**Full Gen Data**

* **EDA:**

**Here is a list of the important attributes:**

**• country: Country name**

**• article: 6-digit article number, as unique identifier of an article**

**• sales: total number of units sold in respective retail week**

**• regular\_price: recommended retail price of the article**

**• current\_price: current selling price (weighted average over the week)**

**• ratio: price ratio as current\_price/regular\_price, such that price discount is 1-ratio**

**• retailweek: start date of the retailweek**

**• promo1: indicator for media advertisement, taking 1 in weeks of activation and 0 otherwise**

**• promo2: indicator for store events, taking 1 in weeks with events and 0 otherwise**

**• customer\_id: customer unique identifier, one id per customer**

**• productgroup: product group the article belongs to**

**• category: product category the article belongs to**

**• cost: total costs of the article (assumed to be fixed over time)**

**• style: description of article design**

**• sizes: size range in which article is available**

**• gender: gender of target consumer of the article**

**• rgb\_\*\_main\_color: intensity of the red (r), green (g), and blue (b) primaries of the article ‘s main color, taking values [0,250]**

**• rgb\_\*\_sec\_color: intensity of the red (r), green (g), and blue (b) primaries of the article ‘s secondary color, taking values [0,250]**

**• label: advertisement result after offering/sending/presenting the offer to the customer. 0 means the customer did not buy and 1 means the customer did buy.**

**This data doesn’t have any null values.**

**Some Insights:**

* **Visualize the distribution of numerical variables.**

**A group of blue and white graphs

Description automatically generated**

* **Dominant gender and customers did buy Women**A green circle with blue and orange triangles

  Description automatically generated
* Dominant article **BR3179.**
* Dominant article in sales **MA7179.**
* Dominant retailweek and customers did buy **2017-01-22**.
* Dominant category and customers did buy **Training**.
* Dominant Product and customers did buy **Shoes**.
* Dominant Style and customers did buy **Regular**.
* the most gender affected by the color is **women** then **kids** then **men.**
* Count plot of labelA blue and orange bar graph

  Description automatically generated

**Data Preprocessing:**

1. **Apply Over Sampling (SMOTEN)**

* SMOTEN is Synthetic Minority Over-sampling Technique for Nominal.

1. **Split Data**

* 80% train – 20% test

1. **Apply Normalization (StandardScaler)**

**Data Modeling:**

Random Forest Classifier

* Test Accuracy: 91.18064516129031%