**No normalized data**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | | | | | |
| Model | 1ª | 2ª | 3ª | 4ª | 5ª | 6ª |
| Correlation | V4 | V11 | V12 | V10 | V16 | V3 |
| Kbest Select with F\_Classif Method | V14 | V4 | V12 | V11 | V10 | V3 |
| KBest with Mutual Information Classification Method | V14 | V3 | V17 | V10 | V12 | V4 |
| Feature importance | V14 | V10 | V16 | V3 | V12 | V4 |

Balance Dataset

**Feature Selected: V10, V12, V4, V3, V14, V11,** V16, V17

Imbalance Dataset

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | | | | | |
| Model | 1ª | 2ª | 3ª | 4ª | 5ª | 6ª |
| Correlation | V14 | V12 | V10 | V16 | V3 | V7 |
| Kbest Select with F\_Classif Method | V17 | V14 | V12 | V10 | V16 | V3 |
| KBest with Mutual Information Classification Method | V17 | V14 | V12 | V10 | V11 | V16 |
| Feature importance | V16 | V12 | V10 | V17 | V7 | V11 |

Imbalance Dataset

**Feature Selected: V16, V12, V10, V17, V14, V3**, V7, V11

Imbalance Dataset

**Features selected: V16, V12, V17, V14, V10, V3**

Balance Dataset

**Features selected: V10, V12, V16, V14, V4, V11**

**# df1\_new dataset to be used**

**from matplotlib.ticker import FuncFormatter**

**from pyspark.ml import Pipeline, PipelineModel**

**from pyspark.ml.classification import LogisticRegression, DecisionTreeClassifier, RandomForestClassifier, GBTClassifier, LinearSVC, NaiveBayes**

**from pyspark.ml.evaluation import MulticlassClassificationEvaluator, BinaryClassificationEvaluator**

**from pyspark.ml.feature import StandardScaler**

**from pyspark.ml.feature import StringIndexer, VectorIndexer, IndexToString, VectorAssembler**

**from pyspark.sql import SparkSession**

**from scipy import stats**

**from scipy.stats import norm**

**from sklearn.linear\_model import LogisticRegression**

**from sklearn.metrics import accuracy\_score,f1\_score,precision\_score, recall\_score**

**from sklearn.model\_selection import cross\_val\_score**

**from sklearn.model\_selection import train\_test\_split**

**from sklearn.naive\_bayes import GaussianNB**

**from sklearn.neighbors import KNeighborsClassifier**

**from sklearn.preprocessing import LabelEncoder**

**from sklearn.tree import DecisionTreeClassifier**

**import matplotlib.pyplot as plt**

**import numpy as np**

**import pandas as pd**

**import pyspark.sql.functions as f**

**import pyspark.sql.types as t**

**import seaborn as sns**