Projet Capstone

Rapport phase 1

# Description

Projet : **Détection de mélanome basée sur une analyse d’images de la peau**

**Pourquoi : Problème de prédiction chez les dermatologues**

**But : Pour aider les dermatologues.**

**Symptômes du mélanome :** **Signes à surveiller**

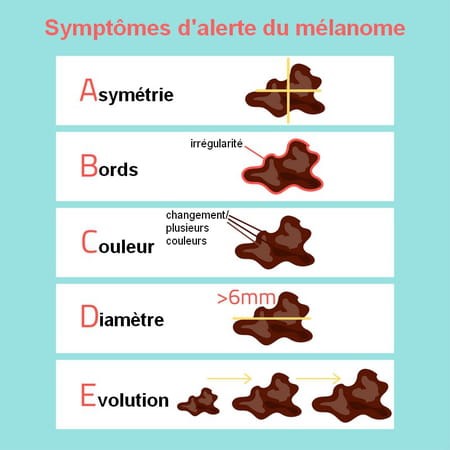
****

Table of Contents

[Description 1](#_Toc137805192)

[Table des matières 1](#_Toc137805193)

[A. Data Engineering Pipelines 1](#_Toc137805194)

[1. Collecte des données 1](#_Toc137805195)

[2. Prétraitement des données 5](#_Toc137805196)

[a. Prétraitement des données numériques 6](#_Toc137805197)

[b. Prétraitement des images 8](#_Toc137805198)

[B. Model pipelines 14](#_Toc137805199)

[1. Machine Learning 14](#_Toc137805200)

[2. AutoKeras 14](#_Toc137805201)

[3. Deep MachineLearning 18](#_Toc137805202)

[Présentation des travaux futurs 19](#_Toc137805203)

# Data Engineering Pipelines

## Collecte des données

**PLANNING DU TRAVAIL POUR LA SEMAINE (1) 8 MAI à 12 MAI**

|  |  |  |
| --- | --- | --- |
| **Les Tâches à faire : collecte des données** | | |
| **Evelyne** | **Hassan** | **EL Hadji** |
| **Collecte des données** | **Collecte des données** | **Collecte des données** |
| **Recherche :**  Des méthodes pour extraire les caractéristiques ABCDE du mélanome à partir des images | **Recherche :**  Des méthodes pour extraire les caractéristiques ABCDE du mélanome à partir des images | **Rechercher :**  D’autres méthodes pour récupérer les données |

**Bilan de la semaine 1 :**

**Source**: <https://www.kaggle.com/competitions/siim-isic-melanoma-classification/data>

* Durant toute la semaine, on a essayé de récupérer toutes les images de notre dataset à partir de kaggle.
* **Difficultés rencontrées lors de la collecte des données :**
* Taille importante du **dataset(116 Go).**
* Notre Google Colab initial n'avait pas suffisamment d'espace de stockage pour héberger ces données massives.
* ***Pour avancer :***
* ***Face*** à ces difficultés, nous avons cherché des solutions pour continuer à progresser dans notre projet.
* Notre objectif principal était **d'extraire les caractéristiques ABCDE du mélanome** à partir des images, en nous concentrant notamment sur le traitement des images pour éliminer les poils et autres artefacts.

**PLANNING DU TRAVAIL POUR LA SEMAINE (2) 15 MAI à 19 MAI**

* Heureusement, nous avons reçu une aide précieuse du professeur jean Gabriel pour surmonter les défis. Il nous a aidés en transférant et en stockant le dataset sur son propre Google Colab, qui disposait de suffisamment d'espace de stockage pour accueillir ces données massives

|  |  |  |
| --- | --- | --- |
| **Les Tâches à faire : collecte des données et Étude des données** | | |
| **Evelyne** | **Hassan** | **EL Hadji** |
| Récupérer les données  Comprendre les données  Continuer les recherches sur l’extraction (ABCDE) | Récupérer les données  Comprendre les données  Continuer les recherches sur l’extraction (ABCDE).  Préparer le compte rendu | Récupérer les données  Comprendre les données  Continuer les recherches sur l’extraction (ABCDE). |

**Bilan de la semaine :**

|  |  |
| --- | --- |
| **Les tâches** | Remarque |
| **Récupérer les données**  **Comprendre les données**  **\*Numérique** | Différences entre le nombre réel d'images et le nombre de références d'images présentes dans le dataset.  **Une image contenant texte, Police, capture d’écran  Description générée automatiquement**  Correction de la différence :  Une image contenant texte, capture d’écran, Police, affichage  Description générée automatiquement  Une image contenant texte, reçu, Police, blanc  Description générée automatiquement |
| Certains patients possèdent plusieurs images, tandis que d'autres en possèdent moins.  **Une image contenant texte, Police, capture d’écran  Description générée automatiquement** |
| **\*image** |  |

## Prétraitement des données

**PLANNING DU TRAVAIL POUR LA SEMAINE (3)** **22 MAI à 26 MAI**

#### 

|  |  |  |
| --- | --- | --- |
| **Les Tâches à faire : prétraitement des données** | | |
| **Evelyne** | **Hassan** | **EL Hadji** |
| Continuer les recherches sur l’extraction (ABCDE)  Prétraitement des images   * + Reduce image size | Continuer les recherches sur l’extraction (ABCDE).  Proposition d’un brouillon du compte rendu  Prétraitement des données numériques | Continuer les recherches sur l’extraction (ABCDE).  Prétraitement des images   * + Reduce image size |

**Bilan de la semaine :**

|  |  |  |
| --- | --- | --- |
| **Les tâches** | **REMARQUE** | |
| **Continuer les recherches sur l’extraction (ABCDE).** | Les difficultés d'étudier l’évolution (E ) du mélanome:   * Nombre limité d'images pour certains patients.   **Une image contenant texte, capture d’écran, Police, ligne  Description générée automatiquement**   * Intervalle entre les images qui n'est pas précis. * Surtout, l'absence de référence initiale : sans une image de référence initiale pour chaque patient.   **Une image contenant texte, capture d’écran, nombre, menu  Description générée automatiquement** | |
|  | **Terminer :Proposition d’un brouillon du compte rendu** | |
| Prétraitement des données numériques |  | |
| **Code** | | **Nombre de valeurs uniques par colonne :** |
| **print("\n")**  **print("Nb Valeurs Uniques par Colonne")**  **for col in df.columns:**  **nb\_valeurs = len(pd.unique(df[col]))**  **print( "{} : {} valeurs".format(col,nb\_valeurs) )** | | Nb Valeurs Uniques par Colonne  image\_name : 33098 valeurs  patient\_id : 2056 valeurs  sex : 3 valeurs  age\_approx : 19 valeurs  anatom\_site\_general\_challenge : 7 valeurs  diagnosis : 9 valeurs  benign\_malignant : 2 |
| **Code** | | **Nb Valeurs NaN par colonne:** |
| print("\n")  print("Nb Valeurs NaN par colonne")  for col in df.columns:    nb\_valeurs = df[col].isna().sum()    print( "{} : {} NaN".format(col,nb\_valeurs) ) | | image\_name : 0 NaN  patient\_id : 0 NaN  sex : 65 NaN  age\_approx : 68 NaN  anatom\_site\_general\_challenge : 527 NaN  diagnosis : 0 NaN  benign\_malignant :0 |

* **Graphique pour visualiser la distribution des données avant le remplacement des valeurs manquantes**

**anatom\_site\_general**

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

**sex**

**Une image contenant texte, capture d’écran, Bleu électrique, Rectangle

Description générée automatiquement**

* **age\_approx**

Une image contenant texte, capture d’écran, diagramme, ligne

Description générée automatiquement

* ***Remplir les valeurs manquantes***

|  |
| --- |
| mode\_sex=df["sex"].mode()  df["sex"]=df["sex"].fillna("mode\_sex")  mode\_anatom\_site\_general\_challenge=df["anatom\_site\_general\_challenge"].mode()  df["anatom\_site\_general\_challenge"]=df["anatom\_site\_general\_challenge"].fillna("mode\_anatom\_site\_general\_challenge")  moyenne\_age\_approx=df["age\_approx"].mean()  df["age\_approx"]=df["age\_approx"].fillna(moyenne\_age\_approx) |

* ***Vérification***

|  |
| --- |
| # 4. Verifier si des colonnes ont des NaN  print("\n")  print("Nb Valeurs NaN par colonne")  for col in df.columns:    nb\_valeurs = df[col].isna().sum()    print( "{} : {} NaN".format(col,nb\_valeurs) ) |
| Nb Valeurs NaN par colonne  image\_name : 0 NaN  patient\_id : 0 NaN  sex : 0 NaN  age\_approx : 0 NaN  anatom\_site\_general\_challenge : 0 NaN  diagnosis : 0 NaN  benign\_malignant : 0 NaN  target : 0 NaN |

### Prétraitement des images

|  |  |  |
| --- | --- | --- |
| **Les Tâches à faire : prétraitement des données** | | |
| **Evelyne** | **Hassan** | **EL Hadji** |
| Continuer les recherches sur l’extraction (ABCDE) | Continuer les recherches sur l’extraction (ABCDE).  Prétraitement des données numériques   * + Dataset\_separation | Continuer les recherches sur l’extraction (ABCDE).  Prétraitement des images   * + Dataset\_separation |
| Prétraitement des images   * + Prerocessing   + Image\_separation   + Malignant Images augmentation |  | Prétraitement des images   * + Prerocessing   + Image\_separation   -Malignant Images augmentation |
|  | * Malignant dataset augmentation * Faire new dasaset |  |

|  |  |
| --- | --- |
| **Les tâches** | **REMARQUE** |
| * + **Reduce image size** | **A screenshot of a computer program  Description automatically generated with medium confidence** |
| * + **Preprocessing** | **A screenshot of a computer code  Description automatically generated with medium confidence** |
| * + **Dataset\_separation** | **Terminer**  **Une image contenant texte, Police, ligne, nombre  Description générée automatiquement**    **Une image contenant texte, capture d’écran, Police, ligne  Description générée automatiquement** |
| **Malignant dataset augmentation**  **Faire new dasaset** | **Terminer :**  **A screenshot of a computer program  Description automatically generated with medium confidence**   * **Ajouter les référence des images aux dataset** * **Vérification de pourcentage après augmentation**   **Une image contenant texte, Police, logiciel, Page web  Description générée automatiquement**  **Une image contenant texte, Police, ligne, capture d’écran  Description générée automatiquement** |

**PLANNING DU TRAVAIL POUR LA SEMAINE (5) 5 à 9 JUIN**

|  |  |  |
| --- | --- | --- |
| **Les Tâches à faire : prétraitement des données et création des modèles** | | |
| **Evelyne** | **Hassan** | **EL Hadji** |
| Améliorer la segmentation  Améliorer extraction ABD | Améliorer la segmentation  Améliorer extraction ABD | Faire une recherche sur le serveur |
| Faire un 1er modèle DML | Utiliser Autokeras pour créer un modèle | Faire un 1er modèle ML |
| Comparer les résultats  Du modèle de hassan  Faire le recherche extraction de C | Faire un 1er modèle DML  Choisir et sélectionné le meilleur : segmentation et extraction | Faire un 2eme modèle ML  Avec données extraites  Comparer les résultats |

|  |  |
| --- | --- |
| **Les tâches** | **REMARQUE** |
| **Améliorer la segmentation** | **Améliorer si possible :**   |  |  | | --- | --- | | **Une image contenant acarien, rose, pêche  Description générée automatiquement** |  | |
| **Améliorer extraction ABD** | |  |  | | --- | --- | | **Une image contenant texte, Police, capture d’écran, blanc  Description générée automatiquement** | **Une image contenant Caractère coloré, cœur, capture d’écran  Description générée automatiquement** | |
| Effectuer une segmentation sur l'ensemble du jeu de données. | **Une image contenant texte, capture d’écran, Police, nombre  Description générée automatiquement**   |  |  |  |  | | --- | --- | --- | --- | |  | **Une image contenant carte, texte  Description générée automatiquement** |  |  | |
| **Insérer dans dataset** | **Une image contenant texte, Police, capture d’écran  Description générée automatiquement** |

# Model pipelines

## Machine Learning

A screenshot of a computer code

Description automatically generated with low confidence

## AutoKeras

|  |  |
| --- | --- |
| **Utiliser Autokeras pour créer un modèle** | **RESULTAT AVEC DEUX IMAGES DÉJÀ VU**      **RESULTAT AVEC SIX IMAGE NON VU** |

## Deep MachineLearning

A picture containing text, font, screenshot, line

Description automatically generated

A screenshot of a computer code

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

Le model ne performe pas bien pour le moment l’entrainement n’a pas encore été correctement fait

A picture containing text, sketch, black and white

Description automatically generated

**Recherche sur le serveur**

Après plusieurs tentatives en vain, on a pu finalement importer toutes les images dans le serveur.

# Présentation des travaux futurs

* Recommencer le pretraitement des image dans le serveur
* Entrainement des modèles au niveau du serveur
* Amélioration des modèles
* Ajouter notre travail sur Gitlab