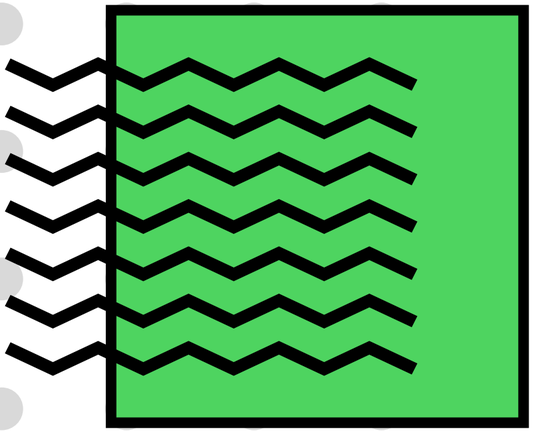


SKIN CONDITION CLASSIFICATION HACATHON

GROUP 3

TEAM

NIKHIL GAIKWAD

ABDUL HADI

ADETUTU ADEBAYO

AMIT JOSHI

TABLE OF CONTENT

INTRODUCTION

OBJECTIVE

Develop a machine learning model to classify skin conditions based on beauty product descriptions.

GOAL

Automatically classify skin conditions from product descriptions to improve product recommendations and customer satisfaction.

CONTEXT

- Skin conditions (e.g., oily, dry, acne-prone) require tailored skincare solutions.
- Beauty product descriptions contain valuable information about their intended use for specific skin conditions.

DATA

6,241 unlabelled data
33 Attributes

PIPELINE

01

AWS FOR PROMPTING

02

LABEL DATA USING LLM & PROMPTS

03

TRAIN CLASSIFICATION MODEL

04

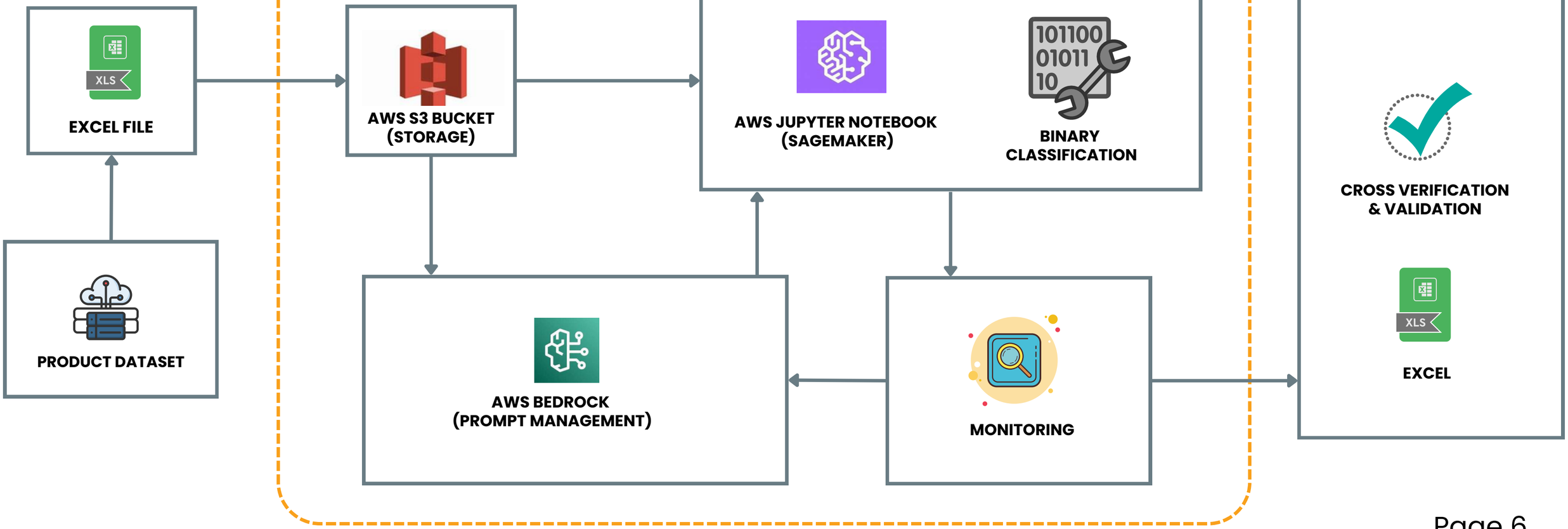
EVALUATE CARBON EMISSIONS

DATA LABELING

INGEST

EXTRACT

ENRICH



CODE & PROMPT FLOW

TOP LLM PICKS

We tested multiple LLMs and found the best performers:

Claude 3.5 Sonnet
Llama 3.1 70B Instruct
Claude 3 Haiku
Nova Pro

PROMPT TEMPLATE

- **Insert Description**: The prompt takes a product description as input.
- **Extract Attributes:** It directs the model to identify specific attributes using keywords and synonyms.
- **Binary Classification:** For each attribute, it assigns a "1" if present, or "0" if absent.
- **Formatted Output**: The model returns a comma-separated list of binary values in a fixed order.

```
Extract relevant attributes from the product description using synonyms and context.
Assign **1 (present) or 0 (absent)** for each attribute.

**Example Attributes:**
- **Skin Concerns**: Dark pigmentation, acne, wrinkles, etc.
- **Age Groups**: 18-34, 35-54, 55-99
- **Skin Types**: Dry, oily, combination
- **Sensitivity Levels**: High, low, none
- **Functions**: Cleanse, treat, moisturize, protect, etc.

**Guidelines:**
- **Contextual Mapping**: Identify attributes using synonyms (e.g., "hyperpigmentation" →
  "dark pigmentation").
- **Ingredient-Based Recognition**: Match skincare ingredients to concerns (e.g., "Retinol" →
  "Wrinkles").
- **Categorization**: Classify based on age, skin type, and sensitivity.

**Output Format:**
A comma-separated list of **1s and 0s**, following the attribute order.
_No extra explanations._
```

	A	B	C	D	E	F	G	H	I
1	text_raw	_pigmenta	acne	ye_contou	omogeneit	ck_firmne	ck_radianc	pores	fine_lines
2	"This advanced skincare solution is	1	1	1	1	1	1	1	1
3	"This innovative formula enhances	1	1	1	1	1	1	1	1
4	"This product targets uneven skin t	1	0	0	1	1	0	0	1
5	"This treatment helps reduce dark	1	1	0	1	0	0	0	0
6	"This moisturizer is formulated for	0	0	0	0	0	0	0	0
7	"This skincare range is designed to	0	0	0	0	0	0	0	0
8	"This is a high-quality skincare proc	0	0	0	0	0	0	0	0
9	"A luxurious beauty product formu	0	0	0	0	0	0	0	0
10	"A versatile skincare product that f	0	0	0	0	0	0	0	0
11	Australian Gold Sunscreen Spray G	0	0	0	0	0	0	0	0
12	Australian Gold Sunscreen Spray G	0	0	0	0	0	0	0	0
13	Aveeno Positively Radiant Daily Fa	1	0	0	1	0	1	0	0
14	Aveeno Positively Radiant Daily Fa	1	0	0	1	0	1	0	0

CROSS VERIFICATION VALIDATION &

Since manual verification of 6000+ product descriptions was impractical, we implemented an automated validation approach

STRUCTURED TEST DATASET

- ✅ **All attributes** included – tests complete extraction.
- ✅ No attributes (with **synonyms/chemical** terms) – checks false positives.
- ✅ **Partial** attributes – validates partial detection.
- ✅ Only "**Age**" – ensures specific extraction.
- ✅ **Duplicate text** – confirms consistent processing.

CLASSIFICATION STEPS:

IMPORT LIBRARIES

- **Import** necessary libraries like pandas, sklearn, LightGBM, nltk, and codecarbon.

LOAD DATA

- Load the **dataset** using pd.read_excel.

FEATURE ENGINEERING

- **Text Cleaning:** Remove special characters, stopwords, and apply lemmatization to improve data quality.
- **Vectorization:** Use TfidfVectorizer with character n-grams to convert text to numerical features.

TRAIN MODEL

- Train an ensemble model
- Tune **Hyperparameters**: Adjust parameters like max_depth, n_estimators, learning_rate

EVALUATE

- Evaluate using **accuracy, F1 score,** recall, precision.
- Track Carbon Emissions: Measure CO2 emissions using the codecarbon library during training

OPTIMIZE

- Optimize for performance and sustainability by **tuning** hyperparameters and reducing CO2 emissions.

ALL MODELS

Model	Accuracy	F1-score	Carbon Emission (kg CO2)
RoBERTa + LightGBM	0.90	0.84	0.0069
LightGBM	0.88	0.84	0.0005
SVM	0.87	0.82	0.0151
XGBoost	0.89	0.82	0.0020
Gradient Boosting	0.88	0.82	0.0128
Random Forest	0.79	0.76	0.0001
RoBerta + XGBoost	0.90	0.83	0.0069

LOW CARBON EMMISON

HIGH ACCURCY MODEL

LOGISTIC REGRESSION

Accuracy - 0.875704
F1-score - 0.82
Precision - 0.81
Carbon Emission - 0.0001 kg Co2

ROBERTA

Accuracy - 0.900325
F1-score - 0.85
Precision - 0.84
Carbon Emission - 0.0060 kg CO2

RECOMMENDED MODEL

LOGISTIC REGRESSION + LIGHTGBM

CARBON EMISSION

0.0002 kg CO2

ACCURACY

0.881459

F1

0.84

This model optimizes max_depth, learning_rate, and subsample for better generalization. It uses sparse matrices to reduce memory overhead and n_jobs=-1 for parallel processing. The model achieves high accuracy with lower emissions compared to others.

```
29                moisturize  0.892628  0.917454  0.927741   0.907392
30                   protect  0.883013  0.817043  0.806931   0.827411
31                       day  0.830128  0.894980  0.931271   0.861411
32                     night  0.833333  0.863517  0.920709   0.813015
Average              Average  0.881459  0.820999  0.875522   0.778089

Micro Average:
Precision: 0.79, Recall: 0.89, F1: 0.83

Macro Average:
Precision: 0.78, Recall: 0.88, F1: 0.82

Weighted Average:
Precision: 0.80, Recall: 0.89, F1: 0.84

Samples Average:
Precision: 0.78, Recall: 0.89, F1: 0.82

Total Carbon Emissions: 0.0002 kg CO2
```

RECOMMENDATION

1. **If L'Oréal prioritizes sustainability:** Current model is a strong candidate due to low emissions.
2. **For best of both worlds:** Hybrid approach with efficient resource allocation and targeted optimizations.
3. **Accuracy Optimization**

- Hyperparameter tuning
- Model complexity
- Data augmentation
- Trade-off
- Transfer learning
- Efficiency measures

CONCLUSION

- L'Oréal has a range of machine learning models to choose from, depending on their priorities.
- If the primary focus is on accuracy, RoBERTa + LightGBM and XGBoost are the top performers, with RoBERTa + LightGBM offering the highest accuracy and F1-score.
- If sustainability is a key consideration, Logistic Regression + LightGBM, LightGBM provides an excellent balance of accuracy and low carbon emissions.

For long- Term success, L'Oréal should consider a hybrid approach, leveraging high-accuracy models for critical tasks while gradually integrating more sustainable models into their workflow

THANK YOU