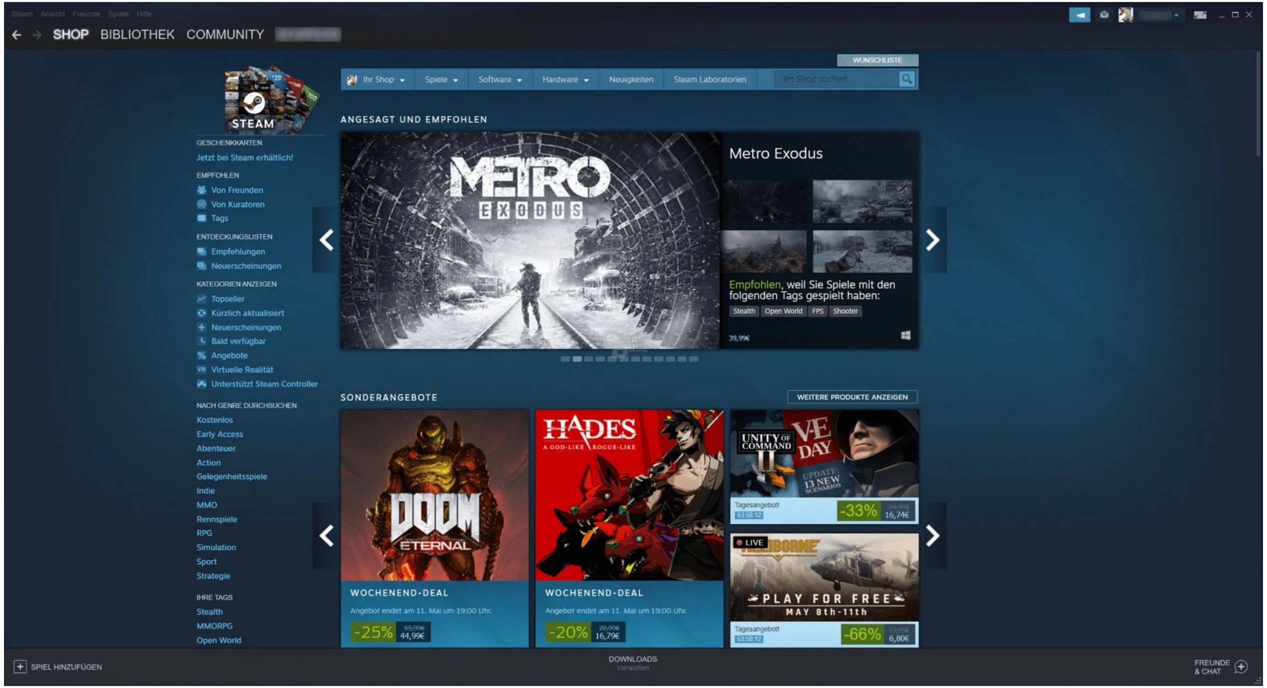
Prédire le succès d'un jeu vidéo Steam avant son lancement

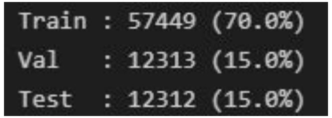
Projet de Fouilles de Données & aide à la décision

BRAHIM KHLIL Chems Eddine

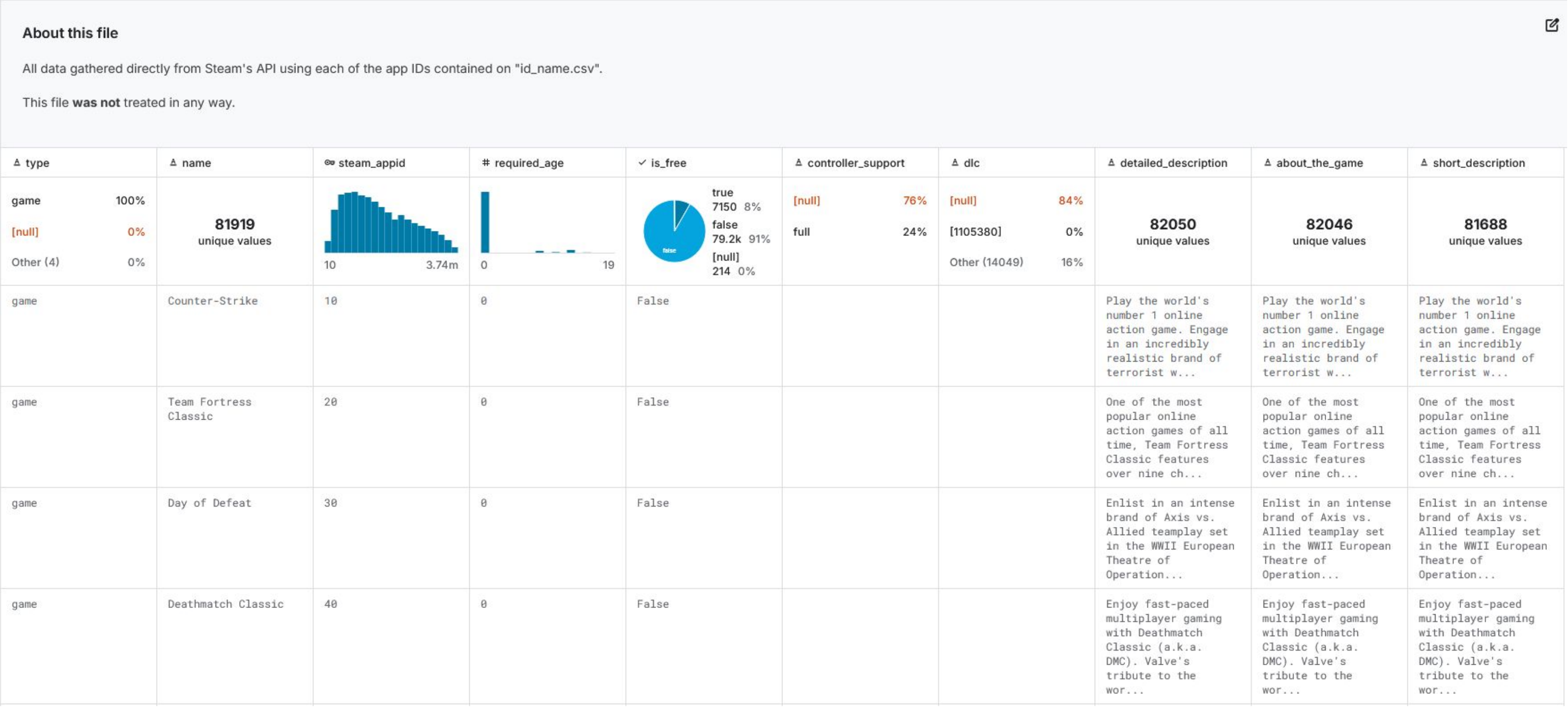
Choix du dataset et de la problématique



Le Dataset



```
Train : 57449 (70.0%)
Val   : 12313 (15.0%)
Test  : 12312 (15.0%)
```



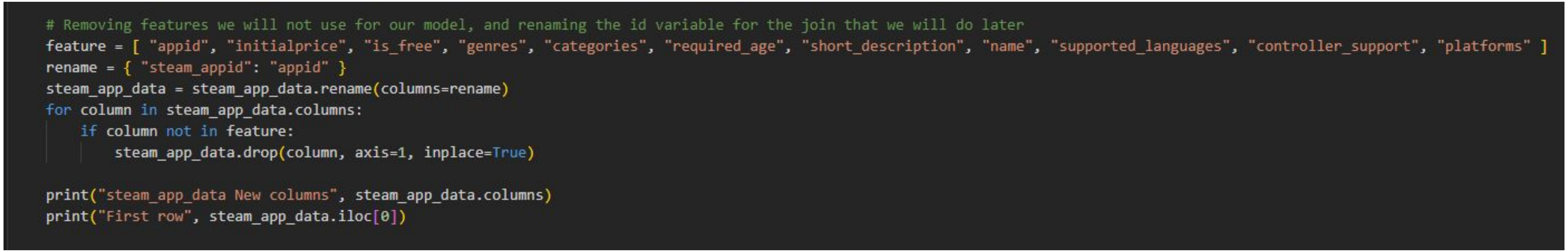
About this file

All data gathered directly from Steam's API using each of the app IDs contained on "id_name.csv".

This file **was not** treated in any way.

type	name	steam_appid	required_age	is_free	controller_support	dlc	detailed_description	about_the_game	short_description
game 100% [null] 0% Other (4) 0%	81919 unique values	10 3.74m	0 19	true 7150 8% false 79.2k 91% [null] 214 0%	[null] 76% full 24%	[null] 84% [1105380] 0% Other (14049) 16%	82050 unique values	82046 unique values	81688 unique values
game	Counter-Strike	10	0	False			Play the world's number 1 online action game. Engage in an incredibly realistic brand of terrorist w...	Play the world's number 1 online action game. Engage in an incredibly realistic brand of terrorist w...	Play the world's number 1 online action game. Engage in an incredibly realistic brand of terrorist w...
game	Team Fortress Classic	20	0	False			One of the most popular online action games of all time, Team Fortress Classic features over nine ch...	One of the most popular online action games of all time, Team Fortress Classic features over nine ch...	One of the most popular online action games of all time, Team Fortress Classic features over nine ch...
game	Day of Defeat	30	0	False			Enlist in an intense brand of Axis vs. Allied teamplay set in the WWII European Theatre of Operation...	Enlist in an intense brand of Axis vs. Allied teamplay set in the WWII European Theatre of Operation...	Enlist in an intense brand of Axis vs. Allied teamplay set in the WWII European Theatre of Operation...
game	Deathmatch Classic	40	0	False			Enjoy fast-paced multiplayer gaming with Deathmatch Classic (a.k.a. DMC). Valve's tribute to the wor...	Enjoy fast-paced multiplayer gaming with Deathmatch Classic (a.k.a. DMC). Valve's tribute to the wor...	Enjoy fast-paced multiplayer gaming with Deathmatch Classic (a.k.a. DMC). Valve's tribute to the wor...

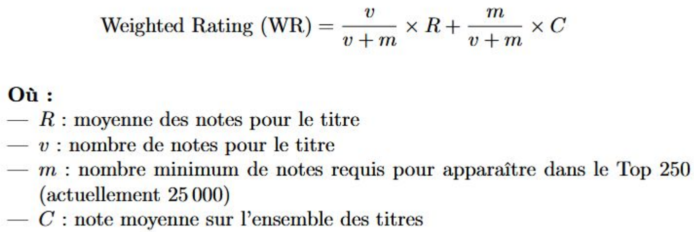
Features



```python
# Removing features we will not use for our model, and renaming the id variable for the join that we will do later
feature = [ "appid", "initialprice", "is_free", "genres", "categories", "required_age", "short_description", "name", "supported_languages", "controller_support", "platforms" ]
rename = { "steam_appid": "appid" }
steam_app_data = steam_app_data.rename(columns=rename)
for column in steam_app_data.columns:
    if column not in feature:
        steam_app_data.drop(column, axis=1, inplace=True)

print("steam_app_data New columns", steam_app_data.columns)
print("First row", steam_app_data.iloc[0])
```

Variable cible : is_hit

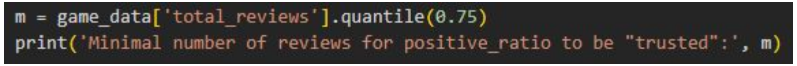
IMDb



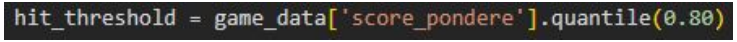
$$\text{Weighted Rating (WR)} = \frac{v}{v+m} \times R + \frac{m}{v+m} \times C$$

Où :

— R : moyenne des notes pour le titre
— v : nombre de notes pour le titre
— m : nombre minimum de notes requis pour apparaître dans le Top 250 (actuellement 25 000)
— C : note moyenne sur l'ensemble des titres

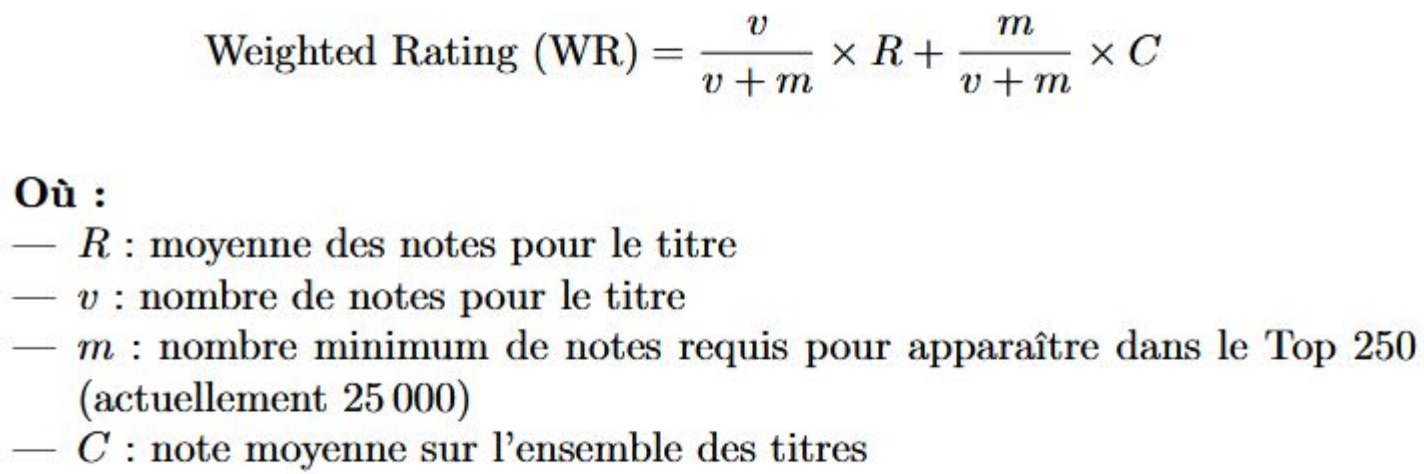


```python
m = game_data['total_reviews'].quantile(0.75)
print('Minimal number of reviews for positive_ratio to be "trusted":', m)
```



```python
hit_threshold = game_data['score_pondere'].quantile(0.80)
```

Variable cible : is_hit



$$\text{Weighted Rating (WR)} = \frac{v}{v+m} \times R + \frac{m}{v+m} \times C$$

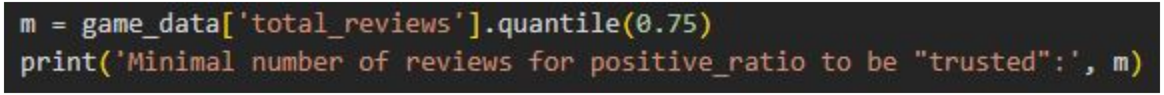
Où :

— R : moyenne des notes pour le titre

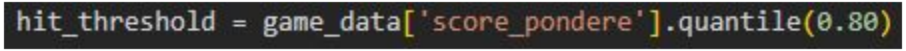
— v : nombre de notes pour le titre

— m : nombre minimum de notes requis pour apparaître dans le Top 250 (actuellement 25 000)

— C : note moyenne sur l'ensemble des titres

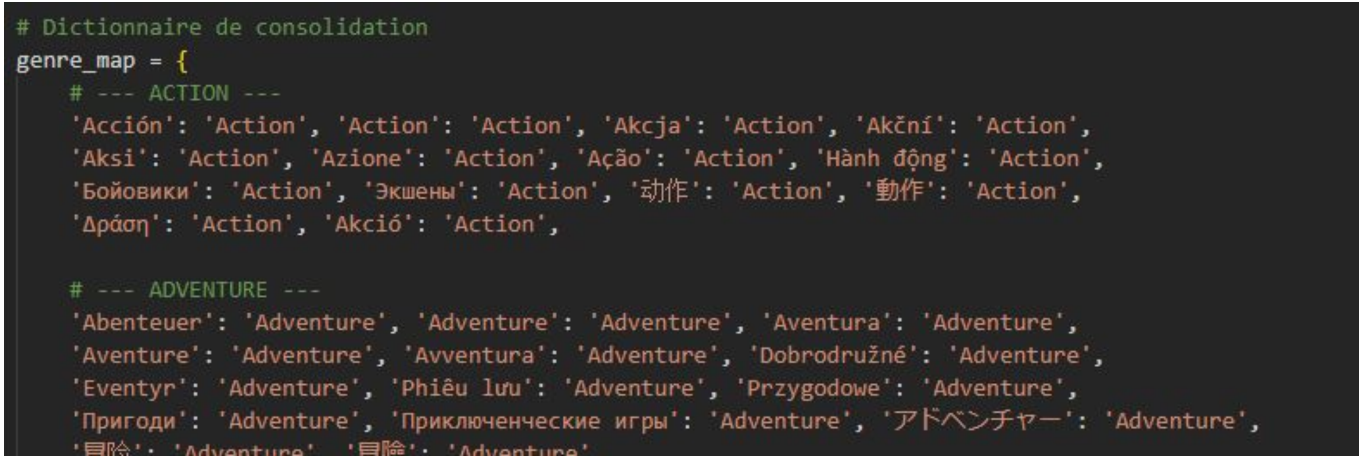


```
m = game_data['total_reviews'].quantile(0.75)
print('Minimal number of reviews for positive_ratio to be "trusted":', m)
```

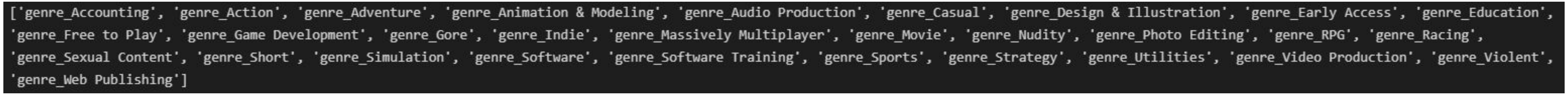


```
hit_threshold = game_data['score_pondere'].quantile(0.80)
```

Feature Engineering : Genres

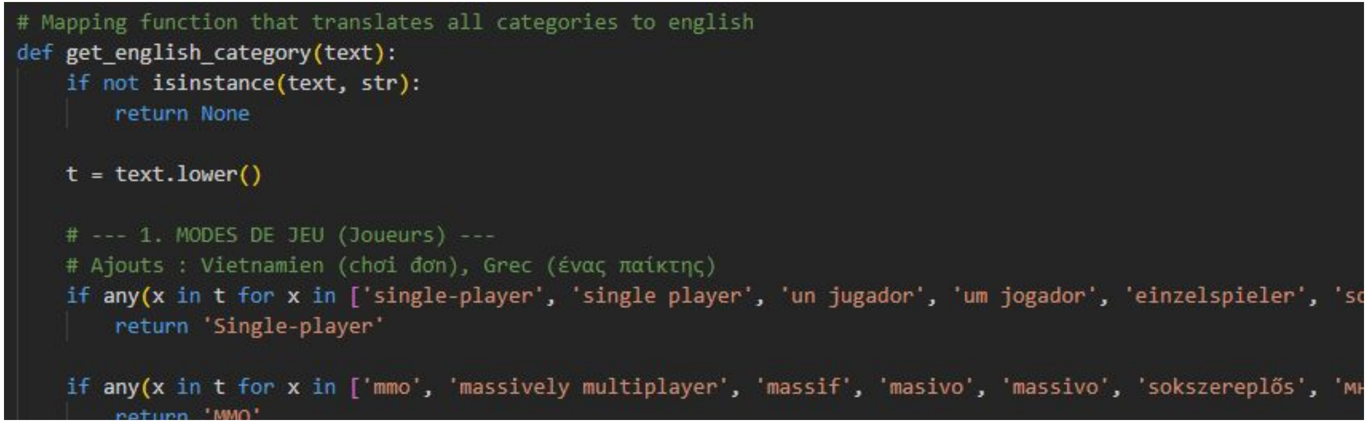


```python
# Dictionnaire de consolidation
genre_map = {
    # --- ACTION ---
    'Acción': 'Action', 'Action': 'Action', 'Akcja': 'Action', 'Akční': 'Action',
    'Aksi': 'Action', 'Azione': 'Action', 'Ação': 'Action', 'Hành động': 'Action',
    'Бойовики': 'Action', 'Экшены': 'Action', '动作': 'Action', '動作': 'Action',
    'Δράση': 'Action', 'Akció': 'Action',

    # --- ADVENTURE ---
    'Abenteuer': 'Adventure', 'Adventure': 'Adventure', 'Aventura': 'Adventure',
    'Aventure': 'Adventure', 'Avventura': 'Adventure', 'Dobrodružné': 'Adventure',
    'Eventyr': 'Adventure', 'Phiêu lưu': 'Adventure', 'Przygodowe': 'Adventure',
    'Пригоди': 'Adventure', 'Приключенческие игры': 'Adventure', 'アドベンチャー': 'Adventure',
    '冒险': 'Adventure', '冒險': 'Adventure',
```

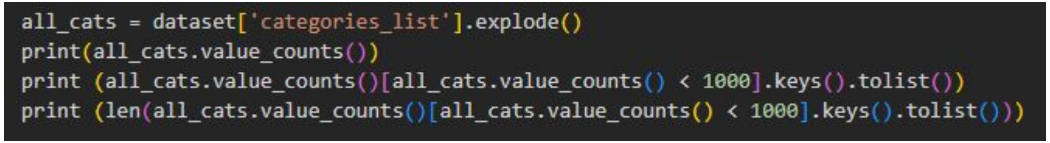


```
['genre_Accounting', 'genre_Action', 'genre_Adventure', 'genre_Animation & Modeling', 'genre_Audio Production', 'genre_Casual', 'genre_Design & Illustration', 'genre_Early Access', 'genre_Education',
'genre_Free to Play', 'genre_Game Development', 'genre_Gore', 'genre_Indie', 'genre_Massively Multiplayer', 'genre_Movie', 'genre_Nudity', 'genre_Photo Editing', 'genre_RPG', 'genre_Racing',
'genre_Sexual Content', 'genre_Short', 'genre_Simulation', 'genre_Software', 'genre_Software Training', 'genre_Sports', 'genre_Strategy', 'genre_Utilities', 'genre_Video Production', 'genre_Violent',
'genre_Web Publishing']
```

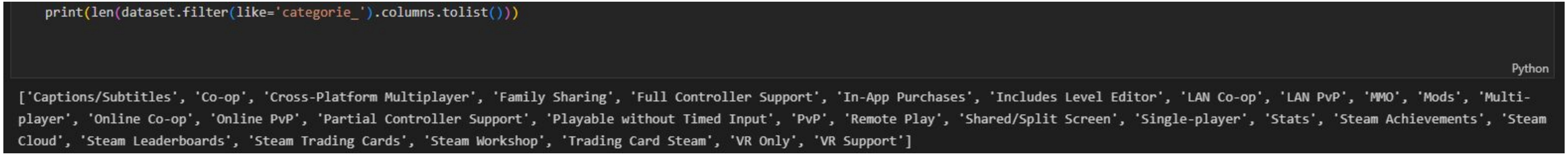
Feature Engineering : Catégories



```
# Mapping function that translates all categories to english
def get_english_category(text):
    if not isinstance(text, str):
        return None

    t = text.lower()

    # --- 1. MODES DE JEU (Joueurs) ---
    # Ajouts : Vietnamien (chơi đơn), Grec (ένας παίκτης)
    if any(x in t for x in ['single-player', 'single player', 'un jugador', 'um jogador', 'einzelspieler', 'sc
        return 'Single-player'

    if any(x in t for x in ['mmo', 'massively multiplayer', 'massif', 'masivo', 'massivo', 'sokszereplős', 'мн
        return 'MMO'
```



```
all_cats = dataset['categories_list'].explode()
print(all_cats.value_counts())
print (all_cats.value_counts()[all_cats.value_counts() < 1000].keys().tolist())
print (len(all_cats.value_counts()[all_cats.value_counts() < 1000].keys().tolist()))
```

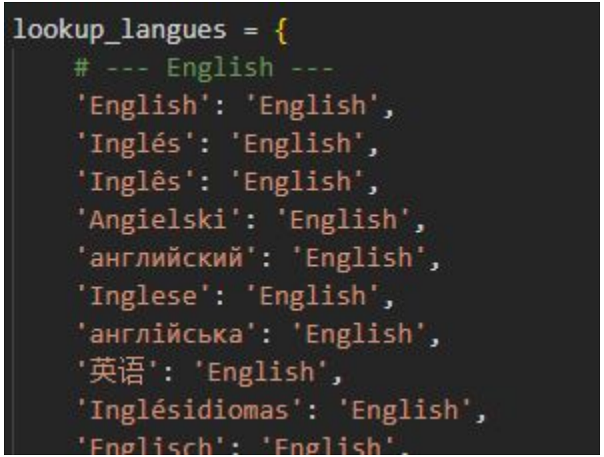


```
print(len(dataset.filter(like='categorie_').columns.tolist()))
```

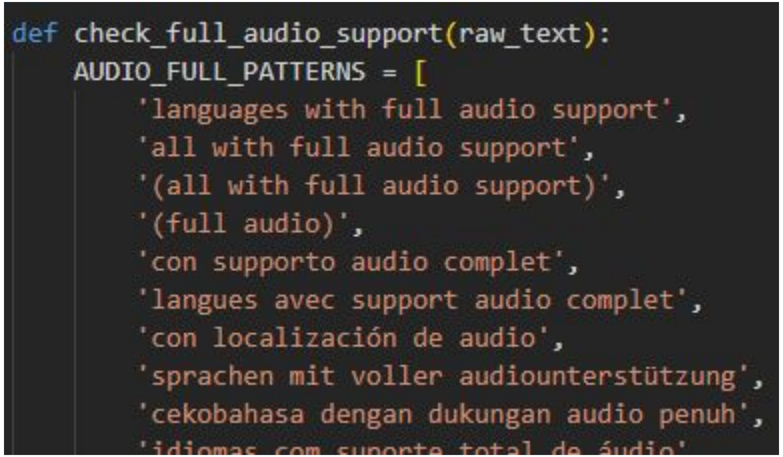
Python

```
['Captions/Subtitles', 'Co-op', 'Cross-Platform Multiplayer', 'Family Sharing', 'Full Controller Support', 'In-App Purchases', 'Includes Level Editor', 'LAN Co-op', 'LAN PvP', 'MMO', 'Mods', 'Multi-player', 'Online Co-op', 'Online PvP', 'Partial Controller Support', 'Playable without Timed Input', 'PvP', 'Remote Play', 'Shared/Split Screen', 'Single-player', 'Stats', 'Steam Achievements', 'Steam Cloud', 'Steam Leaderboards', 'Steam Trading Cards', 'Steam Workshop', 'Trading Card Steam', 'VR Only', 'VR Support']
```

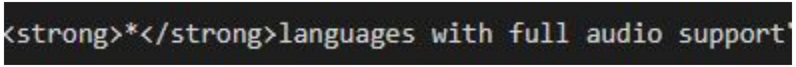
Feature Engineering : Langues supportées par le jeu



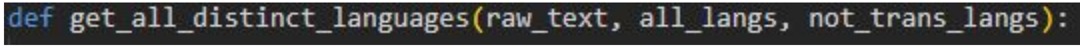
```
lookup_langues = {
    # --- English ---
    'English': 'English',
    'Inglés': 'English',
    'Inglês': 'English',
    'Angielski': 'English',
    'английский': 'English',
    'Inglese': 'English',
    'англійська': 'English',
    '英语': 'English',
    'Inglésidiomas': 'English',
    'Englisch': 'English',
```



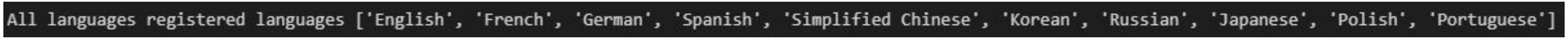
```
def check_full_audio_support(raw_text):
    AUDIO_FULL_PATTERNS = [
        'languages with full audio support',
        'all with full audio support',
        '(all with full audio support)',
        '(full audio)',
        'con supporto audio complet',
        'langues avec support audio complet',
        'con localización de audio',
        'sprachen mit voller audiounterstützung',
        'cekobahasa dengan dukungan audio penuh',
        'idiomas com suporte total de áudio',
```



```
<strong>*</strong>languages with full audio support
```

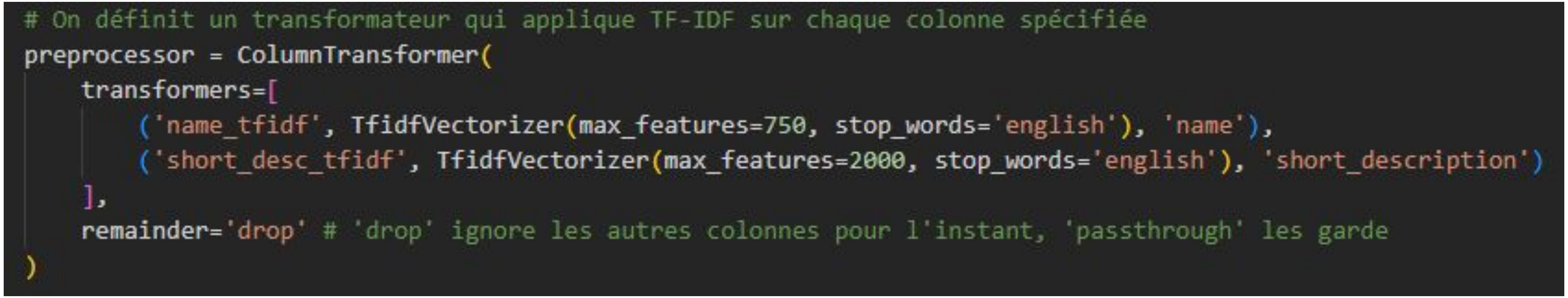


```
def get_all_distinct_languages(raw_text, all_langs, not_trans_langs):
```



```
All languages registered languages ['English', 'French', 'German', 'Spanish', 'Simplified Chinese', 'Korean', 'Russian', 'Japanese', 'Polish', 'Portuguese']
```

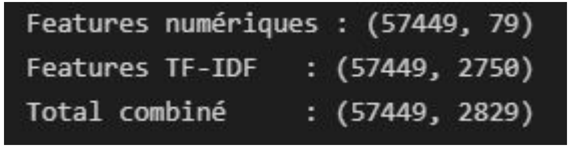
Feature Engineering : TF-IDF : short_description & name



```
# On définit un transformateur qui applique TF-IDF sur chaque colonne spécifiée
preprocessor = ColumnTransformer(
    transformers=[
        ('name_tfidf', TfidfVectorizer(max_features=750, stop_words='english'), 'name'),
        ('short_desc_tfidf', TfidfVectorizer(max_features=2000, stop_words='english'), 'short_description')
    ],
    remainder='drop' # 'drop' ignore les autres colonnes pour l'instant, 'passthrough' les garde
)
```

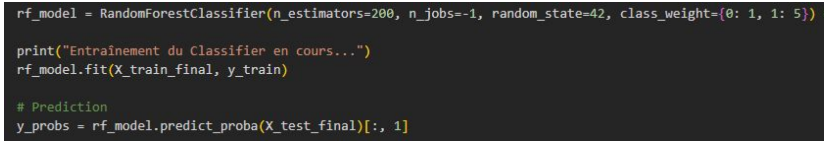
Une description courte (comme un “pitch” bien rédigée d’un jeu fournit souvent des données intéressantes sur son contenu et ses caractéristiques.

Jeu sans description -> peu de chance de succès

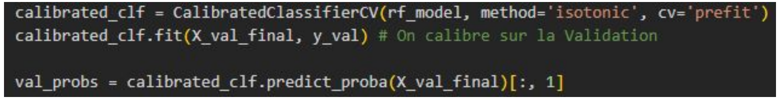


```
Features numériques : (57449, 79)
Features TF-IDF     : (57449, 2750)
Total combiné       : (57449, 2829)
```

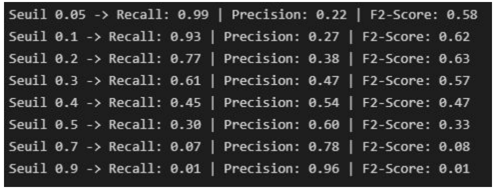
Modélisation 1 : Random Forest



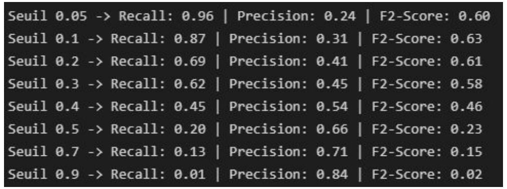
```
rf_model = RandomForestClassifier(n_estimators=200, n_jobs=-1, random_state=42, class_weight={0: 1, 1: 5})

print("Entraînement du Classifier en cours...")
rf_model.fit(X_train_final, y_train)

# Prediction
y_probs = rf_model.predict_proba(X_test_final)[:, 1]
```



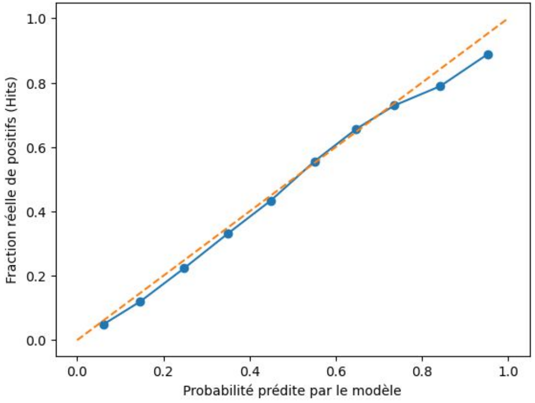
```
calibrated_clf = CalibratedClassifierCV(rf_model, method='isotonic', cv='prefit')
calibrated_clf.fit(X_val_final, y_val) # On calibre sur la Validation

val_probs = calibrated_clf.predict_proba(X_val_final)[:, 1]
```



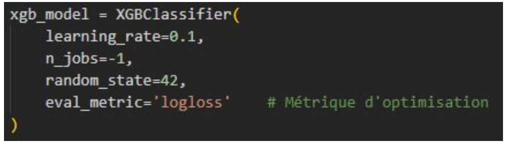
```
Seuil 0.05 -> Recall: 0.99 | Precision: 0.22 | F2-Score: 0.58
Seuil 0.1 -> Recall: 0.93 | Precision: 0.27 | F2-Score: 0.62
Seuil 0.2 -> Recall: 0.77 | Precision: 0.38 | F2-Score: 0.63
Seuil 0.3 -> Recall: 0.61 | Precision: 0.47 | F2-Score: 0.57
Seuil 0.4 -> Recall: 0.45 | Precision: 0.54 | F2-Score: 0.47
Seuil 0.5 -> Recall: 0.30 | Precision: 0.60 | F2-Score: 0.33
Seuil 0.7 -> Recall: 0.07 | Precision: 0.78 | F2-Score: 0.08
Seuil 0.9 -> Recall: 0.01 | Precision: 0.96 | F2-Score: 0.01
```



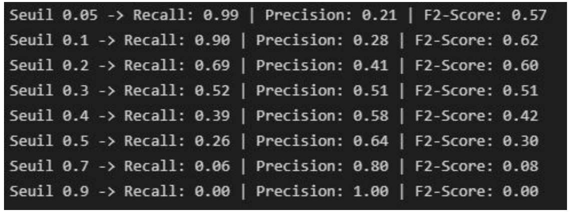
```
Seuil 0.05 -> Recall: 0.96 | Precision: 0.24 | F2-Score: 0.60
Seuil 0.1 -> Recall: 0.87 | Precision: 0.31 | F2-Score: 0.63
Seuil 0.2 -> Recall: 0.69 | Precision: 0.41 | F2-Score: 0.61
Seuil 0.3 -> Recall: 0.62 | Precision: 0.45 | F2-Score: 0.58
Seuil 0.4 -> Recall: 0.45 | Precision: 0.54 | F2-Score: 0.46
Seuil 0.5 -> Recall: 0.20 | Precision: 0.66 | F2-Score: 0.23
Seuil 0.7 -> Recall: 0.13 | Precision: 0.71 | F2-Score: 0.15
Seuil 0.9 -> Recall: 0.01 | Precision: 0.84 | F2-Score: 0.02
```



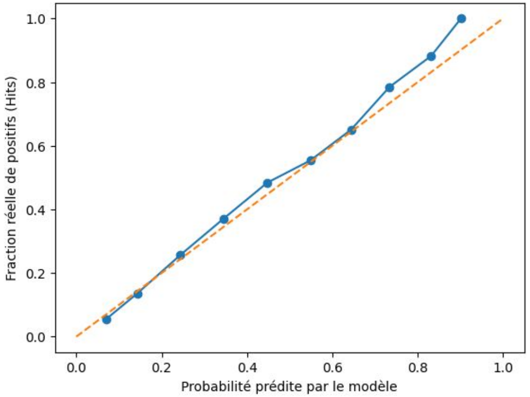
Modélisation 2 : Gradient Boosting



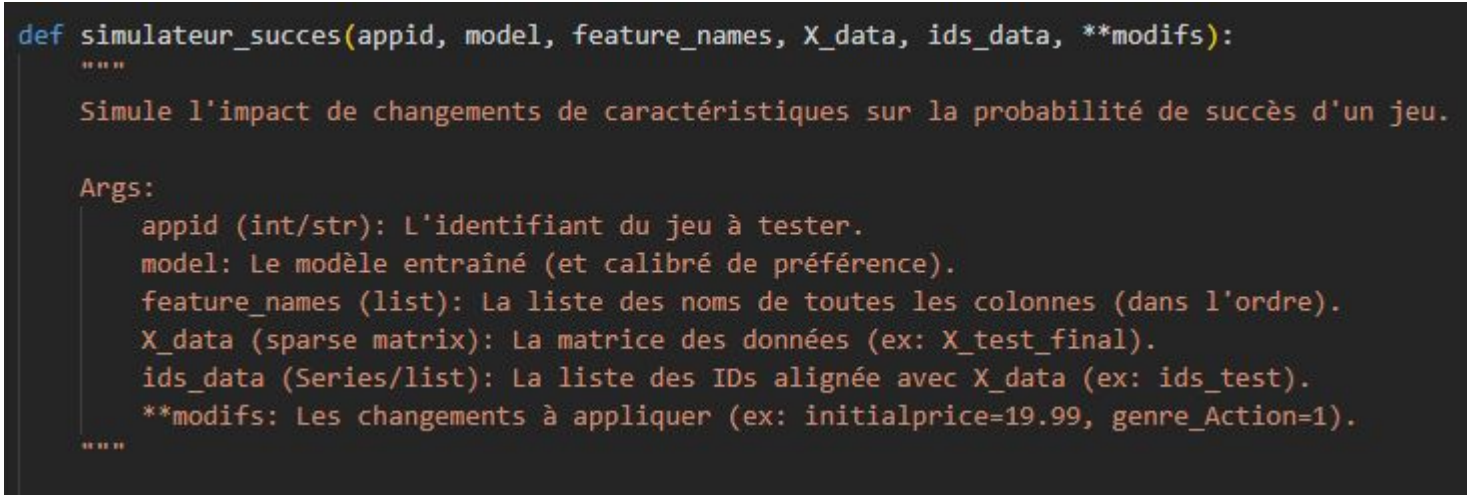
```
xgb_model = XGBClassifier(
    learning_rate=0.1,
    n_jobs=-1,
    random_state=42,
    eval_metric='logloss'    # Métrique d'optimisation
)
```



```
Seuil 0.05 -> Recall: 0.99 | Precision: 0.21 | F2-Score: 0.57
Seuil 0.1 -> Recall: 0.90 | Precision: 0.28 | F2-Score: 0.62
Seuil 0.2 -> Recall: 0.69 | Precision: 0.41 | F2-Score: 0.60
Seuil 0.3 -> Recall: 0.52 | Precision: 0.51 | F2-Score: 0.51
Seuil 0.4 -> Recall: 0.39 | Precision: 0.58 | F2-Score: 0.42
Seuil 0.5 -> Recall: 0.26 | Precision: 0.64 | F2-Score: 0.30
Seuil 0.7 -> Recall: 0.06 | Precision: 0.80 | F2-Score: 0.08
Seuil 0.9 -> Recall: 0.00 | Precision: 1.00 | F2-Score: 0.00
```



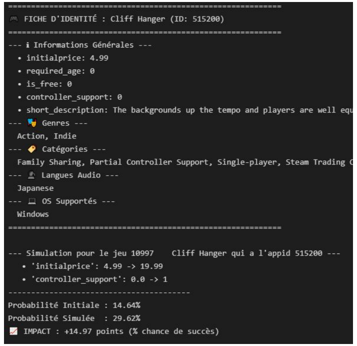
Application : importance des features et interactions entre les features



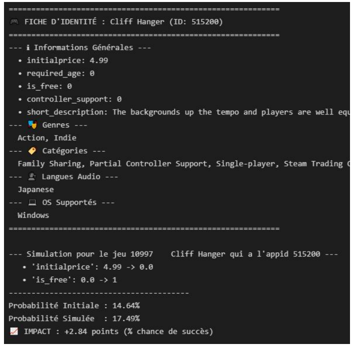
```
def simulateur_succes(appid, model, feature_names, X_data, ids_data, **modifs):
    """
    Simule l'impact de changements de caractéristiques sur la probabilité de succès d'un jeu.

    Args:
        appid (int/str): L'identifiant du jeu à tester.
        model: Le modèle entraîné (et calibré de préférence).
        feature_names (list): La liste des noms de toutes les colonnes (dans l'ordre).
        X_data (sparse matrix): La matrice des données (ex: X_test_final).
        ids_data (Series/list): La liste des IDs alignée avec X_data (ex: ids_test).
        **modifs: Les changements à appliquer (ex: initialprice=19.99, genre_Action=1).
    """
```

Tests de modifications de caractéristiques

Prix

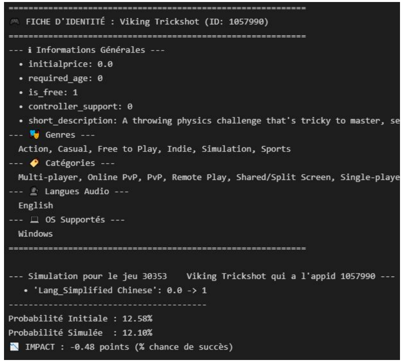


```
============================================================
🎮 FICHE D'IDENTITÉ : Cliff Hanger (ID: 515200)
============================================================
--- ℹ Informations Générales ---
  • initialprice: 4.99
  • required_age: 0
  • is_free: 0
  • controller_support: 0
  • short_description: The backgrounds up the tempo and players are well equ
--- 🎭 Genres ---
  Action, Indie
--- 🏷️ Catégories ---
  Family Sharing, Partial Controller Support, Single-player, Steam Trading C
--- 🗣️ Langues Audio ---
  Japanese
--- 💻 OS Supportés ---
  Windows
============================================================


--- Simulation pour le jeu 10997    Cliff Hanger qui a l'appid 515200 ---
   • 'initialprice': 4.99 -> 19.99
   • 'controller_support': 0.0 -> 1
----------------------------------------
Probabilité Initiale : 14.64%
Probabilité Simulée  : 29.62%
📈 IMPACT : +14.97 points (% chance de succès)
```

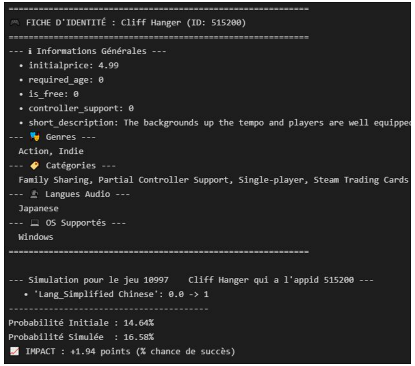


```
============================================================
🎮 FICHE D'IDENTITÉ : Cliff Hanger (ID: 515200)
============================================================
--- ℹ Informations Générales ---
  • initialprice: 4.99
  • required_age: 0
  • is_free: 0
  • controller_support: 0
  • short_description: The backgrounds up the tempo and players are well equ
--- 🎭 Genres ---
  Action, Indie
--- 🏷️ Catégories ---
  Family Sharing, Partial Controller Support, Single-player, Steam Trading C
--- 🗣️ Langues Audio ---
  Japanese
--- 💻 OS Supportés ---
  Windows
============================================================


--- Simulation pour le jeu 10997    Cliff Hanger qui a l'appid 515200 ---
   • 'initialprice': 4.99 -> 0.0
   • 'is_free': 0.0 -> 1
----------------------------------------
Probabilité Initiale : 14.64%
Probabilité Simulée  : 17.49%
📈 IMPACT : +2.84 points (% chance de succès)
```

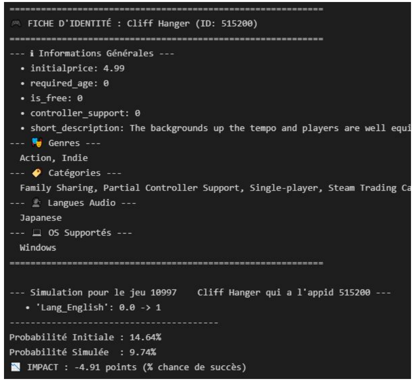
Langues



```
============================================================
🎮 FICHE D'IDENTITÉ : Viking Trickshot (ID: 1057990)
============================================================
--- ℹ Informations Générales ---
  • initialprice: 0.0
  • required_age: 0
  • is_free: 1
  • controller_support: 0
  • short_description: A throwing physics challenge that's tricky to master, se
--- 🎭 Genres ---
  Action, Casual, Free to Play, Indie, Simulation, Sports
--- 🏷️ Catégories ---
  Multi-player, Online PvP, PvP, Remote Play, Shared/Split Screen, Single-playe
--- 🗣️ Langues Audio ---
  English
--- 💻 OS Supportés ---
  Windows
============================================================


--- Simulation pour le jeu 30353    Viking Trickshot qui a l'appid 1057990 ---
   • 'Lang_Simplified Chinese': 0.0 -> 1
----------------------------------------
Probabilité Initiale : 12.58%
Probabilité Simulée  : 12.10%
📉 IMPACT : -0.48 points (% chance de succès)
```

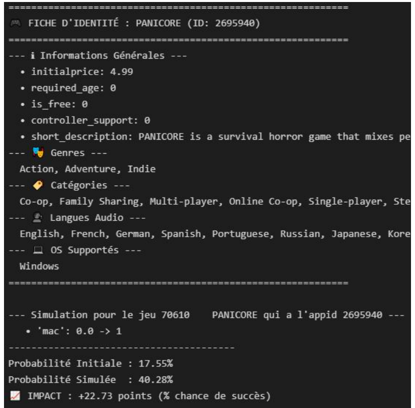


```
============================================================
🎮 FICHE D'IDENTITÉ : Cliff Hanger (ID: 515200)
============================================================
--- ℹ Informations Générales ---
  • initialprice: 4.99
  • required_age: 0
  • is_free: 0
  • controller_support: 0
  • short_description: The backgrounds up the tempo and players are well equipped
--- 🎭 Genres ---
  Action, Indie
--- 🏷️ Catégories ---
  Family Sharing, Partial Controller Support, Single-player, Steam Trading Cards
--- 🗣️ Langues Audio ---
  Japanese
--- 💻 OS Supportés ---
  Windows
============================================================


--- Simulation pour le jeu 10997    Cliff Hanger qui a l'appid 515200 ---
   • 'Lang_Simplified Chinese': 0.0 -> 1
----------------------------------------
Probabilité Initiale : 14.64%
Probabilité Simulée  : 16.58%
📈 IMPACT : +1.94 points (% chance de succès)
```

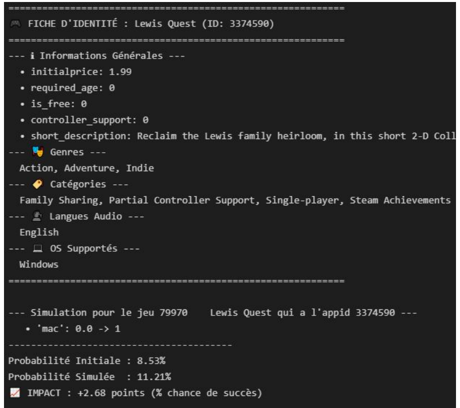


```
============================================================
🎮 FICHE D'IDENTITÉ : Cliff Hanger (ID: 515200)
============================================================
--- ℹ Informations Générales ---
  • initialprice: 4.99
  • required_age: 0
  • is_free: 0
  • controller_support: 0
  • short_description: The backgrounds up the tempo and players are well equi
--- 🎭 Genres ---
  Action, Indie
--- 🏷️ Catégories ---
  Family Sharing, Partial Controller Support, Single-player, Steam Trading Ca
--- 🗣️ Langues Audio ---
  Japanese
--- 💻 OS Supportés ---
  Windows
============================================================


--- Simulation pour le jeu 10997    Cliff Hanger qui a l'appid 515200 ---
   • 'Lang_English': 0.0 -> 1
----------------------------------------
Probabilité Initiale : 14.64%
Probabilité Simulée  : 9.74%
📉 IMPACT : -4.91 points (% chance de succès)
```

Plateforme



```
============================================================
🎮 FICHE D'IDENTITÉ : PANICORE (ID: 2695940)
============================================================
--- ℹ Informations Générales ---
  • initialprice: 4.99
  • required_age: 0
  • is_free: 0
  • controller_support: 0
  • short_description: PANICORE is a survival horror game that mixes pe
--- 🎭 Genres ---
  Action, Adventure, Indie
--- 🏷️ Catégories ---
  Co-op, Family Sharing, Multi-player, Online Co-op, Single-player, Ste
--- 🗣️ Langues Audio ---
  English, French, German, Spanish, Portuguese, Russian, Japanese, Kore
--- 💻 OS Supportés ---
  Windows
============================================================


--- Simulation pour le jeu 70610    PANICORE qui a l'appid 2695940 ---
   • 'mac': 0.0 -> 1
----------------------------------------
Probabilité Initiale : 17.55%
Probabilité Simulée  : 40.28%
📈 IMPACT : +22.73 points (% chance de succès)
```



```
============================================================
🎮 FICHE D'IDENTITÉ : Lewis Quest (ID: 3374590)
============================================================
--- ℹ Informations Générales ---
  • initialprice: 1.99
  • required_age: 0
  • is_free: 0
  • controller_support: 0
  • short_description: Reclaim the Lewis family heirloom, in this short 2-D Coll
--- 🎭 Genres ---
  Action, Adventure, Indie
--- 🏷️ Catégories ---
  Family Sharing, Partial Controller Support, Single-player, Steam Achievements
--- 🗣️ Langues Audio ---
  English
--- 💻 OS Supportés ---
  Windows
============================================================


--- Simulation pour le jeu 79970    Lewis Quest qui a l'appid 3374590 ---
   • 'mac': 0.0 -> 1
----------------------------------------
Probabilité Initiale : 8.53%
Probabilité Simulée  : 11.21%
📈 IMPACT : +2.68 points (% chance de succès)
```

Merci pour votre écoute

Code disponible ici :

https://github.com/chemsss/steam-app-data/

Questions ?