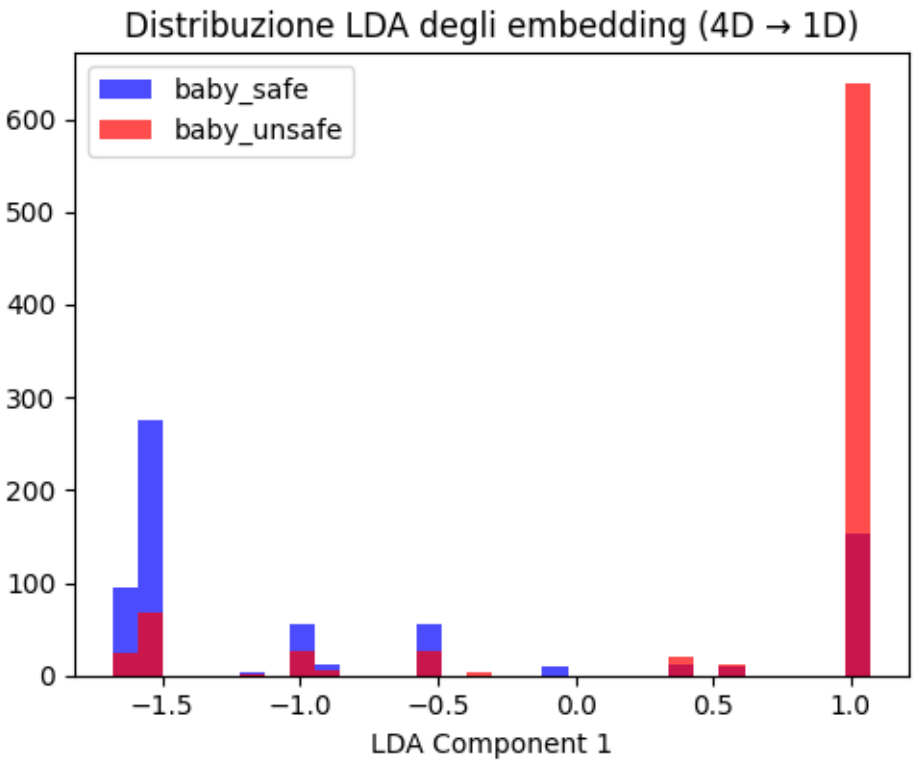
```
Precision at different
k:----------------------------------------------------------------------
```

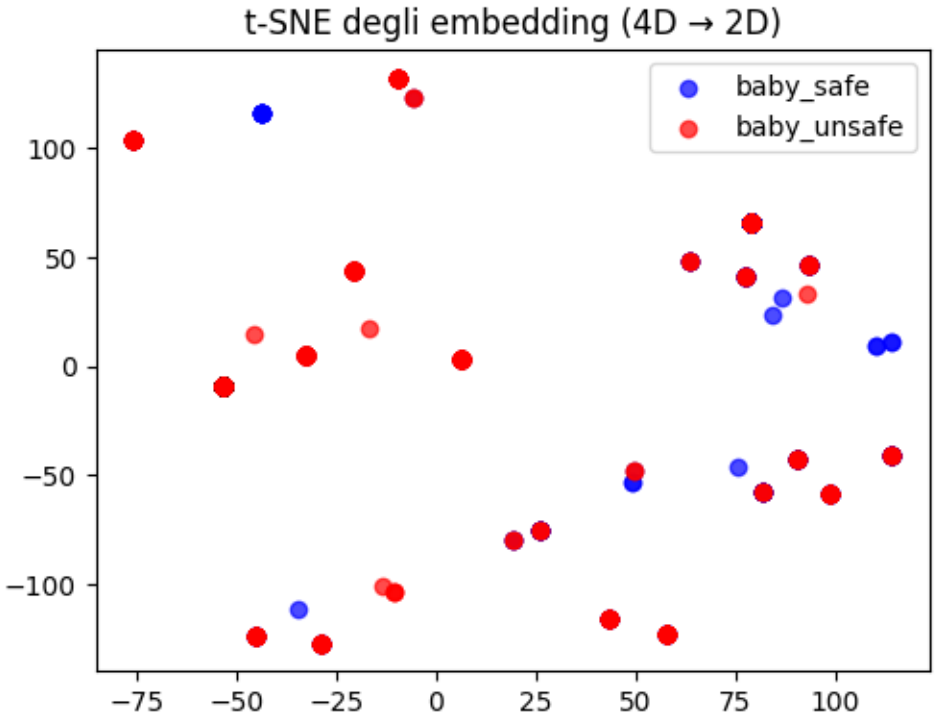



```
Recall at
R-------------------------------------------------------------------------------
0.6502013355033254

Silhouette
score-----------------------------------------------------------------------------
```

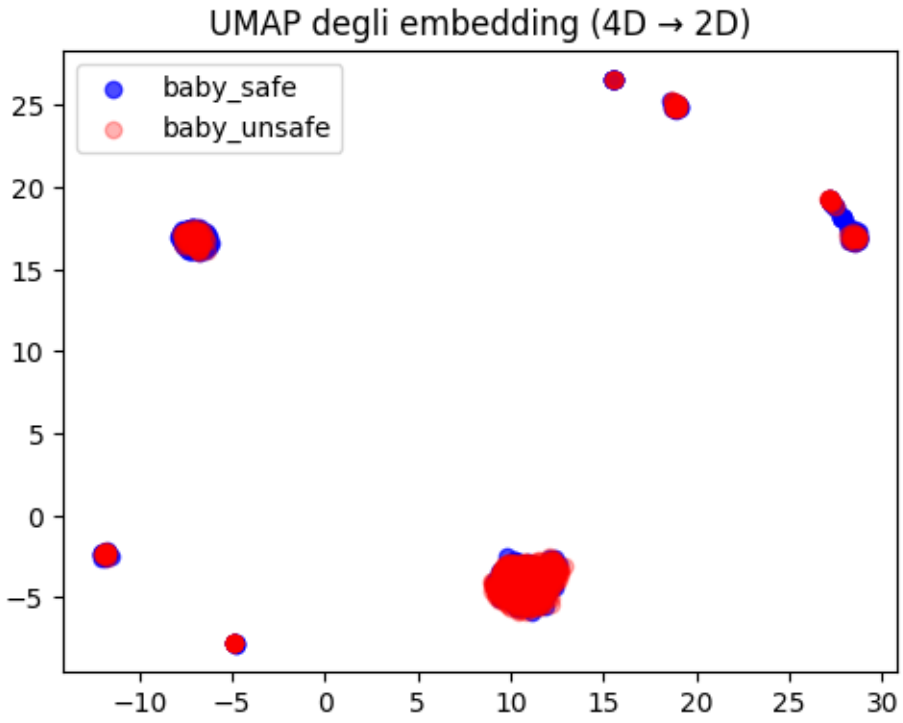


Embeddings
distributions----------


Distribuzione LDA degli embedding (4D → 1D)
baby_safe
baby_unsafe
600
500
400
300
200
100
0
−1.5
−1.0
−0.5
0.0
0.5
1.0
LDA Component 1


t-SNE degli embedding (4D → 2D)
baby_safe
baby_unsafe
100
50
0
−50
−100
−75
−50
−25
0
25
50
75
100

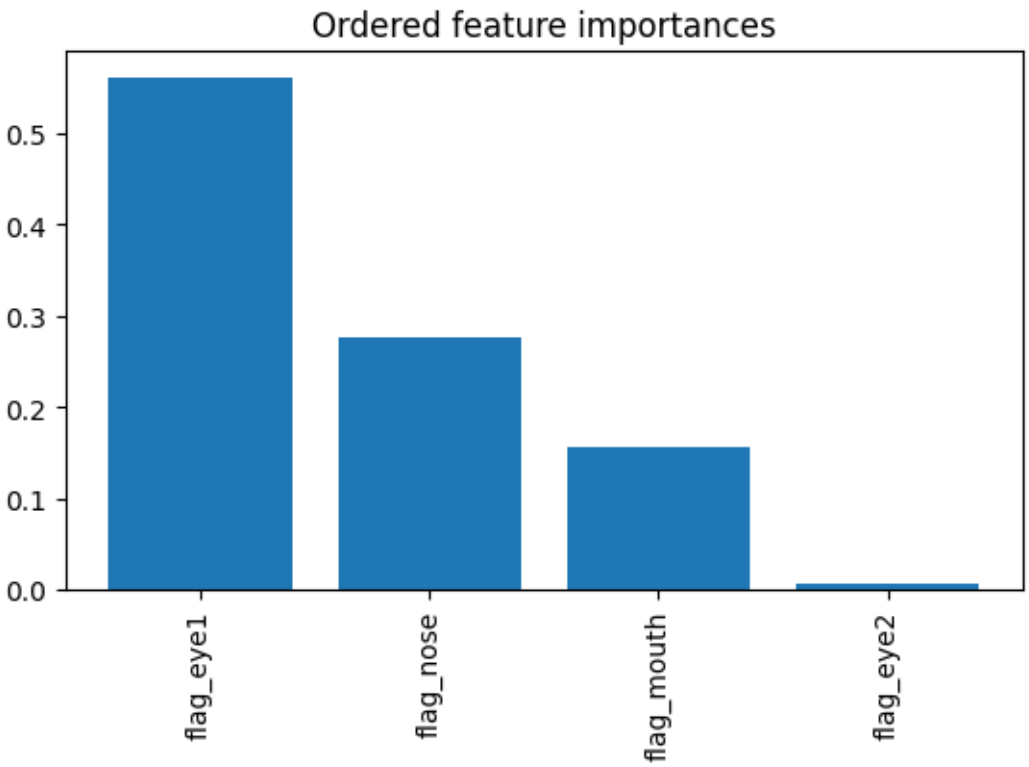
```
/home/terra/anaconda3/envs/SIDS_revelation_project/lib/python3.10/site-
packages/umap/umap_.py:1952: UserWarning: n_jobs value 1 overridden to 1 by
setting random_state. Use no seed for parallelism.
  warn(
```

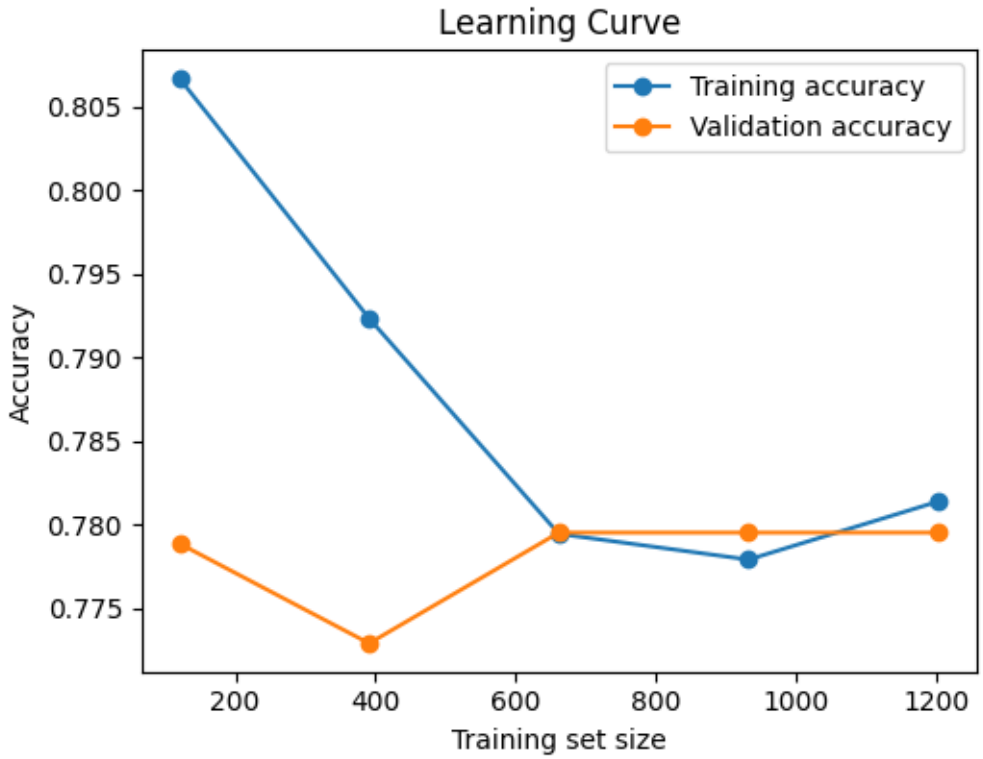


1.5 4. Train model classification

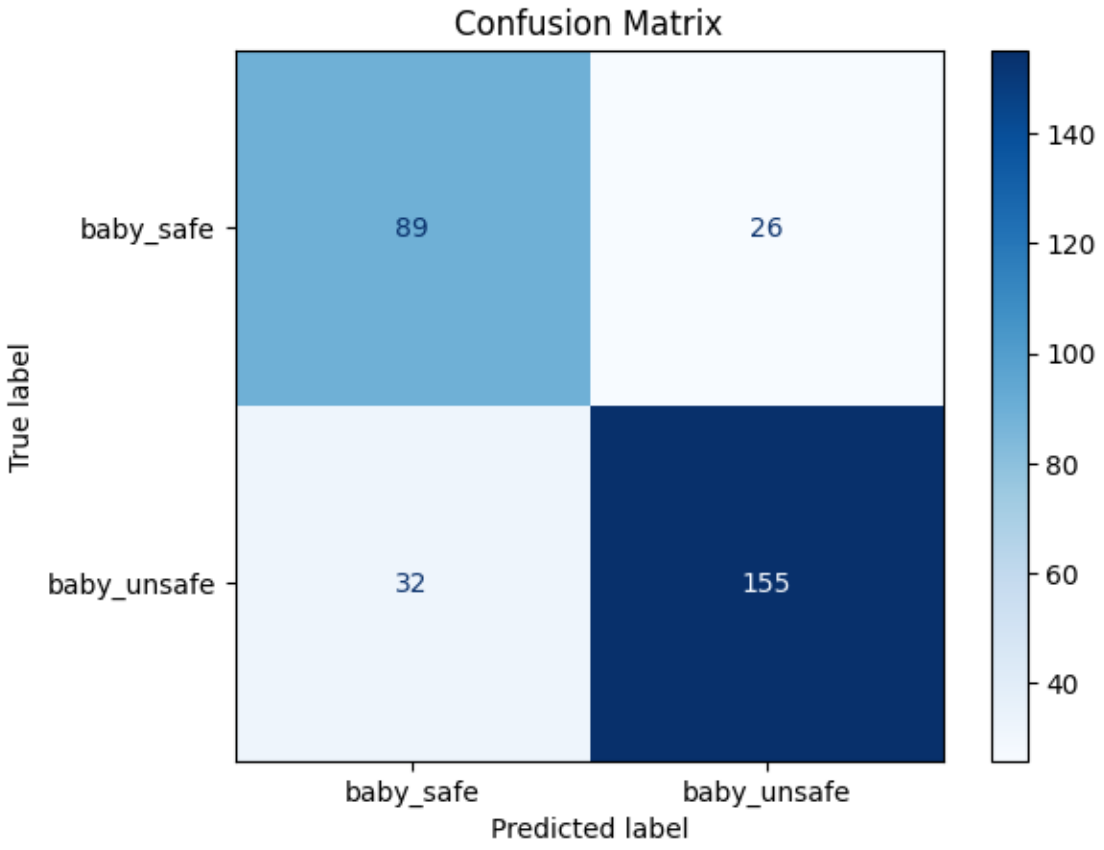
[8]:
```
clf = Classifier(embeddings, emb_builder.y, emb_builder.classes_bs)
clf.random_forest()
```

```
----------------------------------------------------------------------------------
----------
-------------------------------------FIRST
ANALYSIS-------------------------------------
----------------------------------------------------------------------------------
----------
```


Ordered feature importances
0.5
0.4
0.3
0.2
0.1
0.0
flag_eye1
flag_nose
flag_mouth
flag_eye2


Learning Curve
Training accuracy
Validation accuracy
Accuracy
0.805
0.800
0.795
0.790
0.785
0.780
0.775
200
400
600
800
1000
1200
Training set size

```
Dataset labels:----------------------------------------
{'baby_safe': 0, 'baby_unsafe': 1}

Report-------------------------------------------------
              precision    recall  f1-score   support

   baby_safe       0.74      0.77      0.75       115
 baby_unsafe       0.86      0.83      0.84       187

    accuracy                           0.81       302
   macro avg       0.80      0.80      0.80       302
weighted avg       0.81      0.81      0.81       302

Confusion matrix---------------------------------------
```



1.6 2. Extract embeddings from dataset

Create embeddings

[9]:
```
embeddings = emb_builder.embedding_all_features()
```

```
Creation of all features
embedding-------------------------------------------------------
1506 embedding created
-------------------------------------------------------------------------------
----------
```

[10]:
```
embeddings.head()
```

[10]:

	flag_eye1	flag_eye2	flag_nose	flag_mouth	x_eye1	y_eye1	x_eye2
0	0	0	0	0	-1.000000	-1.000000	-1.000000
1	1	1	1	1	0.747866	0.955937	0.746319
2	0	0	0	0	-1.000000	-1.000000	-1.000000
3	1	1	1	1	0.533123	0.143157	0.374687
4	1	1	1	1	0.859706	0.598094	0.846316

	y_eye2	x_nose	y_nose	x_mouth	y_mouth	eye_distance
0	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000
1	0.680452	0.709361	0.853581	0.650598	0.840860	0.275490
2	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000
3	0.148591	0.457184	0.163462	0.469082	0.201713	0.158529
4	0.728694	0.827010	0.658031	0.788980	0.648129	0.131284

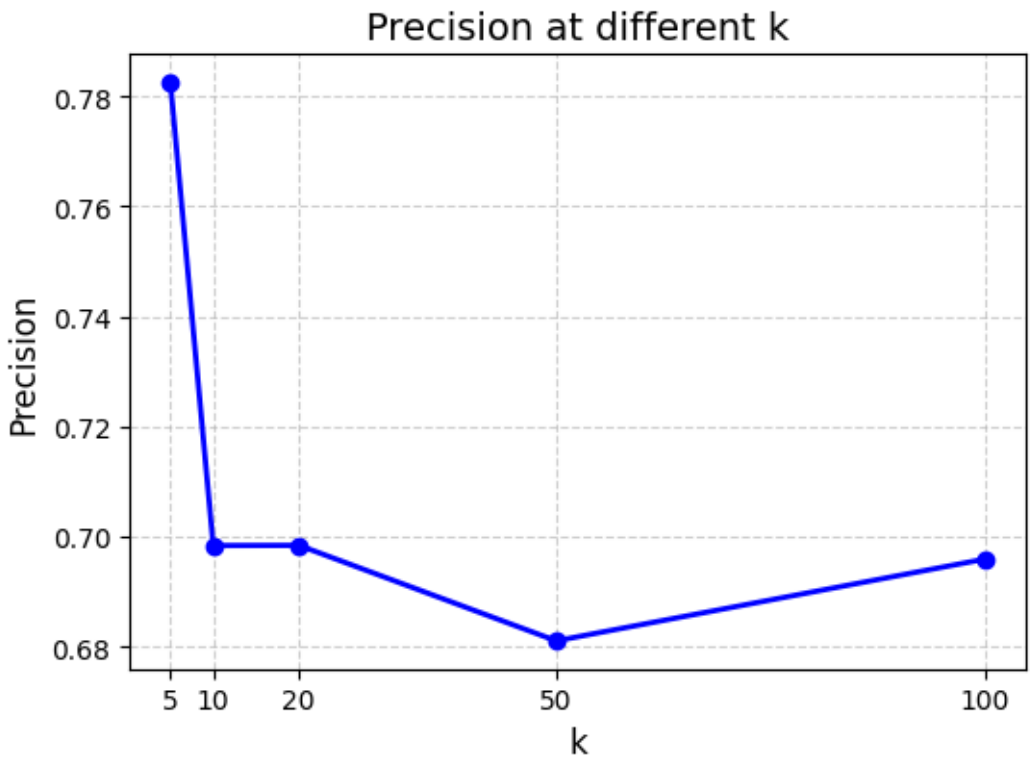
	face_vertical_length	face_angle_vertical	face_angle_horizontal
0	-1.000000	-1.000000	-1.000000
1	0.060125	122.830696	147.334481
2	-1.000000	-1.000000	-1.000000
3	0.040059	87.692268	154.811363
4	0.039298	104.018161	136.106213

	symmetry_diff
0	0.000000
1	0.085138
2	0.000000
3	0.004185
4	0.002724

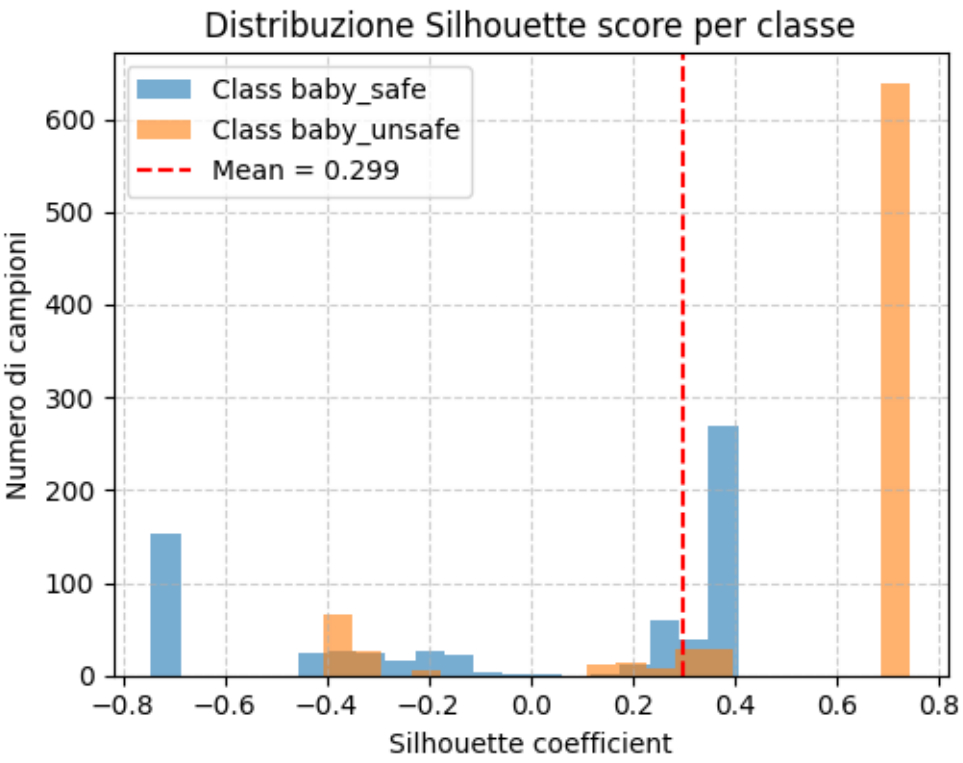
1.7 3. Retrieval to evaluate embedding goodness

[11]:
```
ret = ImageRetrieval( embeddings, emb_builder.y, emb_builder.image_paths,␣
 ↪image_dataset_path, emb_builder.classes_bs)
ret.report("euclidean")
```

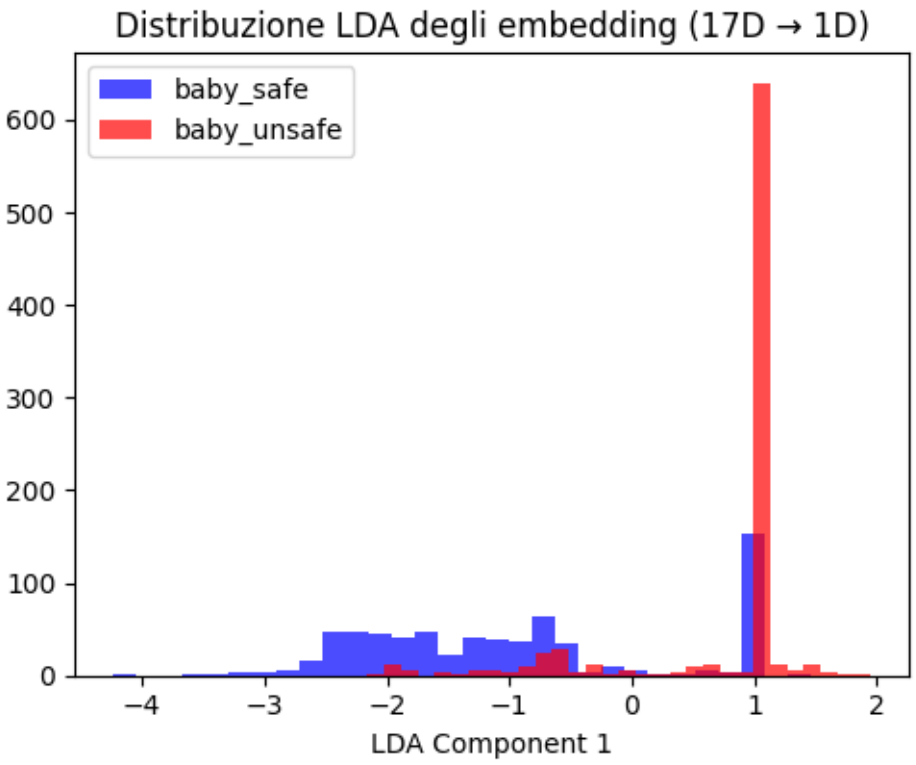
```
Precision at different
k:----------------------------------------------------------------------
```

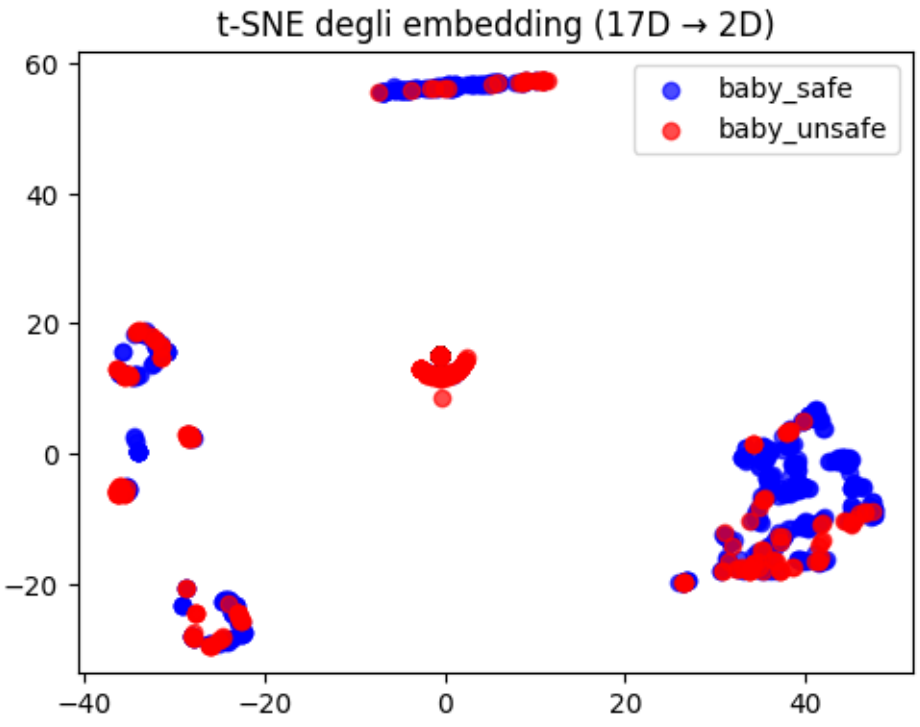



```
Recall at
R-----------------------------------------------------------------------------
0.6126016574337287

Silhouette
score--------------------------------------------------------------------------
```

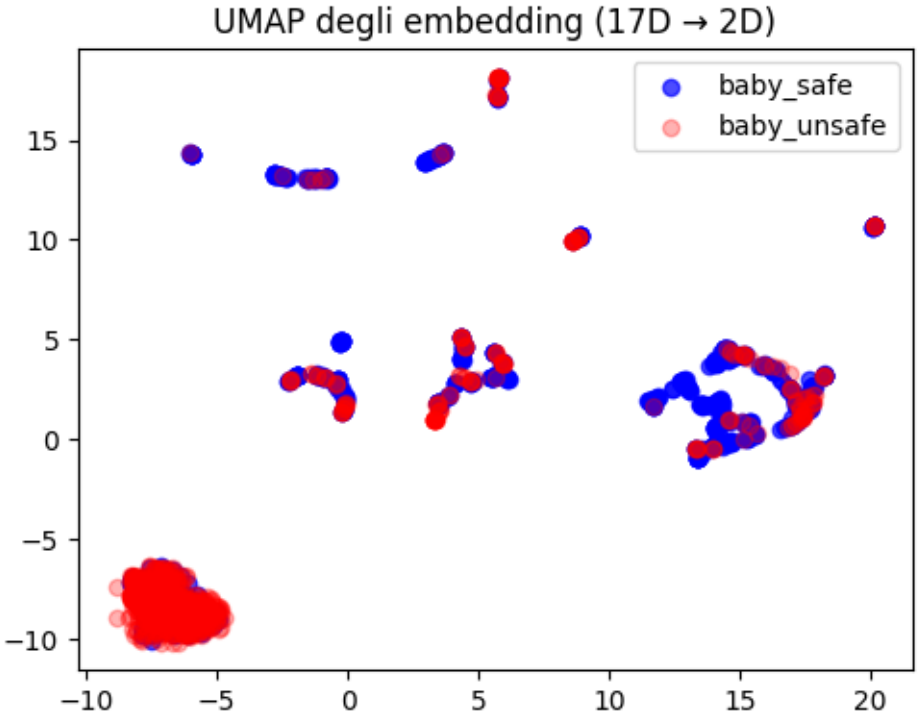


Embeddings
distributions----------


Distribuzione LDA degli embedding (17D → 1D)
baby_safe
baby_unsafe
600
500
400
300
200
100
0
−4
−3
−2
−1
0
1
2
LDA Component 1

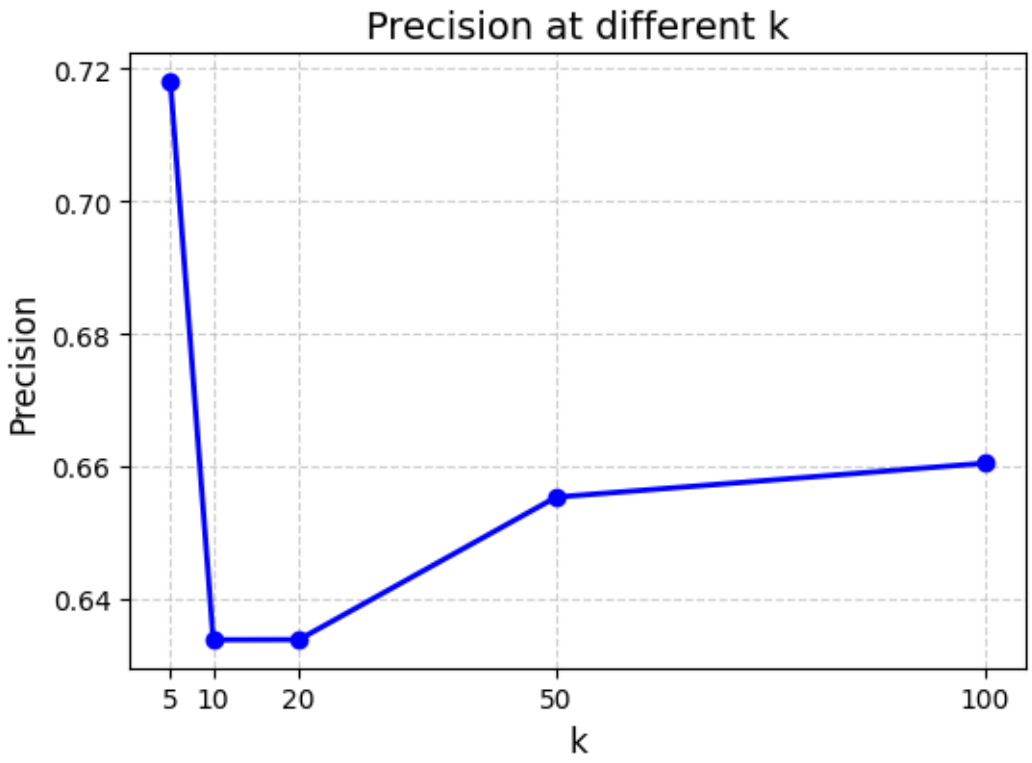

t-SNE degli embedding (17D → 2D)
baby_safe
baby_unsafe
60
40
20
0
−20
−40
−20
0
20
40

```
/home/terra/anaconda3/envs/SIDS_revelation_project/lib/python3.10/site-
packages/umap/umap_.py:1952: UserWarning: n_jobs value 1 overridden to 1 by
setting random_state. Use no seed for parallelism.
  warn(
/home/terra/anaconda3/envs/SIDS_revelation_project/lib/python3.10/site-
packages/sklearn/manifold/_spectral_embedding.py:328: UserWarning: Graph is not
fully connected, spectral embedding may not work as expected.
  warnings.warn(
```

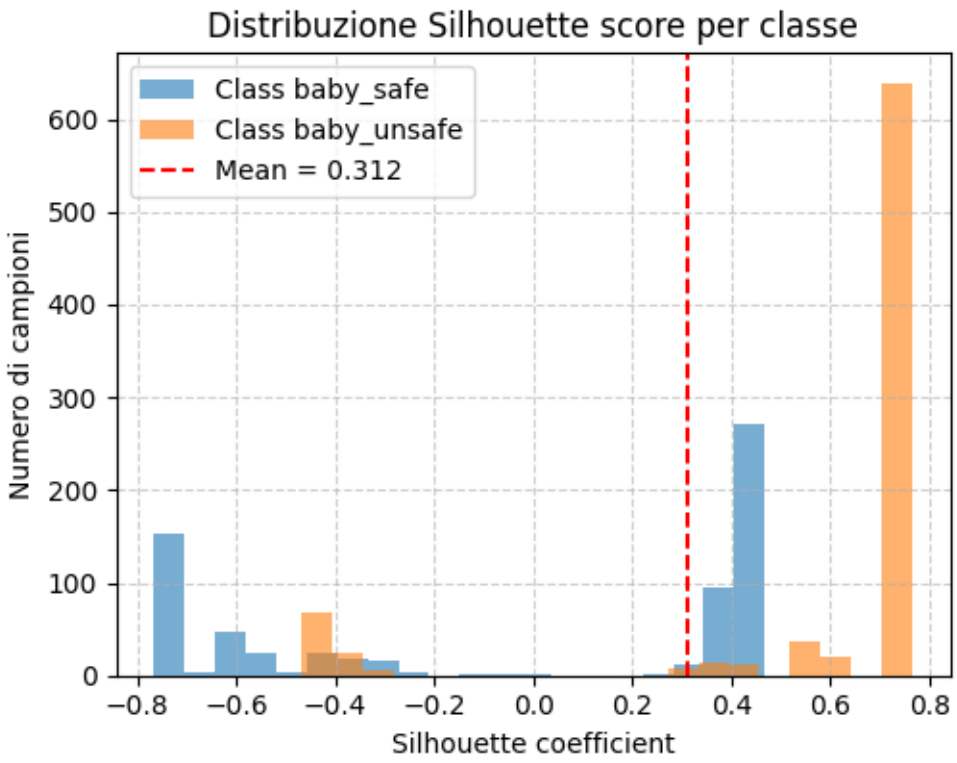


[12]:
```
ret.report("cosine")
```

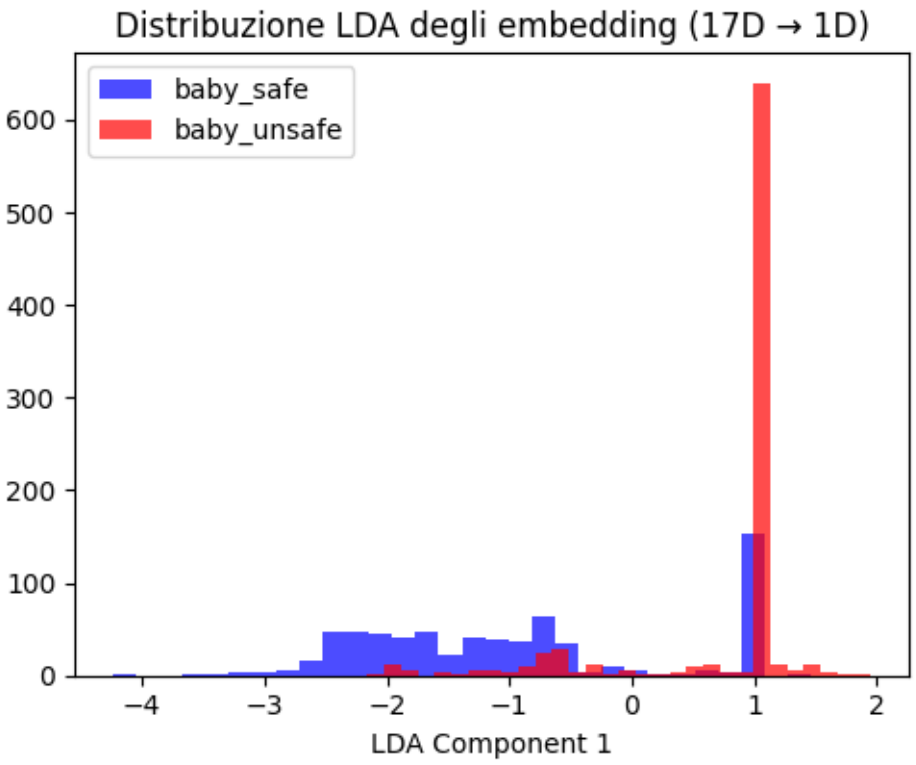
```
Precision at different
k:----------------------------------------------------------------------
```

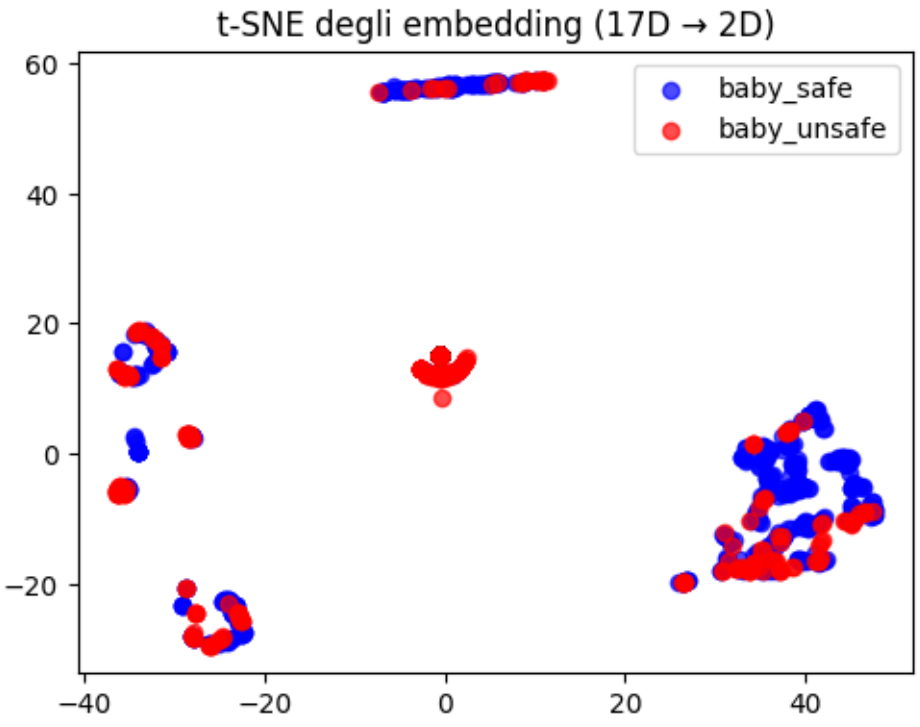



```
Recall at
R-------------------------------------------------------------------------------
0.6158959434877811

Silhouette
score-----------------------------------------------------------------------
```

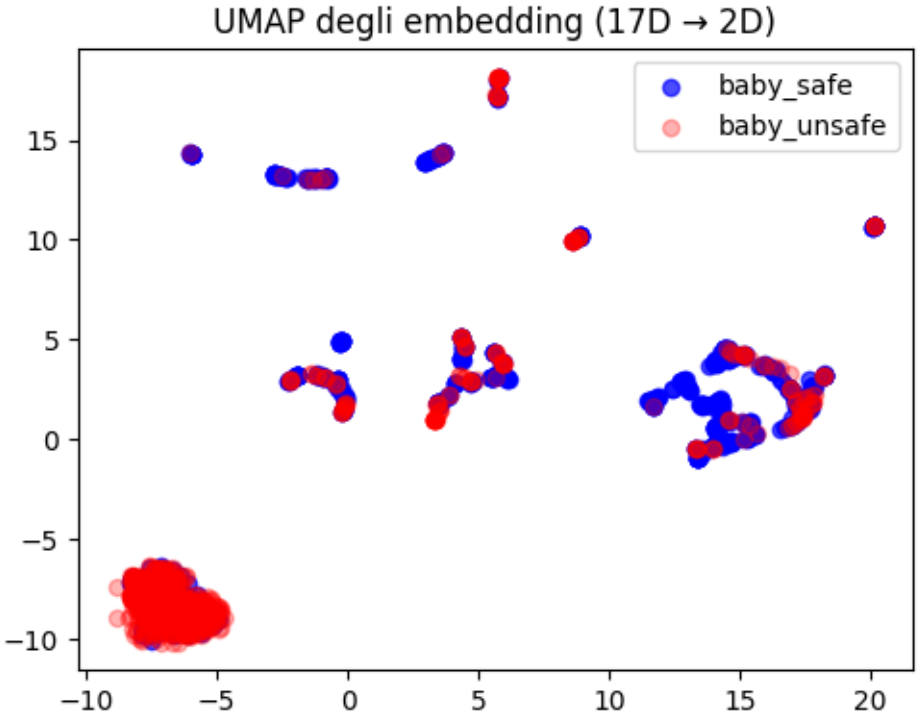


Embeddings
distributions----------


Distribuzione LDA degli embedding (17D → 1D)
baby_safe
baby_unsafe
600
500
400
300
200
100
0
−4
−3
−2
−1
0
1
2
LDA Component 1

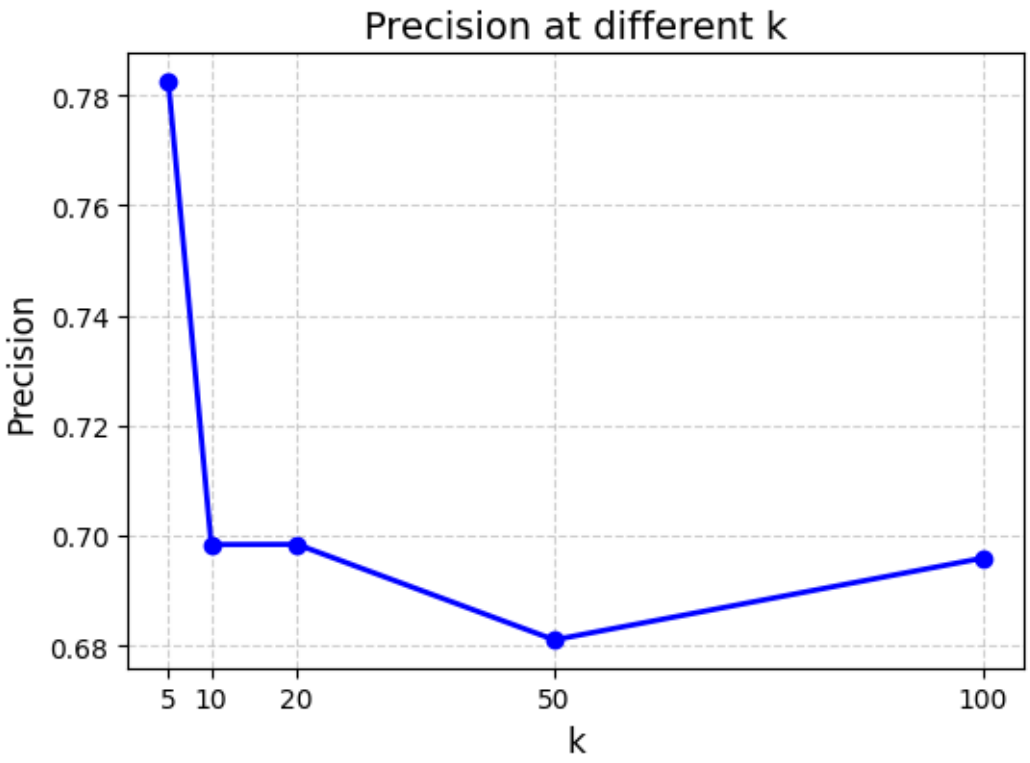

t-SNE degli embedding (17D → 2D)
baby_safe
baby_unsafe
60
40
20
0
−20
−40
−20
0
20
40

```
/home/terra/anaconda3/envs/SIDS_revelation_project/lib/python3.10/site-
packages/umap/umap_.py:1952: UserWarning: n_jobs value 1 overridden to 1 by
setting random_state. Use no seed for parallelism.
  warn(
/home/terra/anaconda3/envs/SIDS_revelation_project/lib/python3.10/site-
packages/sklearn/manifold/_spectral_embedding.py:328: UserWarning: Graph is not
fully connected, spectral embedding may not work as expected.
  warnings.warn(
```



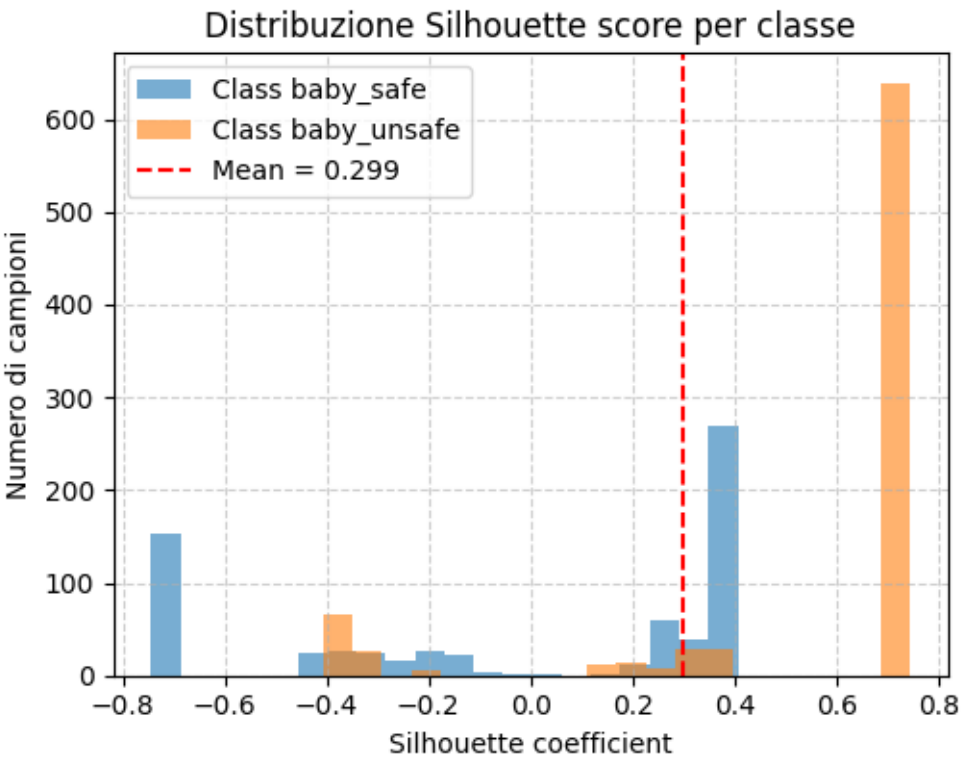
[13]:
```
ret.report("minkowski")
```

```
Precision at different
k:----------------------------------------------------------------------
```

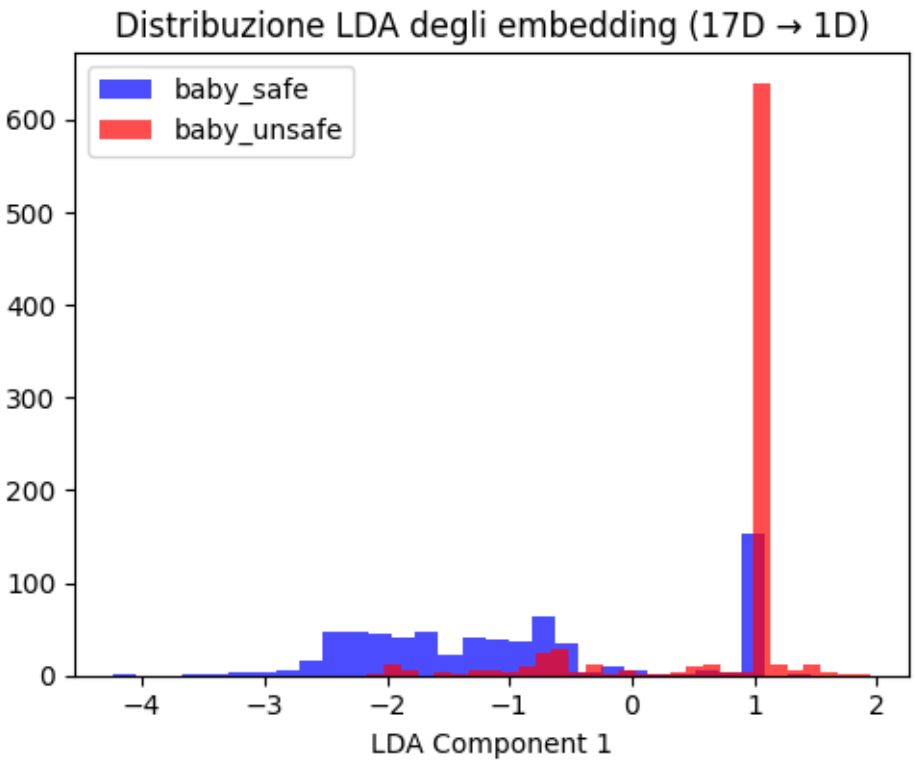


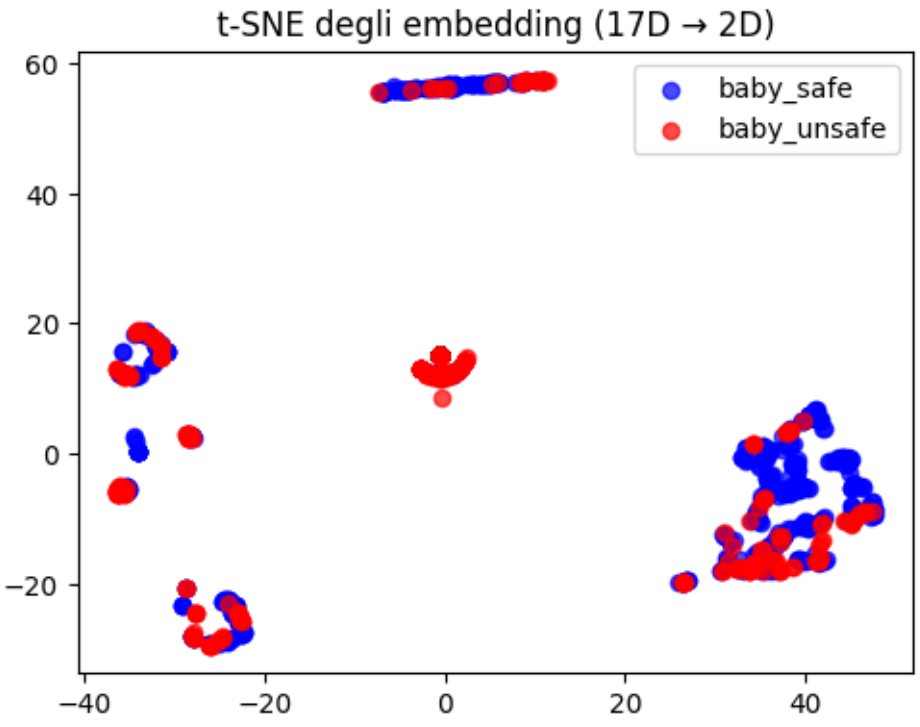
Recall at
R----------
0.6126016574337287

Silhouette
score----------

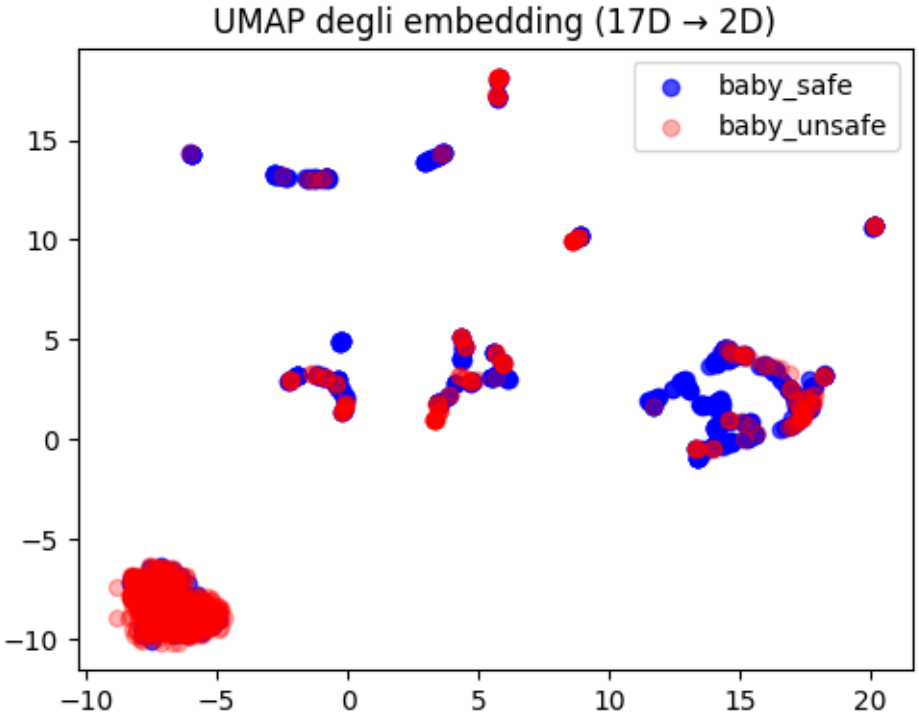


Embeddings
distributions----------


Distribuzione LDA degli embedding (17D → 1D)
baby_safe
baby_unsafe
600
500
400
300
200
100
0
−4
−3
−2
−1
0
1
2
LDA Component 1


t-SNE degli embedding (17D → 2D)
baby_safe
baby_unsafe
60
40
20
0
−20
−40
−20
0
20
40

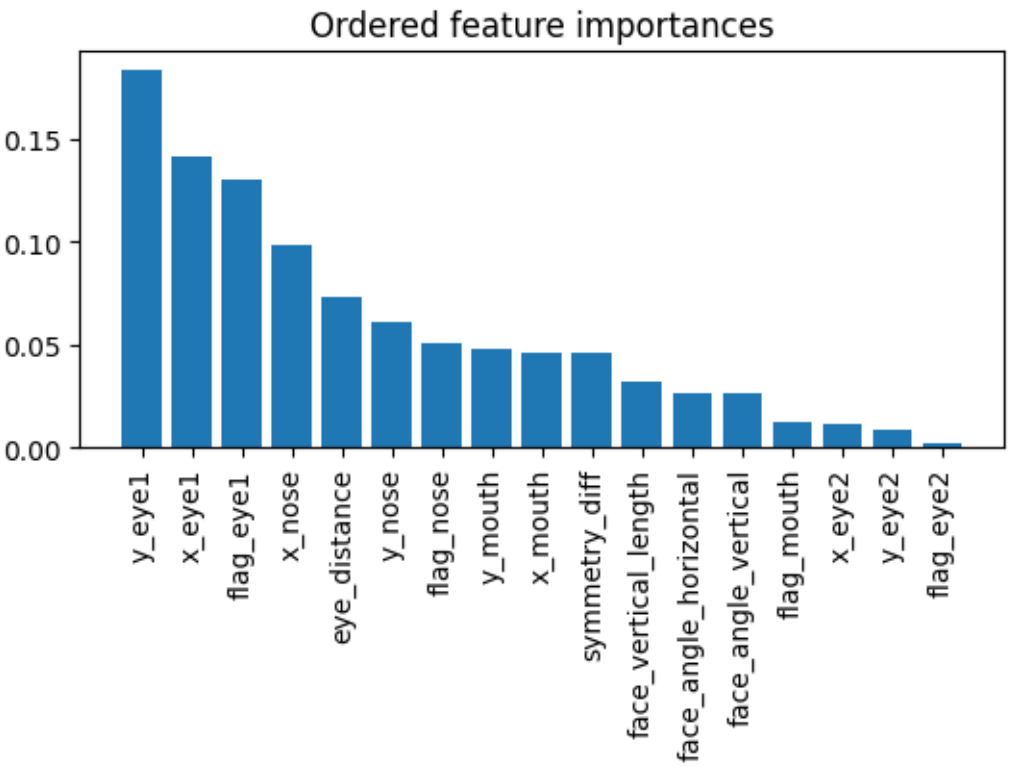
```
/home/terra/anaconda3/envs/SIDS_revelation_project/lib/python3.10/site-
packages/umap/umap_.py:1952: UserWarning: n_jobs value 1 overridden to 1 by
setting random_state. Use no seed for parallelism.
  warn(
/home/terra/anaconda3/envs/SIDS_revelation_project/lib/python3.10/site-
packages/sklearn/manifold/_spectral_embedding.py:328: UserWarning: Graph is not
fully connected, spectral embedding may not work as expected.
  warnings.warn(
```

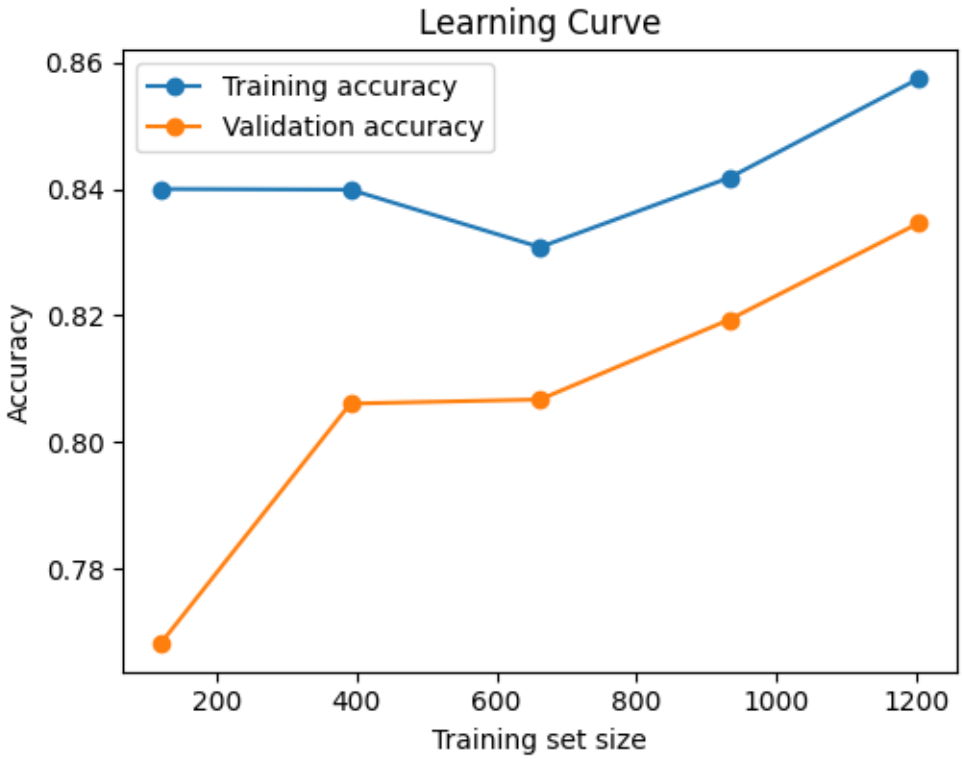


1.8 4. Train model classification

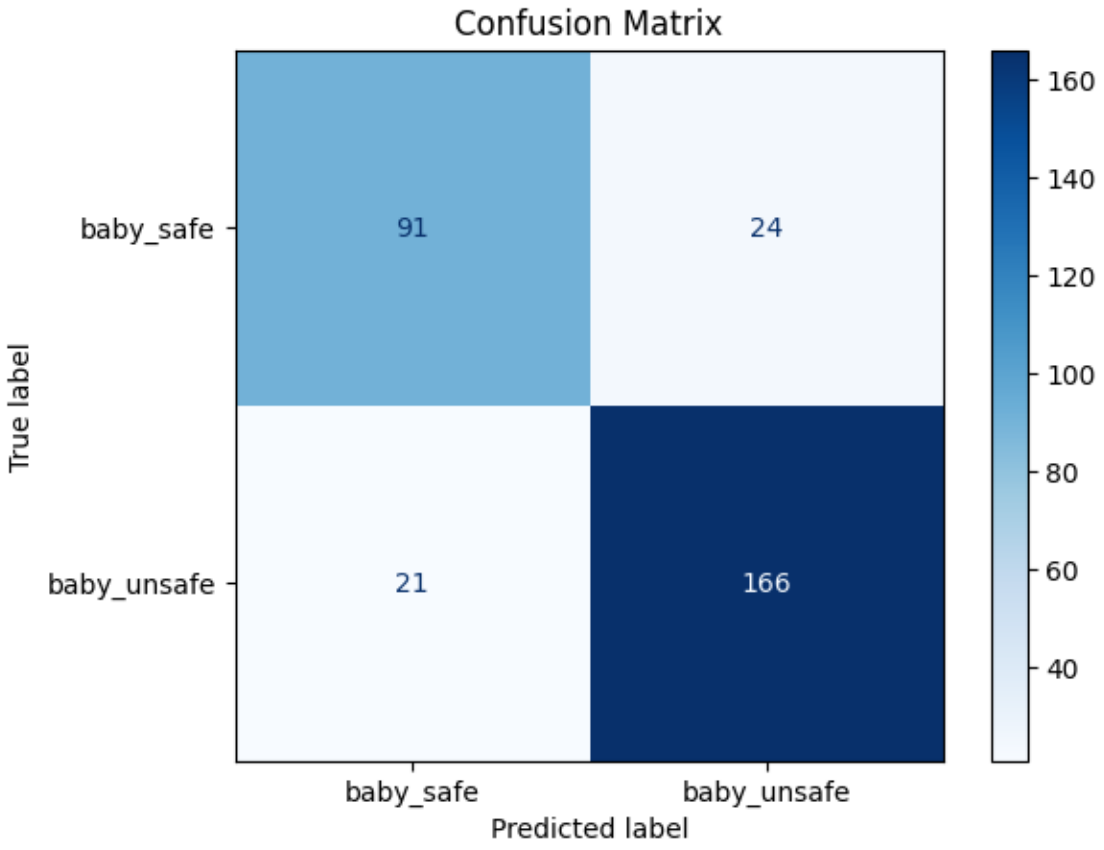
[14]:
```
clf = Classifier(embeddings, emb_builder.y, emb_builder.classes_bs)
clf.random_forest()
```

```
----------------------------------------------------------------------------------
----------
---------------------------------------FIRST
ANALYSIS---------------------------------------
----------------------------------------------------------------------------------
----------
```

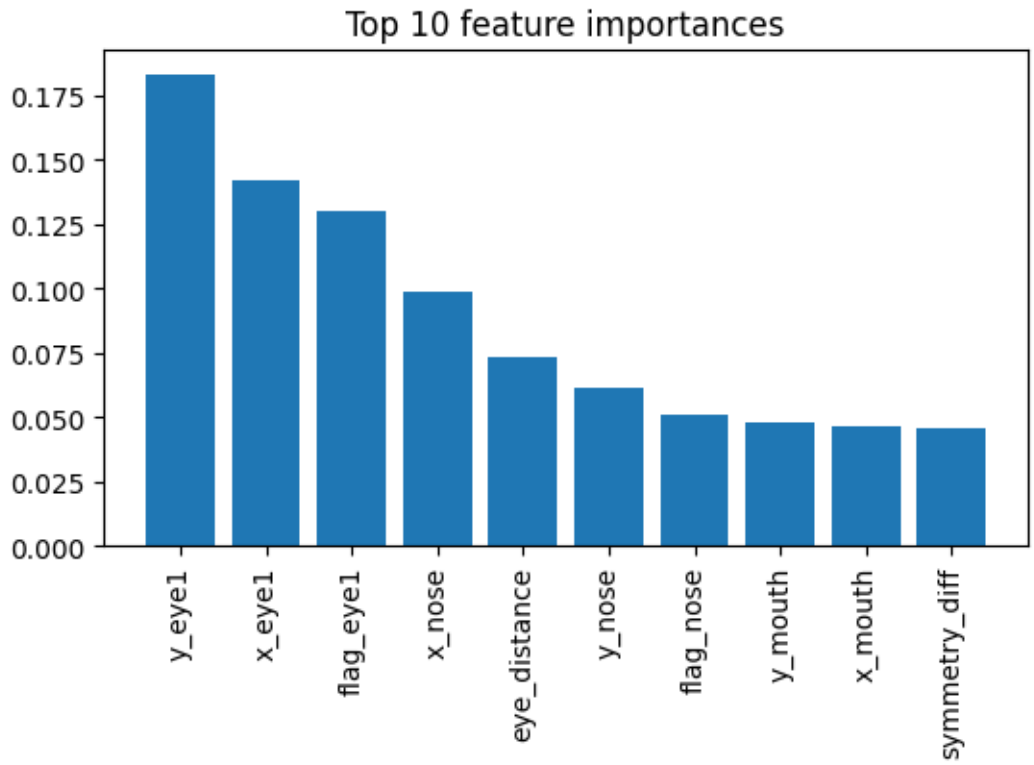

Ordered feature importances
0.15
0.10
0.05
0.00
y_eye1
x_eye1
flag_eye1
x_nose
eye_distance
y_nose
flag_nose
y_mouth
x_mouth
symmetry_diff
face_vertical_length
face_angle_horizontal
face_angle_vertical
flag_mouth
x_eye2
y_eye2
flag_eye2

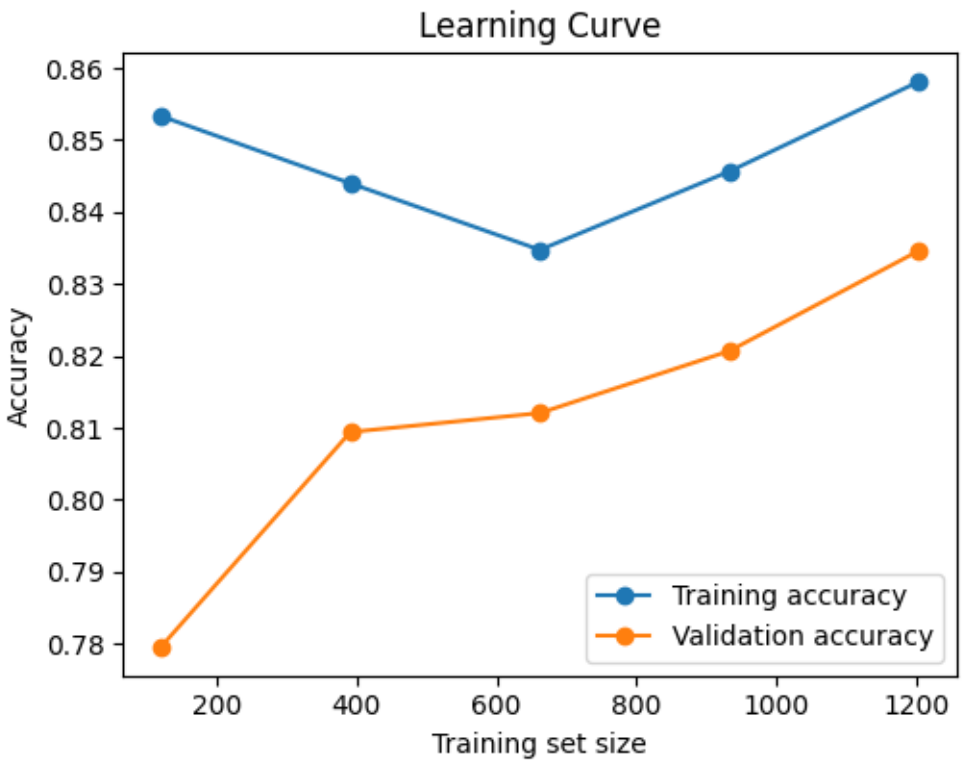

Learning Curve
Training accuracy
Validation accuracy
0.86
0.84
0.82
0.80
0.78
Accuracy
200
400
600
800
1000
1200
Training set size

```
Dataset labels:----------------------------------------
{'baby_safe': 0, 'baby_unsafe': 1}

Report-------------------------------------------------
              precision    recall  f1-score   support

   baby_safe       0.81      0.79      0.80       115
 baby_unsafe       0.87      0.89      0.88       187

    accuracy                           0.85       302
   macro avg       0.84      0.84      0.84       302
weighted avg       0.85      0.85      0.85       302

Confusion matrix---------------------------------------
```

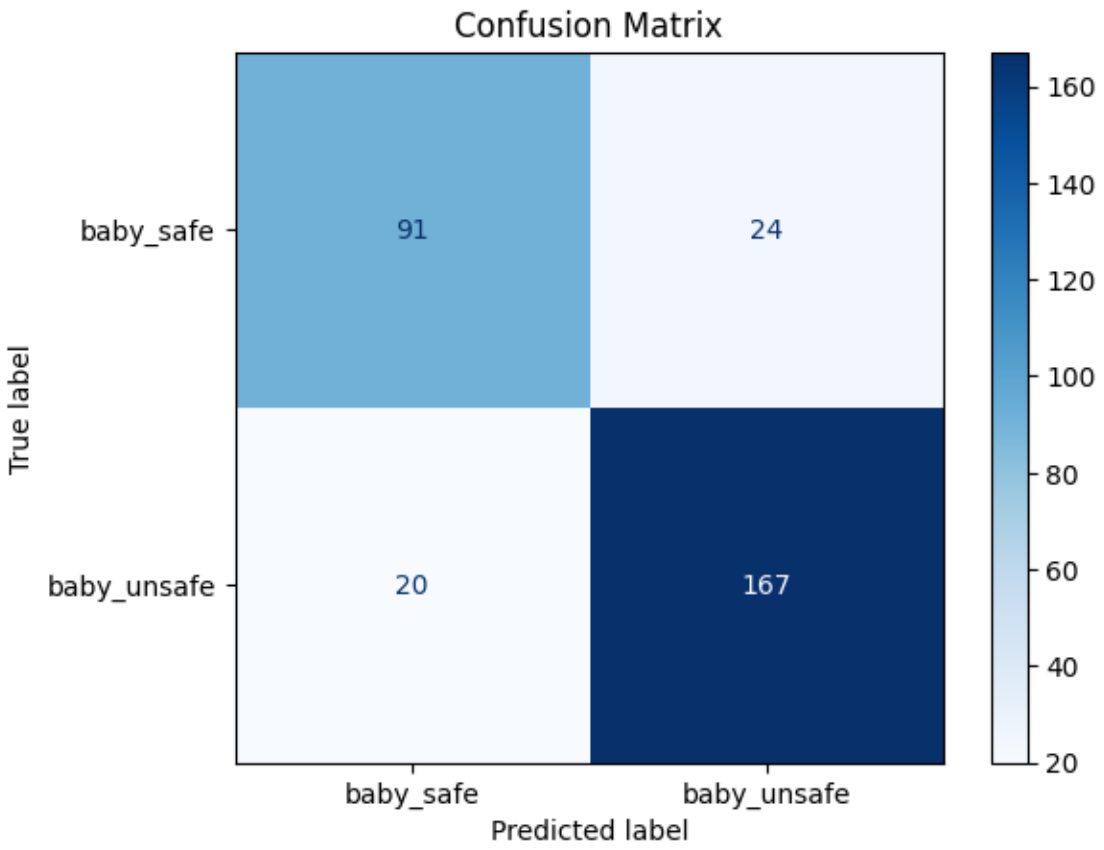


----------TOP 10 FEATURES ANALYSIS----------






```
Dataset labels:----------------------------------------
{'baby_safe': 0, 'baby_unsafe': 1}

Report-------------------------------------------------
              precision    recall  f1-score   support

   baby_safe       0.82      0.79      0.81       115
 baby_unsafe       0.87      0.89      0.88       187

    accuracy                           0.85       302
   macro avg       0.85      0.84      0.84       302
weighted avg       0.85      0.85      0.85       302

Confusion matrix---------------------------------------
```



1.9 2. Extract embeddings from dataset

Create embeddings

[15]:
```
embeddings = emb_builder.embedding_all_features_norm()
```

```
Creation of all features
embedding-----------------------------------------------------------
FINISHED: 1506 embedding created
```

[16]:
```
embeddings.head()
```

[16]:

	flag_eye1	flag_eye2	flag_nose	flag_mouth	x_eye1	y_eye1	x_eye2	\
0	0	0	0	0	-1.000000	-1.000000	-1.000000	
1	1	1	1	1	0.747866	0.955937	0.746319	
2	0	0	0	0	-1.000000	-1.000000	-1.000000	
3	1	1	1	1	0.533123	0.143157	0.374687	
4	1	1	1	1	0.859706	0.598094	0.846316	

```
     y_eye2    x_nose    y_nose  …  x_mouth_norm  y_mouth_norm  \
0 -1.000000 -1.000000 -1.000000  …     -7.346064     -3.560087
1  0.680452  0.709361  0.853581  …      0.814792      1.258686
2 -1.000000 -1.000000 -1.000000  …     -1.126908     -1.251334
3  0.148591  0.457184  0.163462  …      1.014402      1.536646
4  0.728694  0.827010  0.658031  …      0.911147      0.985550

   eye_distance  eye_distance_norm  face_vertical_length  \
0     -1.000000          -1.000000             -1.000000
1      0.275490           0.685025              0.060125
2     -1.000000          -1.000000             -1.000000
3      0.158529           0.397289              0.040059
4      0.131284           0.496643              0.039298

   face_vertical_length_norm  face_angle_vertical  face_angle_horizontal  \
0                  -1.000000            -1.000000              -1.000000
1                   0.090708           122.830696             147.334481
2                  -1.000000            -1.000000              -1.000000
3                   0.172300            87.692268             154.811363
4                   0.121073           104.018161             136.106213

   symmetry_diff  head_ration
0       0.000000     1.292278
1       0.085138     1.648186
2       0.000000     1.728926
3       0.004185     0.582652
4       0.002724     1.227875

[5 rows x 28 columns]
```

1.10 3. Retrieval to evaluate embedding goodness

[ ]:
```
ret = ImageRetrieval( embeddings, emb_builder.y, emb_builder.image_paths,␣
↪image_dataset_path, emb_builder.classes_bs)
ret.report("euclidean")
```

[ ]:
```
ret.report("cosine")
```

[ ]:
```
ret.report("minkowski")
```

1.11 4. Train model classification

[ ]:
```
clf = Classifier(embeddings, emb_builder.y, emb_builder.classes_bs)
clf.random_forest()
```

[ ]:
```
clf.XGBC()
```

[ ]:
```
clf.logistic_regression()
```

1.12 Save notebook

[ ]:
```
file_manager.save_as_pdf(ipynbname.path())
```