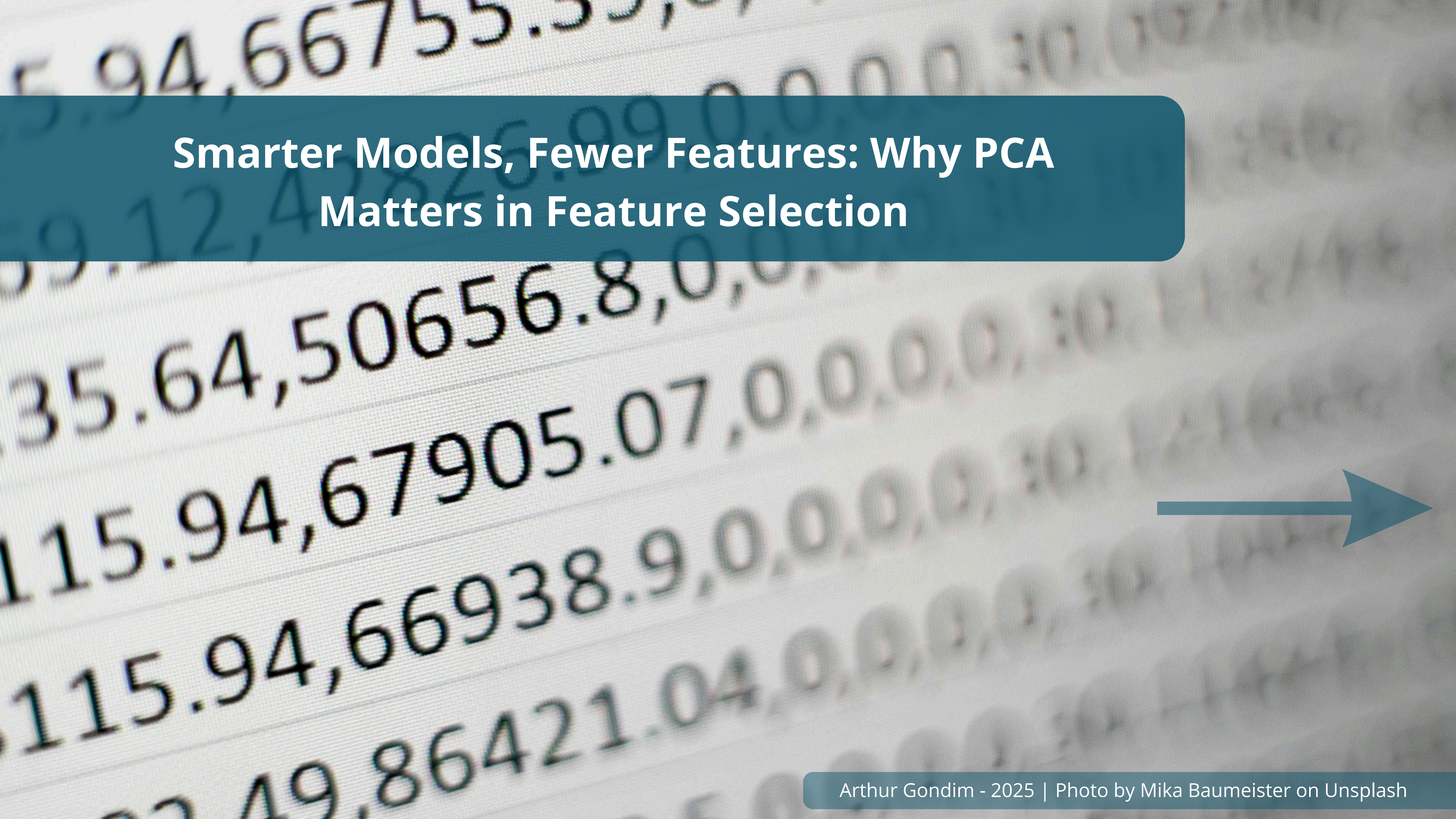
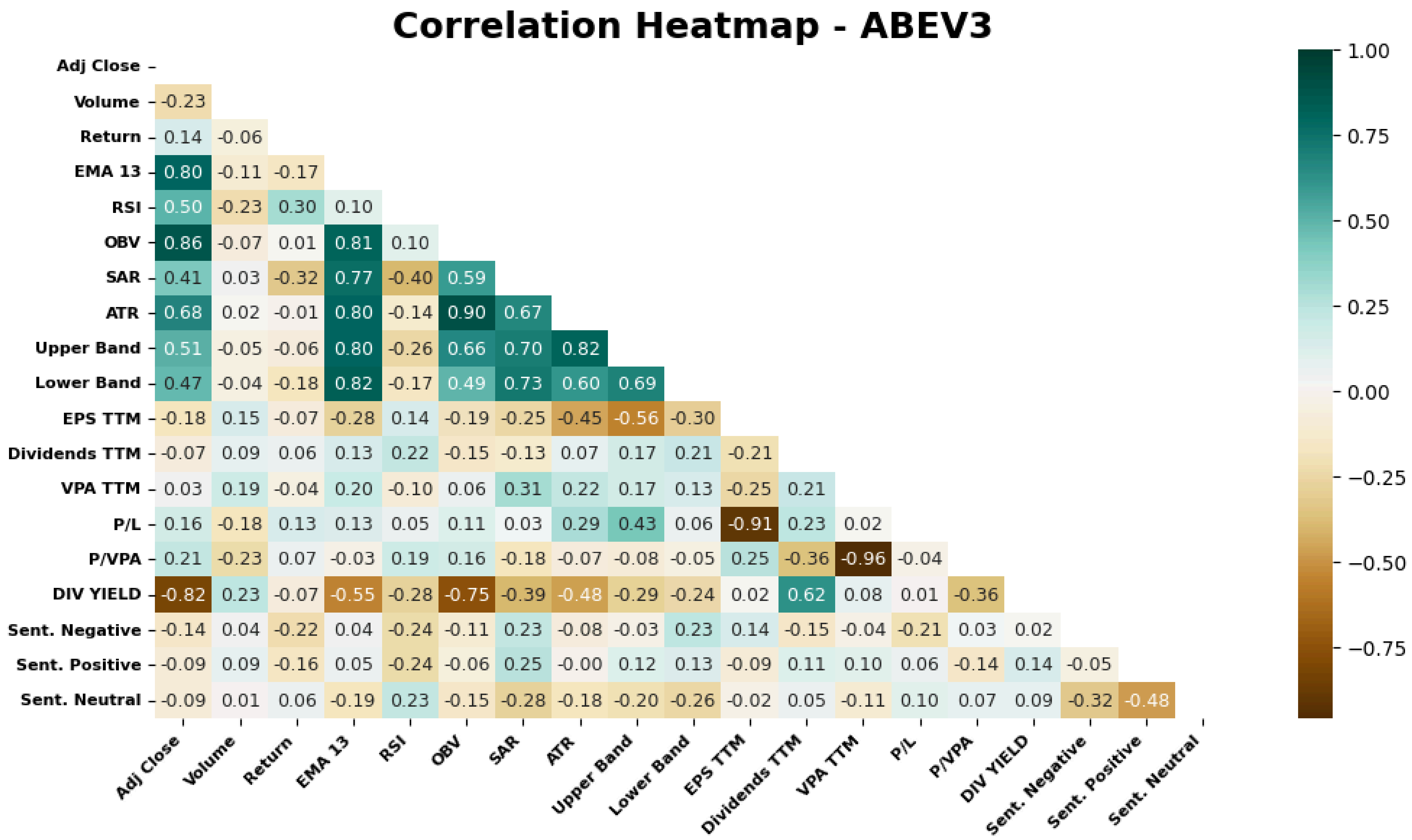
Smarter Models, Fewer Features: Why PCA Matters in Feature Selection



Arthur Gondim - 2025 | Photo by Mika Baumeister on Unsplash

Dataset: ABEV3 - Brazilian Market



Correlation Heatmap - ABEV3

	Adj Close	Volume	Return	EMA 13	RSI	OBV	SAR	ATR	Upper Band	Lower Band	EPS TTM	Dividends TTM	VPA TTM	P/L	P/VPA	DIV YIELD	Sent. Negative	Sent. Positive	Sent. Neutral
Adj Close																			
Volume	-0.23																		
Return	0.14	-0.06																	
EMA 13	0.80	-0.11	-0.17																
RSI	0.50	-0.23	0.30	0.10															
OBV	0.86	-0.07	0.01	0.81	0.10														
SAR	0.41	0.03	-0.32	0.77	-0.40	0.59													
ATR	0.68	0.02	-0.01	0.80	-0.14	0.90	0.67												
Upper Band	0.51	-0.05	-0.06	0.80	-0.26	0.66	0.70	0.82											
Lower Band	0.47	-0.04	-0.18	0.82	-0.17	0.49	0.73	0.60	0.69										
EPS TTM	-0.18	0.15	-0.07	-0.28	0.14	-0.19	-0.25	-0.45	-0.56	-0.30									
Dividends TTM	-0.07	0.09	0.06	0.13	0.22	-0.15	-0.13	0.07	0.17	0.21	-0.21								
VPA TTM	0.03	0.19	-0.04	0.20	-0.10	0.06	0.31	0.22	0.17	0.13	-0.25	0.21							
P/L	0.16	-0.18	0.13	0.13	0.05	0.11	0.03	0.29	0.43	0.06	-0.91	0.23	0.02						
P/VPA	0.21	-0.23	0.07	-0.03	0.19	0.16	-0.18	-0.07	-0.08	-0.05	0.25	-0.36	-0.96	-0.04					
DIV YIELD	-0.82	0.23	-0.07	-0.55	-0.28	-0.75	-0.39	-0.48	-0.29	-0.24	0.02	0.62	0.08	0.01	-0.36				
Sent. Negative	-0.14	0.04	-0.22	0.04	-0.24	-0.11	0.23	-0.08	-0.03	0.23	0.14	-0.15	-0.04	-0.21	0.03	0.02			
Sent. Positive	-0.09	0.09	-0.16	0.05	-0.24	-0.06	0.25	-0.00	0.12	0.13	-0.09	0.11	0.10	0.06	-0.14	0.14	-0.05		
Sent. Neutral	-0.09	0.01	0.06	-0.19	0.23	-0.15	-0.28	-0.18	-0.20	-0.26	-0.02	0.05	-0.11	0.10	0.07	0.09	-0.32	-0.48	

Example 1: Raw Data Classifier

‘Raw’ data as input:

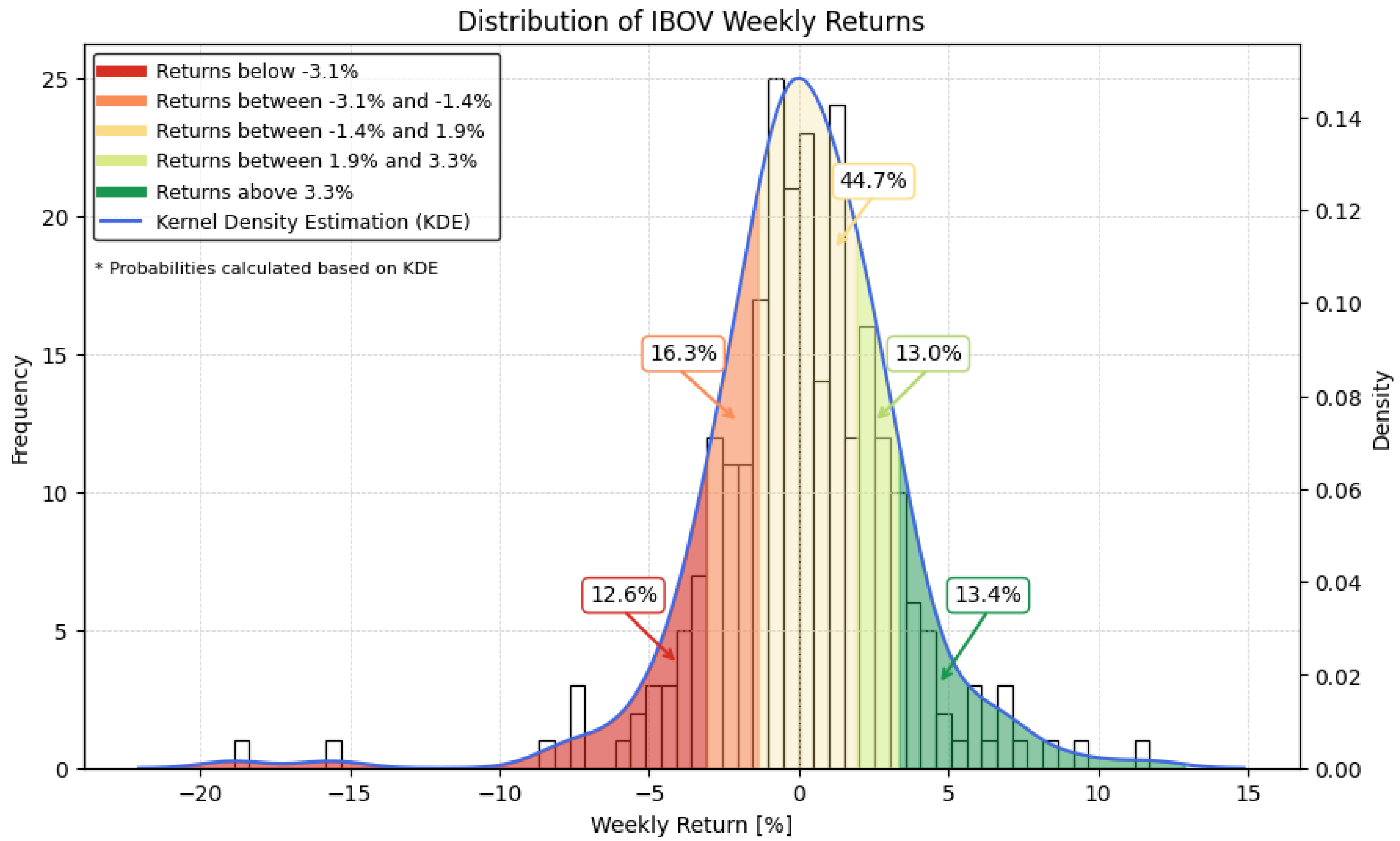
- Total of 19 features.
- Dataset combines technical, fundamental, and sentiment analysis data from the ABEV3 stock.

Random Forest algorithm:

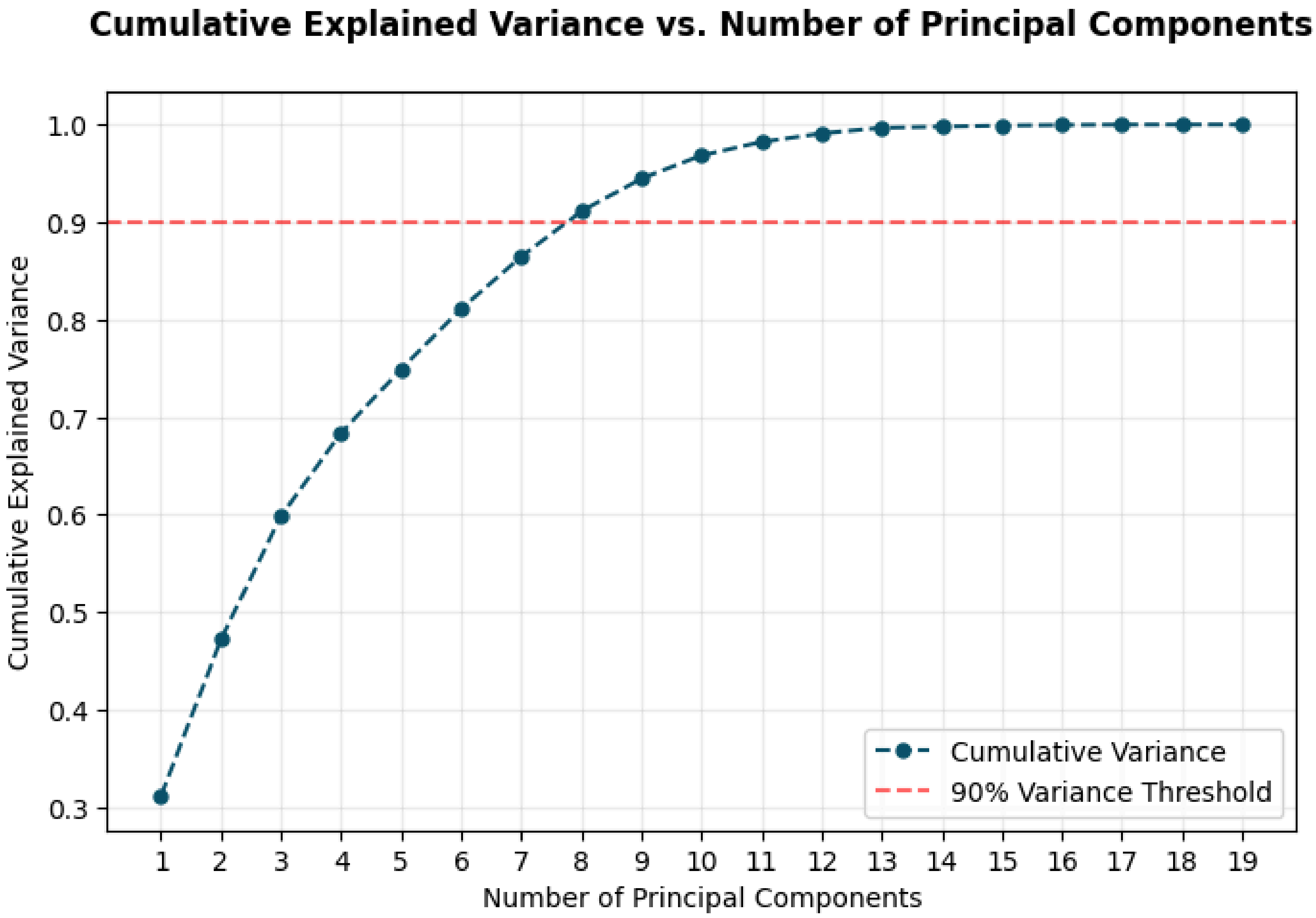
- Since the dataset is small, parameters are set to **stress-test**, simulating computational intensity.

Output

- (Encoded) categorical variable representing the next week's return based on histogram-defined regions.



Principal Component Analysis (PCA)



- The **cumulative explained variance** is the **sum** of the variance proportions explained by each principal component (PC).

- It indicates how much information is retained in **lower-dimensional space**.

- Helps decide the **optimal number** of principal components to retain.

Example 2: Principal Components Classifier

8 Principal Components as input:

- ~90% of the variance.

Random Forest algorithm

- Same parameters as before.

	Overall Results	
	RAW inputs	PCA
Training time	6.38 s	4.42 s
Overall Accuracy	23.8 %	33.3 %