3.2.2. Data Preprocessing

```
df = df.withColumn('num', when(col('num') > 0, 1).otherwise(0))
df = df.withColumn('fbs', col('fbs').cast(StringType()))
df = df.withColumn('exang', col('exang').cast(StringType()))
```

```
df = df.drop('id', 'dataset')
df = df.drop('ca', 'thal')
```

df = df.withColumn('num', when(col('num') > 0, 1).otherwise(0))

df = df.withColumn('fbs', col('fbs').cast(StringType()))

df = df.withColumn('exang', col('exang').cast(StringType()))

df = df.drop('id', 'dataset')

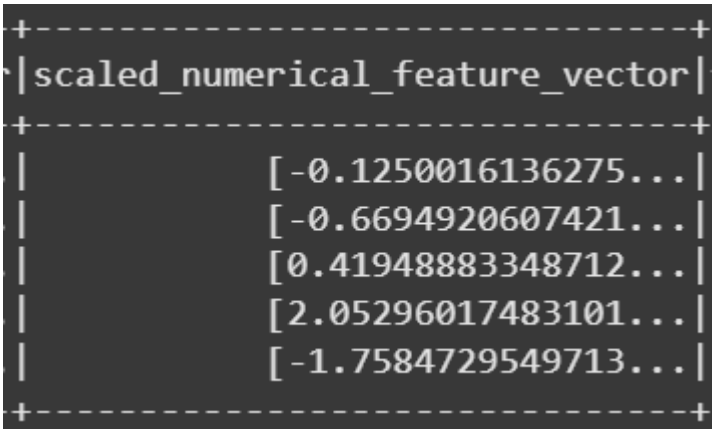
df = df.drop('ca', 'thal')

```
Imputer(
    inputCols=numerical_cols,
    outputCols=numerical_cols,
    strategy="mean"
)
```

Gambar 3.2.2.1: Preprocessing umum

```
+-----------------------+
|numerical_feature_vector|
+-----------------------+
|    [130.0,132.0,0.0,...|
|    [120.0,243.0,0.0,...|
|    [140.0,197.0,0.0,...|
|    [170.0,237.0,0.0,...|
|    [100.0,219.0,0.0,...|
+-----------------------+
```

Gambar 3.2.2.2: Gabungan Fitur Numerik



scaled_numerical_feature_vector
[-0.1250016136275...
[-0.6694920607421...
[0.41948883348712...
[2.05296017483101...
[-1.7584729549713...

Gambar 3.2.2.3: Standard Scaler fitur numerik

fbs_index	restecg_index	slope_index	exang_index	cp_index	sex_index
0.0	1.0	0.0	0.0	2.0	0.0
0.0	0.0	0.0	0.0	2.0	0.0
0.0	0.0	0.0	0.0	2.0	0.0
0.0	2.0	0.0	0.0	3.0	1.0
0.0	2.0	0.0	0.0	2.0	1.0

Gambar 3.2.2.4: String Indexing

fbs_one_hot	restecg_one_hot	slope_one_hot	exang_one_hot	cp_one_hot	sex_one_hot
(1,[0],[1.0])	(2,[1],[1.0])	(2,[0],[1.0])	(1,[0],[1.0])	(3,[2],[1.0])	(1,[0],[1.0])
(1,[0],[1.0])	(2,[0],[1.0])	(2,[0],[1.0])	(1,[0],[1.0])	(3,[2],[1.0])	(1,[0],[1.0])
(1,[0],[1.0])	(2,[0],[1.0])	(2,[0],[1.0])	(1,[0],[1.0])	(3,[2],[1.0])	(1,[0],[1.0])
(1,[0],[1.0])	(2,[],[])	(2,[0],[1.0])	(1,[0],[1.0])	(3,[],[])	(1,[],[])
(1,[0],[1.0])	(2,[],[])	(2,[0],[1.0])	(1,[0],[1.0])	(3,[2],[1.0])	(1,[],[])

Gambar 3.2.2.5: One Hot Encoding

categorical_feature_vector
[1.0,0.0,1.0,1.0,...
[1.0,1.0,0.0,1.0,...
[1.0,1.0,0.0,1.0,...
(10,[0,3,5],[1.0,...
(10,[0,3,5,8],[1....

Gambar 3.2.2.6: Gabungan fitur kategorikal

```
+--------------------+
|final_feature_vector|
+--------------------+
|[-2.7704572709341...|
|[-2.6626014617658...|
+--------------------+
```

Gambar 3.2.2.7: Gabungan semua fitur

3.2.3. Data Pipeline

```
print("Pipeline Stages:")
for i, stage in enumerate(svm_pipeline.getStages()):
    print(f"Stage {i}: {type(stage).__name__}")
```

```
Pipeline Stages:
Stage 0: Imputer
Stage 1: VectorAssembler
Stage 2: StandardScaler
Stage 3: StringIndexer
Stage 4: StringIndexer
Stage 5: StringIndexer
Stage 6: StringIndexer
Stage 7: StringIndexer
Stage 8: StringIndexer
Stage 9: OneHotEncoder
Stage 10: OneHotEncoder
Stage 11: OneHotEncoder
Stage 12: OneHotEncoder
Stage 13: OneHotEncoder
Stage 14: OneHotEncoder
Stage 15: VectorAssembler
Stage 16: VectorAssembler
Stage 17: LinearSVC
```

JELASKAN PIPELINE PADA SPARK BESERTA TRANSORMER DAN ESTIMATOR

Gambar 3.2.3.1: Pipeline PySpark