# Créer votre Personnage et participer à l’aventure



[***Créer votre Personnage et participer à l’aventure*** 1](#_Toc38625985)

[Introduction 2](#_Toc38625986)

[Recherche du sujet 2](#_Toc38625987)

[Importation et sources des données 3](#_Toc38625988)

[Analyse des Données 3](#_Toc38625989)

[Nettoyage des données 4](#_Toc38625990)

[Entrainement du model 5](#_Toc38625991)

[Régression 6](#_Toc38625992)

[Application des régressions 7](#_Toc38625993)

[Solution finale 7](#_Toc38625994)

Introduction

Dans le cadre de notre cursus en école d’ingénieur nous avons découvert le Machine Learning et nous avons eu pour but de réaliser un projet mettant en évidence toutes les connaissances que nous avons acquises.

Ainsi, nous avons dû faire face à ce challenge, dont la première étape était de trouver un sujet.

Recherche du sujet

Le choix du sujet est une des étapes les plus difficiles et encore plus dans le cas d’un projet de Machine Learning car le choix du dataset est déterminent. De plus, avec le peu d’expérience dont nous disposons, il est difficile d’évaluer un dataset sans effectuer de test. Donc, nous avons commencé par chercher sur différents sites Web comme kaggle. Le but était de trouver un dataset contenant assez de données pour pouvoir associer prédiction et efficacité.

Premièrement, nous avons trouvé un ensemble de données sur la situation actuelle du coronavirus. Malheureusement le coronavirus étant un sujet fortement abordé dans le contexte actuel, nous avons préféré se distinguer et donc explorer d’autres pistes.

Etant fans de la célèbre série Game of Thrones, notre choix s’est orienté vers un Dataset concernant cette série.

Ainsi, notre idée fut de donner la possibilité à un utilisateur de créer un personnage en fonction de différents paramètres et de déterminer si celui-ci était vivant ou non à la fin de la série. Si c’est le cas, nous avions pour but de déterminer dans quel épisode, il était décédé, et de quelle façon.

Comme vous l’avez sans doute compris cela va nécessiter une Classification pour déterminer la survie du personnage et une régression pour déterminer l’épisode de mort et la description de la mort.

Pour réaliser cela, nous avons procéder en suivant ces 5 étapes :

1. Analyse des données
2. Nettoyage des données
3. Entrainement du modèle
4. Régression
5. Application

Importation et sources des données

Pour réussir ce projet, nous avons utilisé deux datasets différents. L’un est un dataset concernant les personnages et l’autre concerne les épisodes de la série.

Le premier donne des informations (au nombre de 35) sur plus de 350 personnages et le deuxième décrit les 73 épisodes de la série avec 8 variables.

Ensuite, nous avons décidé d'utiliser Colab pour développer les programmes en python et la première chose que on a fait c’est l’importation des datasets.

Analyse des Données

Dans cette partie, nous allons réaliser plusieurs tests et analyses dans le but de déterminer quels sont les paramètres utilisables pour la classification.

Pour bien débuter notre analyse de données, il faut prendre connaissance de notre dataset et décider si certaines colonnes doivent être supprimées ou non à l’aide de plusieurs moyens.

Dans notre cas, nous avons affiché le pourcentage de valeurs manquantes par colonne ce qui a permis de supprimer certaines d’entre elles.

Nous avons ensuite utilisé une matrice de corrélation pour obtenir les coefficients de corrélation entre les variables de notre dataset. Le but de cette matrice est de déterminer s’il existe de la colinéarité entre celles-ci.

Ainsi pour résumer l’analyse des données a eu plusieurs objectifs dans notre cas et nous a permis de :

* Réaliser un test pour déterminer quels paramètres nous pouvons utiliser pour la classification.
* Affichage des pourcentages de valeurs manquantes par colonne pour le data cleaning.
* Réaliser une matrice de corrélation permettant d'obtenir les coefficients de corrélation entre les variables de notre dataset.

Nettoyage des données

Nous avons maintenant à notre disposition l'ensemble de données et nous avons donc commencer à supprimer les variables n’étant pas utiles pour rendre notre dataset exploitable lors de la classification.

Ensuite, nous avons décidé de remplir certaines colonnes possédant des valeurs manquantes par la médiane de celle-ci.

Le nettoyage des données possède plusieurs intérêts :

* Amélioration de la prise de décision
* Meilleure efficacité
* Des données propres pour des objectifs

Entrainement du model

Dans cette partie les objectifs sont les suivants :

* Obtenir une classification efficace.
* Prédire si un personnage, décrit par un certain nombre de variable, est vivant ou non à la fin de la série Game of Thrones.

Nous avons commencé par sélectionner les paramètres sur lesquels va se baser la classification et nous avons ensuite séparé notre dataset en deux. Une partie sera utilisé pour l'entrainement du modèle et l'autre partie pour vérifier que cet entraînement est efficace ou non.

Et enfin nous avons utilisé plusieurs algorithmes de classification pour tirer le meilleur de notre dataset.

Grâce à cela, nous avons pu obtenir une classification avec les résultats suivants :

* Decision Tree Classification
* Random Forest Classification



Nous remarquons que les résultats obtenus sont plutôt satisfaisants et nous permettent de réaliser une prédiction précise.

Régression

Dans cette partie, l'objectif est de réaliser une régression pour déterminer dans quel épisode un personnage est mort et de quelle façon.

Cette régression est effectivement uniquement possible pour les personnages dont le décès a été prédit par la classification de la partie précédente ou ceux dont le décès est indiqué dans le dataset. Nous avons supprimé, du dataset, les données correspondantes aux personnes vivantes.

Dans cette partie, nous avons utilisé une régression logistique qui a pour but de faire correspondre un modèle mathématique aux données dont nous disposons.

Pour effectuer cette partie, il est nécessaire de reprendre quelques étapes déjà effectuée lors de la classification :

* Compréhension des données
* Préparation des données
* Découpage du jeu données en Training Set et Test Set.
* Evaluation du modèle de régression linéaire

Grâce à cela, nous avons pu prédire dans quel épisode un personnage était décédé avec les résultats suivants :



Nous remarquons que les résultats sont plutôt médiocres et cela peut être expliqué par le fait que nous avons très peu de données sur lesquelles travailler.

Ensuite, nous réalisons une deuxième régression pour déterminer les causes de la mort.

Application des régressions

En appliquant les régressions ci-dessus à des données, nous obtenons les résultats suivants :





Nous pouvons ensuite transmettre ces résultats à notre application web.

Application web

Dans cette partie, nous allons vous présenter l’application web que nous avons mis en place pour effectuer les prédictions énoncées ci-dessus.

