UNIVERSITÉ
PARIS DESCARTES
UFR DE MATHÉMATIQUES ET INFORMATIQUE

2010

Cahier des charges

L2I1 : Machine Learning from disaster

UNIVERSITE PARIS DESCARTES

UNIVERSITÉ PARIS DESCARTES
UFR DE MATHÉMATIQUES ET INFORMATIQUE

Cahier des charges

Nom du projet

Les informations d'identification du document

Référence du document :	D2
Version du document :	1.01
Date du document :	19/02/23
Auteur(s) :	David Janiszek

Les éléments de vérification du document

Validé par :	David Janiszek
Validé le :	19/2/23
Soumis le :	19/2/23
Type de diffusion :	Document électronique (.odt)
Confidentialité :	Réservé aux étudiants UFR Maths-Info de l'université Paris Descartes

Les éléments d'authentification

Maître d'ouvrage:	**Chef de projet :**
Date / Signature :	**Date / Signature :**

Mots clés : modèle de cahier des charges, ce n'est qu'un modèle !

Sommaire

1. Introduction (ou préambule)

Précise l’objectif du document et en résume le contenu

2. Guide de lecture

Précise, pour chaque type de lecteur, comment utiliser efficacement le document

2.1. Maîtrise d’œuvre

2.1.1. Responsable

2.1.2. Personnel administratif

2.1.3. Personnel technique

2.2. Maîtrise d’ouvrage

2.2.1. Responsable

2.2.2. Personnel technique

3.Concepts de base

Dans ce projet, nous allons utiliser des ensembles de données similaires qui comprennent des informations sur les passagers comme le nom, l'âge, le sexe, la classe socio-économique, la situation familiales, le numéro de cabine, le prix du ticket, le port d’embarquement, etc.

Un dataset est intitulé ‘train.csv’ et l'autre est intitulé ‘test.csv’.

‘train.csv’ contiendra les détails d'un sous-ensemble de passagers à bord (891 pour être exact) et surtout, révélera s'ils ont survécu ou non, aussi connu comme la "vérité de base".

Le dataset ‘test.csv’ contient des informations similaires mais ne révèle pas la "vérité de base" pour chaque passager. C'est à vous de prédire ces résultats.

En utilisant les modèles que nous avons trouvés dans les données train.csv, nous allons prédire si les 418 autres passagers à bord (trouvés dans test.csv) ont survécu.

Lors de l'analyse des données, nous allons utiliser divers outils mathématiques tels que matplotlib et seaborn afin d'analyser les résultats et mieux déterminer le modèle que nous utiliserons pour prévoir la survie des passagers à bord.

Enfin, nous utiliserons Streamlit pour construire une application web permettant de mieux visualiser les résultats de notre analyse de données.

Notre projet sera développé avec les outils suivants :

- Jupyter Notebook (avec Python dans l'environnement d'Anaconda 3)
- Serveur SVN (via le logiciel client Tortoise SVN.)
- Numpy
- Pandas (data management) : https://pandas.pydata.org/
- Seaborn (https://seaborn.pydata.org/examples/index.html)
- Matplotlib (https://matplotlib.org/stable/gallery/index)
- Les différents modèles de Machine Learning : Linear Models, etc.
- Streamlit (Streamlit • The fastest way to build and share data apps)
- Plotly (Plotly : The front end for ML and data science models)

4.Contexte

Le naufrage du Titanic est l'un des naufrages les plus tristement célèbres de l'histoire.

Le 15 avril 1912, lors de son voyage inaugural, le RMS Titanic, largement considéré comme "insubmersible", a coulé après avoir heurté un iceberg. Malheureusement, il n'y avait pas assez de canots de sauvetage pour tout le monde à bord, ce qui a entraîné la mort de 1502 des 2224 passagers et membres d'équipage. Bien qu'il y ait eu une part de chance dans la survie, il semble que certains groupes de personnes aient eu plus de chances de survivre que d'autres.

Dans ce défi, notre équipe va construire un modèle prédictif qui répond à la question

suivante : "Quelles sont les personnes qui ont le plus de chances de survivre ?" en utilisant les données des passagers (nom, âge, sexe, classe socio-économique, etc.).

Pour plus d'informations, veuillez consulter le site :
https://www.kaggle.com/competitions/titanic/overview

5. Historique

S'inscrivant dans un cadre académique, ce projet permet une initiation au « Machine Learning » tout en se rapprochant des réalités du monde professionnel. Ainsi les membres du projet découvrent la plupart des notions abordées en s'appuyant sur les compétences acquises au cours des deux premières années d'études. Le projet est supervisé par un encadrant et soumis à certaines consignes dans la méthodologie.

6. Description de la demande

6.1. Les objectifs

Voici la liste des objectifs fixés :

- Obtenir le meilleur classement possible au concours « Kaggle : Titanic - Machine Learning from Disaster ».
- Trouver les facteurs les plus influents sur les chances de survie des passagers.
- Fournir une documentation (...)
- Permettre une visualisation simple et accessible des données.

6.2. Produit du projet

Le produit est un modèle prédictif codé en python, accompagné d'une documentation type notebook 'Jupyter' explicitant les résultats et d'une application Web offrant une interface viable.

6.3. Les fonctions du produit

Le modèle doit permettre de répondre à la question suivante : « A l'aide des données relatives aux passagers au Titanic (nom, âge, sexe, classe socio-économique, etc.), qui a le plus de chance de survivre ? ».

Ainsi, le produit doit être en mesure de prédire la survie ou non d'un passager à partir de ses données avec un taux de précision maximum.

+documentation

(option) De plus, le produit doit permettre la Visualisation interactive des données et résultats obtenus à travers une interface simple et visuelle.

6.4. Critères d'acceptabilité et de réception

Le modèle prédictif doit être soumis à l'ensemble des règles de compétition « Kaggle ». Ainsi on mesurera la qualité du modèle prédictif avec le score « Kaggle » obtenu pour le produit final. Le principal critère de réussite est donc le rang obtenu au classement.

+Documentation

(OPTION) Concernant l'interface, celle-ci doit être simple et lisible. Elle doit permettre à toute personne la compréhension des résultats obtenus. (application WEB, sans contraintes matérielles, streamlit)

7. Contraintes

7.1. Contraintes de coûts

Spécifier le budget alloué au projet (Coût humain)

7.2. Contraintes de délais

Le cahier des charges (à rendre semaine 3);

Le cahier de recette (à rendre semaine 4);

La conception générale

La conception détaillée (à rendre semaine 5);

Le manuel d'utilisation (à rendre semaine 11);

Le manuel d'installation (à rendre semaine 11);

Le plan de tests (à rendre semaine 11);

La documentation interne du code (à rendre semaine 11).

Le code sources du programme (à rendre semaine 11);

Le rapport de projet (avant la soutenance);

Le résumé en français et en anglais (avant la soutenance);

Les diapositives sonorisées (avant la soutenance);

7.3. Contraintes matérielles

A expliciter

Temps de calcul, hébergement etc

7.4. Autres contraintes

Spécifier les éventuelles contraintes à prendre en compte dans le cadre du projet (normes techniques, clauses juridiques, etc.)

8. Déroulement du projet

8.1. Planification

Représenter les grandes phases du projet et les étapes principales

8.2. Ressources

Lister les ressources humaines et matérielles que le client peut mettre à la disposition du prestataire (donnée, librairies)

9. Annexes

Lister et joindre au cahier des charges les éventuels documents que le client peut mettre à disposition

10. Glossaire

Définit l'ensemble des termes spécialisés du document

11. Références

Indique les références bibliographiques vers d'autres documents apportant des informations complémentaires

12. Index

Liste les mots-clés du document et où les trouver dans celui-ci