Лабораторная работа №3. Визуализация зависимостей целевой переменной от входных признаков.

Используемый набор данных: banknote authentication (https://archive.ics.uci.edu/ml/datasets/banknote+authentication)

In [1]:

```
import pandas as pd
import plotly.express as px
import os
import requests

%matplotlib inline

pd.options.display.max_columns = None
```

Загрузим файл с набором данных, если он отсутствует.

In [2]:

```
def downloadFile(url, filePath):
    if not os.path.exists(filePath):
        req = requests.get(url)
        f = open(filePath, "wb")
        f.write(req.content)
        f.close

url = "https://archive.ics.uci.edu/ml/machine-learning-databases/00267/data_banknote_authentication.txt"
fileName = "dataset/data_banknote_authentication.txt"
downloadFile(url, fileName)
```

Опишем заголовки колонок. Для удобства используем сокращенные названия:

- **Variance** - Variance of Wavelet Transformed image.
- **Skewness** - Skewness of Wavelet Transformed image.
- **Curtosis** - Curtosis of Wavelet Transformed image.
- **Entropy** - Entropy of image.

In [3]:

```
headers = ["Variance", "Skewness", "Curtosis", "Entropy", "Class"]
data = pd.read_csv(fileName, names=headers)
# Выполним преобразование типа категориальной колонки (Class)
data["Class"] = data["Class"].astype("category")
data.sample(40)
```

Out[3]:

	Variance	Skewness	Curtosis	Entropy	Class
1047	1.063700	3.695700	-4.159400	-1.937900	1
273	2.694600	6.797600	-0.403010	0.449120	0
420	-2.484000	12.161100	2.820400	-3.741800	0
710	2.401200	1.622300	3.031200	0.716790	0
1157	-5.204900	7.259000	0.070827	-7.300400	1
517	1.762000	4.368200	2.138400	0.754290	0
302	1.684900	8.748900	-1.264100	-1.385800	0
785	-1.666200	-0.300050	1.423800	0.024986	1
42	-0.006892	9.293100	-0.412430	-1.963800	0
135	4.160500	11.219600	-3.613600	-4.081900	0
134	-1.040100	9.398700	0.859980	-5.333600	0
1160	-1.595100	-6.572000	4.768900	-0.943540	1
751	2.254600	8.099200	-0.248770	-3.269800	0
842	-1.896900	-6.789300	5.276100	-0.325440	1
507	4.601400	5.626400	-2.123500	0.193090	0
632	3.694100	-3.948200	4.262500	1.157700	0
43	0.964410	5.839500	2.323500	0.066365	0
994	-0.873400	1.653300	-2.196400	-0.780610	1
1364	-2.839100	-6.630000	10.484900	-0.421130	1
1319	0.663650	-0.045533	-0.187940	0.234470	1
469	0.184800	6.507900	2.013300	-0.872420	0
717	2.985600	7.267300	-0.409000	-2.243100	0
937	-2.902000	-7.656300	11.831800	-0.842680	1
787	-2.668500	-10.451900	9.113900	-1.732300	1
18	1.447900	-4.879400	8.342800	-2.108600	0
942	-3.379300	-13.773100	17.927400	-2.032300	1
483	0.967880	7.190700	1.279800	-2.456500	0
203	4.175700	10.261500	-3.855200	-4.305600	0
1247	-4.477500	-13.030300	17.083400	-3.034500	1
1332	0.904070	3.370800	-4.498700	-3.696500	1
440	3.435900	0.662160	2.104100	1.892200	0
1223	1.340300	4.132300	-4.701800	-2.598700	1
630	2.598900	3.517800	0.762300	0.811190	0
656	-1.361200	10.694000	1.702200	-2.902600	0
1235	-3.535900	0.304170	0.656900	-0.295700	1
7	2.092200	-6.810000	8.463600	-0.602160	0
449	3.941400	-3.290200	3.167400	1.086600	0

	Variance	Skewness	Curtosis	Entropy	Class
320	0.519500	-3.263300	3.089500	-0.984900	0
464	5.740300	-0.442840	0.380150	1.376300	0
1233	-7.042100	9.200000	0.259330	-4.683200	1

Для визуализации зависимостей между входными признаками и целевой переменной используем диаграммы рассеяния. Исключим из матрицы те диаграммы, которые отображают корреляцию одной и той же величины.

In [4]:

```
plots = px.scatter_matrix(data, dimensions=headers[:len(headers) - 1], color="Class")
plots.update_traces(diagonal_visible=False)
plots.show()
```