PORTFOLIO

DATA SCIENCE

NASSER CHAOUCHI

SUMMARY

1. **WHO AM I?**
2. **MY WORK EXPERIENCE**
3. **MY PROJECTS**
    1. THE MOVIE RECOMMENDER SYSTEM
    2. MULTICLASS CLASSIFICATION FOR DIABETES
    3. THE CKD AND DYALISIS PREDICTION

SUMMARY

📌 My academic journey

I'm Nasser, a **French computer science engineer** passionate about **artificial intelligence**, **data**, and **innovation**. I **graduated from UTC** (Université de Technologie de Compiègne) in 2025 with **a major in AI and Data Science**.

I completed:

- A **dual-focused internship** at **Numberly** as a **Data Engineer and Project Manager**, combining technical and management responsibilities.
- An **exchange semester** at the **Escuela de Ingeniería y Arquitectura in Zaragoza** (Spain), as part of a **Data Science Master's program**.
- A **final-year internship** at **Ubisoft** as a **Data Scientist**, working on **real-world game data** and **predictive models**.

💪 Strengths

Curious **Rigorous** **Positive** **Patient**

❤️ Interests

Artificial Intelligence **Sports** **Literature** **Chess**

MY WORK EXPERIENCE

An **internship** at **Ubisoft** from **October 2024** to **March 2025** as a **Data Scientist**, with the following missions:

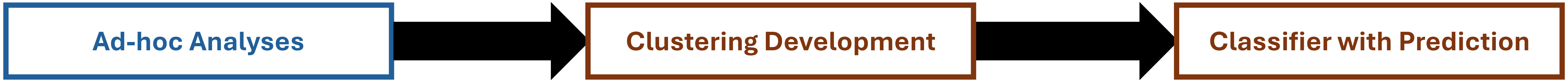
Audiences Understanding

Segmentation Based on Players' Profiles

Player Behavior Prediction

I worked on the game **Avatar: Frontiers of Pandora**. My role was **to understand the game's underperformance** and to **identify and target potential players** within the **Ubisoft ecosystem** who would most likely acquire the game.

The project was divided into **three main phases**:



Ad-hoc Analyses

Clustering Development

Classifier with Prediction

You can contact the team manager for a reference:

- Nicolas Tatin, Associate Director, Data & Analytics

Data Analysis

Data Science

MY PROJECTS – THE MOVIE RECOMMENDER SYSTEM

CONTEXT

- **Dataset**: MovieLens 32M
- **Goal**: Recommend movies users might like, based on behavior and content
- **Type**: Hybrid Recommendation System
    - Collaborative Filtering (ratings)
    - Content-Based Filtering (genres, titles)
- **Size**: 32M+ ratings, ~270k users, 62k movies

BUILT WITH

- Scikit-learn
- Pandas, NumPy
- Seaborn
- Matplotlib
- HuggingFace datasets
- Streamlit

Open the repository

Open the interface
(with MovieLens 1M)

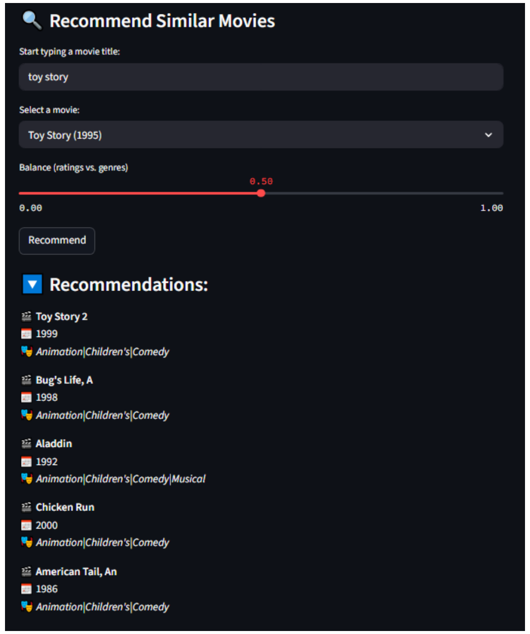
APPROACH

- **Data Cleaning**: Merged movies.csv and ratings.csv, extracted year, processed genres
- **Collaborative Filtering**: Built user-item matrix, applied cosine similarity
- **Content-Based Filtering**: Used TF-IDF/CountVectorizer on genres and titles
- **Hybrid Strategy**: Combined top recommendations from both approaches
- **Implemented multiple strategies**: Most rated movies, Top-rated by genre, Top-rated by year, User-user collaborative hybrid, Item-item collaborative hybrid
- **Profile-Based Recommendation**: Built a user profile from favorite movies to generate personalized suggestions

WHAT I LEARNED

- **Designing** and **comparing recommender strategies**
- Using **similarity metric**s (cosine) on **sparse data**
- **Evaluating trade-offs** between **relevance** and **diversity**

MY PROJECTS – THE MOVIE RECOMMENDER SYSTEM



Recommend Similar Movies

Start typing a movie title:

toy story

Select a movie:

Toy Story (1995)

Balance (ratings vs. genres)

0.50

0.00 1.00

Recommend

Recommendations:

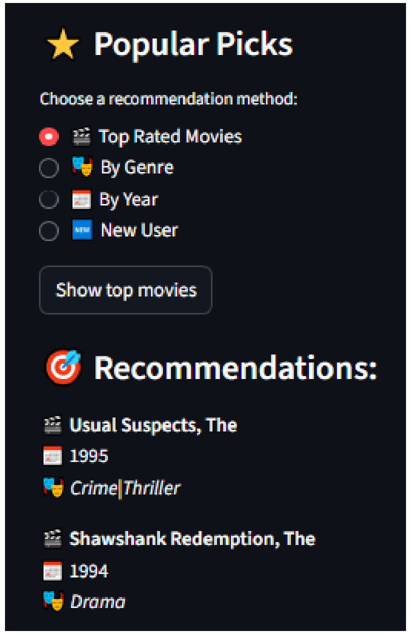
Toy Story 2
1999
Animation|Children's|Comedy

Bug's Life, A
1998
Animation|Children's|Comedy

Aladdin
1992
Animation|Children's|Comedy|Musical

Chicken Run
2000
Animation|Children's|Comedy

American Tail, An
1986
Animation|Children's|Comedy



Popular Picks

Choose a recommendation method:

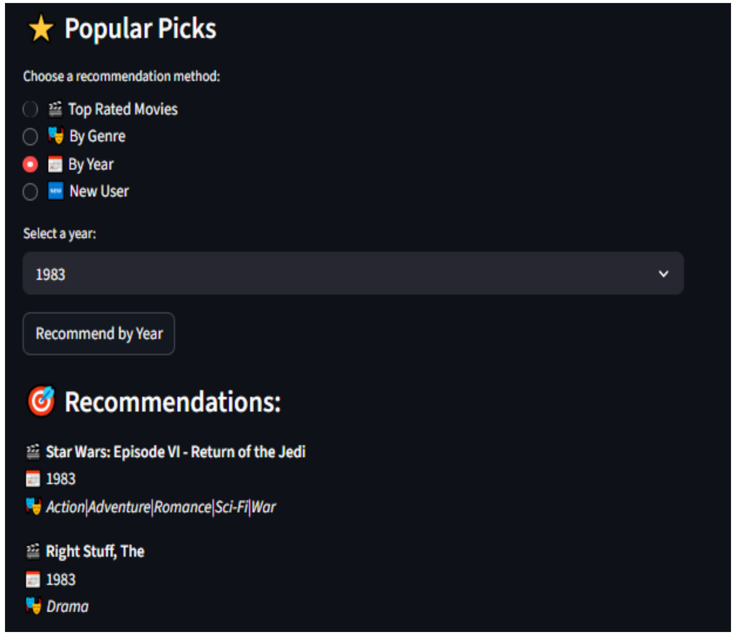
● Top Rated Movies
○ By Genre
○ By Year
○ New User

Show top movies

Recommendations:

Usual Suspects, The
1995
Crime|Thriller

Shawshank Redemption, The
1994
Drama



Popular Picks

Choose a recommendation method:

○ Top Rated Movies
○ By Genre
● By Year
○ New User

Select a year:

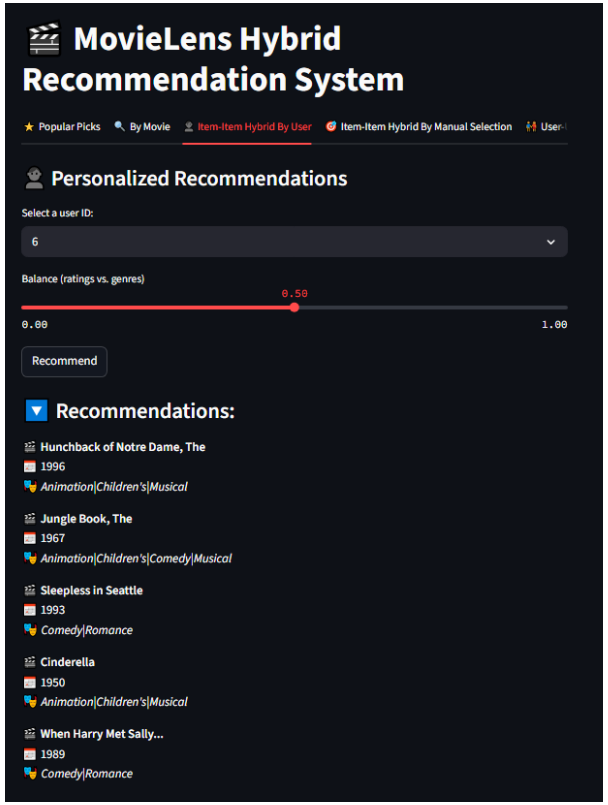
1983

Recommend by Year

Recommendations:

Star Wars: Episode VI - Return of the Jedi
1983
Action|Adventure|Romance|Sci-Fi|War

Right Stuff, The
1983
Drama



MovieLens Hybrid Recommendation System

Personalized Recommendations

Select a user ID:

6

Balance (ratings vs. genres)

0.50

0.00 1.00

Recommend

Recommendations:

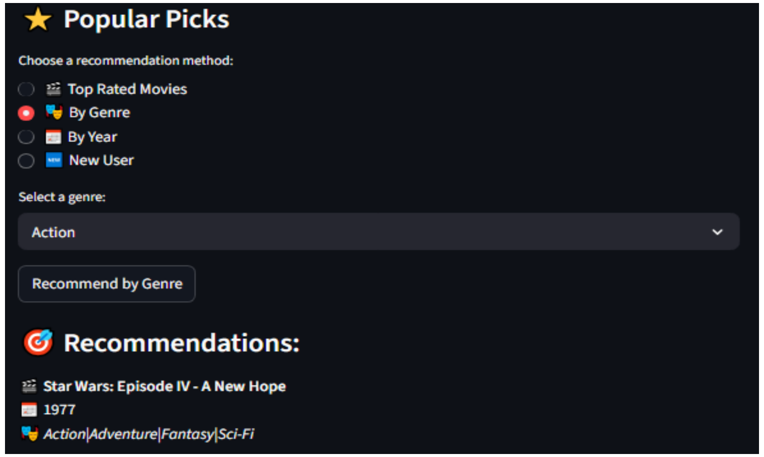
Hunchback of Notre Dame, The
1996
Animation|Children's|Musical

Jungle Book, The
1967
Animation|Children's|Comedy|Musical

Sleepless in Seattle
1993
Comedy|Romance

Cinderella
1950
Animation|Children's|Musical

When Harry Met Sally...
1989
Comedy|Romance



Popular Picks

Choose a recommendation method:

○ Top Rated Movies
● By Genre
○ By Year
○ New User

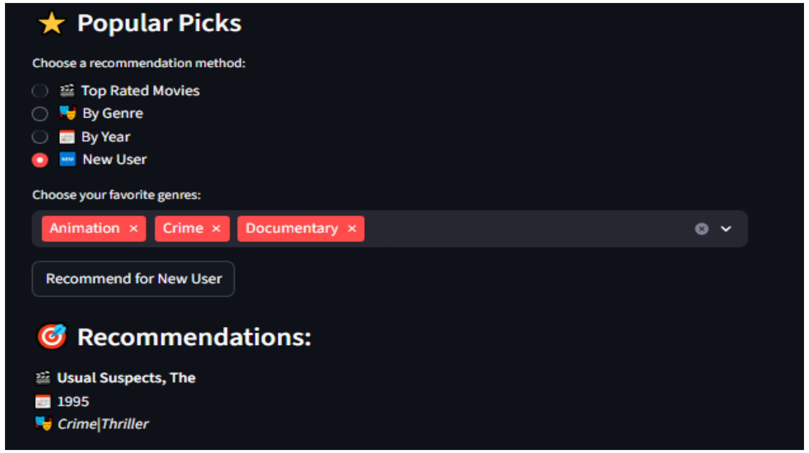
Select a genre:

Action

Recommend by Genre

Recommendations:

Star Wars: Episode IV - A New Hope
1977
Action|Adventure|Fantasy|Sci-Fi



Popular Picks

Choose a recommendation method:

○ Top Rated Movies
○ By Genre
○ By Year
● New User

Choose your favorite genres:

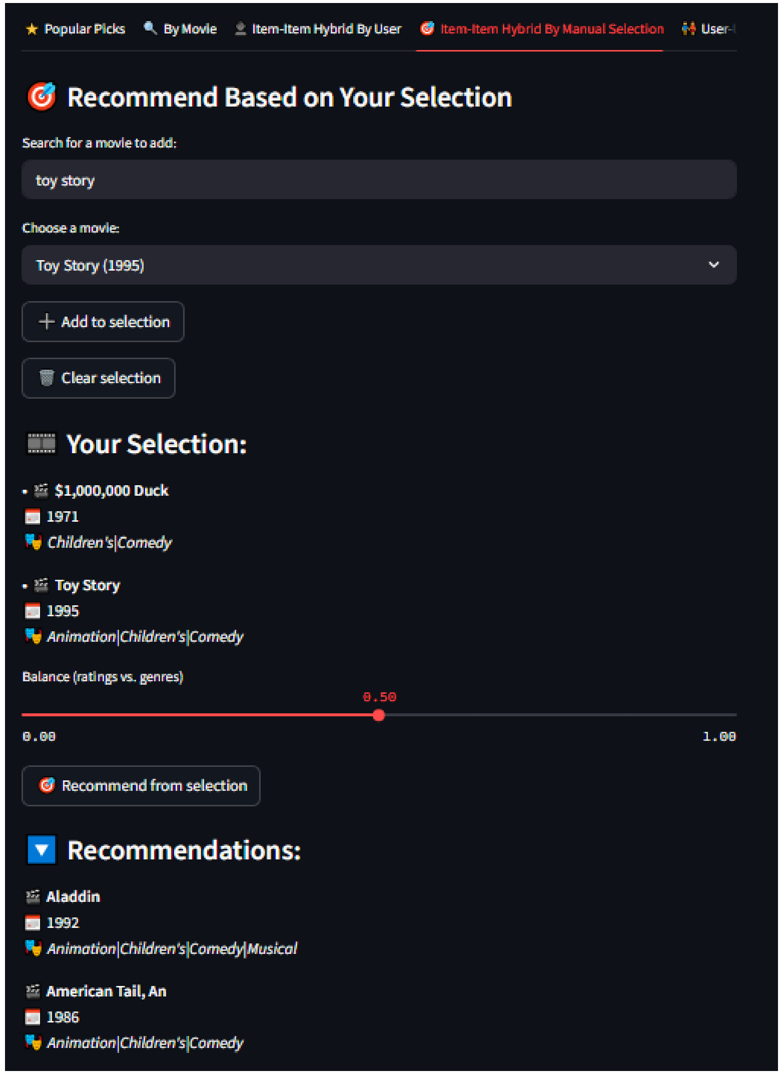
Animation × Crime × Documentary ×

Recommend for New User

Recommendations:

Usual Suspects, The
1995
Crime|Thriller

MY PROJECTS – THE MOVIE RECOMMENDER SYSTEM



Recommend Based on Your Selection

Search for a movie to add: toy story

Choose a movie: Toy Story (1995)

Your Selection:

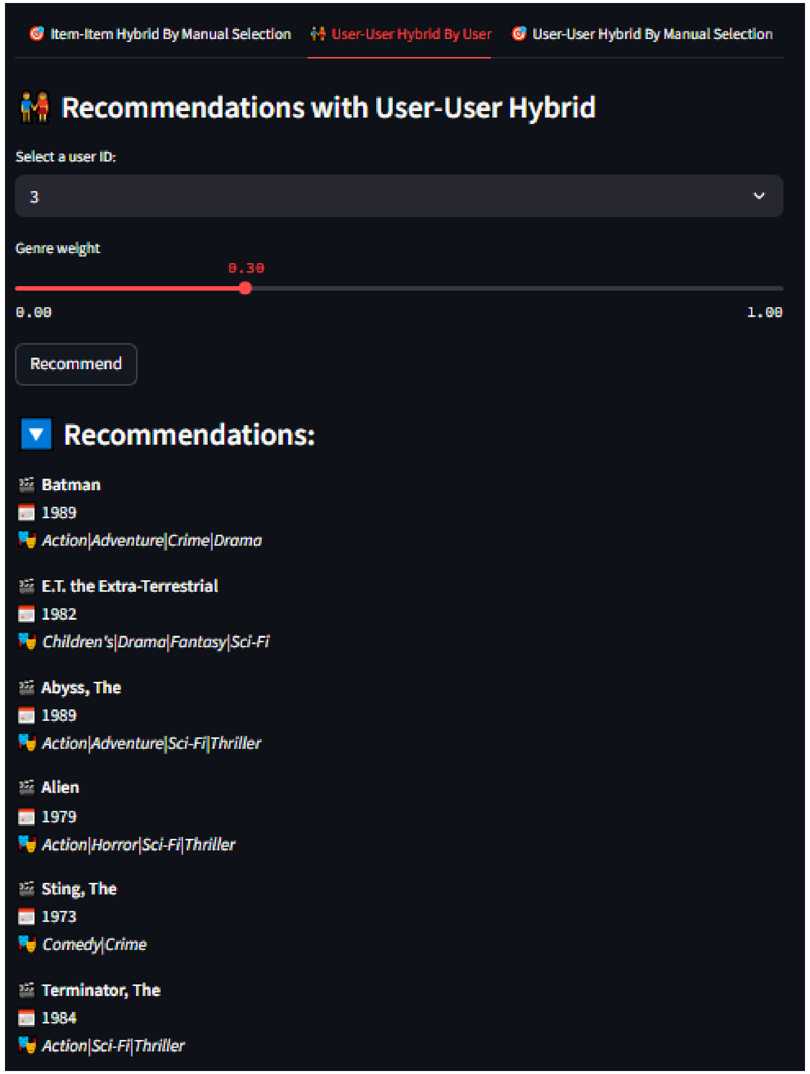
- $1,000,000 Duck
  1971
  Children's|Comedy
- Toy Story
  1995
  Animation|Children's|Comedy

Balance (ratings vs. genres): 0.50 (0.00 – 1.00)

Recommendations:

Aladdin
1992
Animation|Children's|Comedy|Musical

American Tail, An
1986
Animation|Children's|Comedy


Item-Item Hybrid By Manual Selection
User-User Hybrid By User
User-User Hybrid By Manual Selection
Recommend

Recommendations with User-User Hybrid

Select a user ID: 3

Genre weight: 0.30 (0.00 – 1.00)

Recommendations:

Batman
1989
Action|Adventure|Crime|Drama

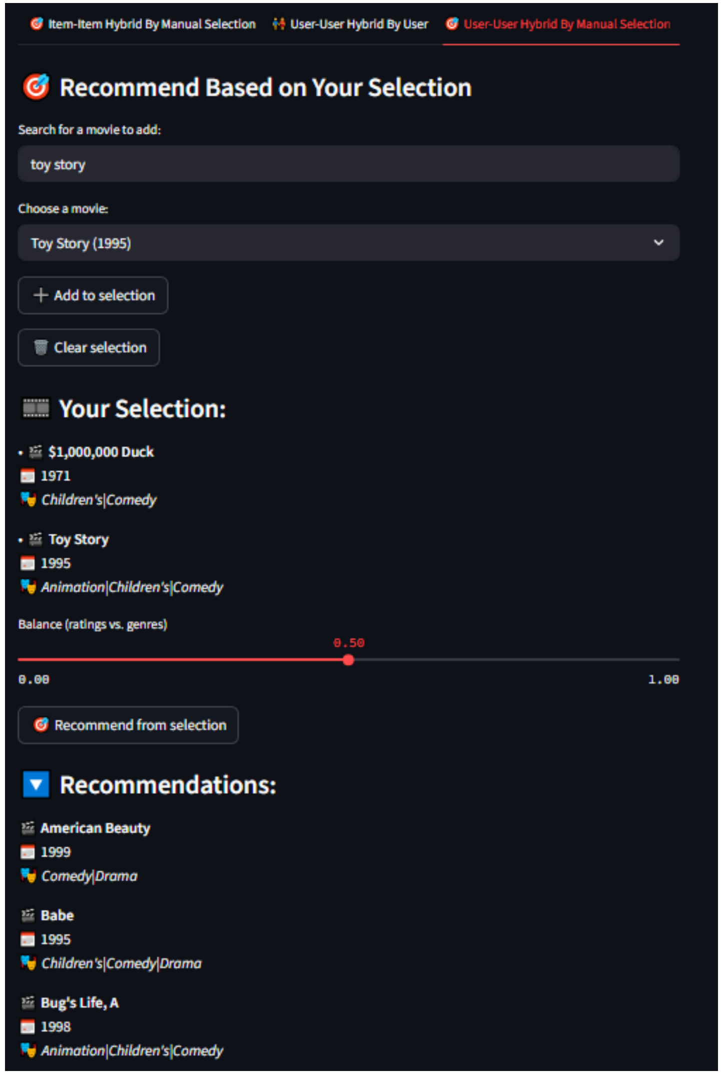
E.T. the Extra-Terrestrial
1982
Children's|Drama|Fantasy|Sci-Fi

Abyss, The
1989
Action|Adventure|Sci-Fi|Thriller

Alien
1979
Action|Horror|Sci-Fi|Thriller

Sting, The
1973
Comedy|Crime

Terminator, The
1984
Action|Sci-Fi|Thriller



Recommend Based on Your Selection

Search for a movie to add: toy story

Choose a movie: Toy Story (1995)

Your Selection:

- $1,000,000 Duck
  1971
  Children's|Comedy
- Toy Story
  1995
  Animation|Children's|Comedy

Balance (ratings vs. genres): 0.50 (0.00 – 1.00)

Recommendations:

American Beauty
1999
Comedy|Drama

Babe
1995
Children's|Comedy|Drama

Bug's Life, A
1998
Animation|Children's|Comedy

MY PROJECTS – MULTICLASS CLASSIFICATION FOR

CONTEXT

- **Dataset**: Multiclass Diabetes Dataset
- **Goal**: Classify patients into several diabetes stages
- **Type**: Supervised, Multiclass classification
- **Size**: 264 patients, 12 features

BUILT WITH

- Scikit-learn
- Pandas, NumPy
- Seaborn
- Matplotlib
- Streamlit

Open the repository

Open the interface

APPROACH

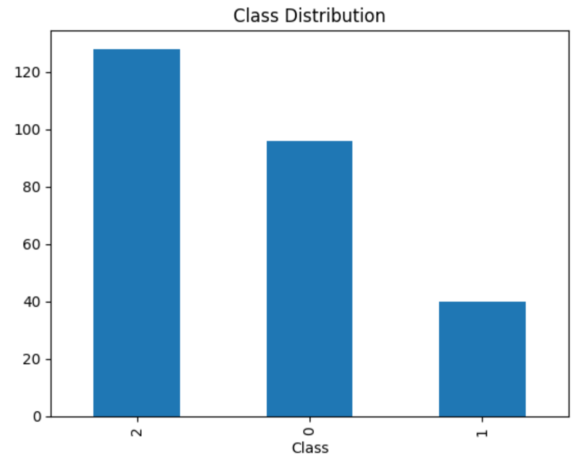
- **EDA & Preprocessing**: Analyzed feature distributions, handled missing values, balanced classes, and scaled data.
- **Model tested**: Logistic Regression, Random Forest and K-Nearest Neighbour
- **Cross Validation**: Ensured robust performance and avoided overfitting
- **Evaluation**: Confusion Matrix, Classification report (F1-Score, Accuracy, Recall)
- **Final model (Random Forest)**
  - **Accuracy**: 97%
  - **Macro F1-score** (better suited to class imbalance): 0.98

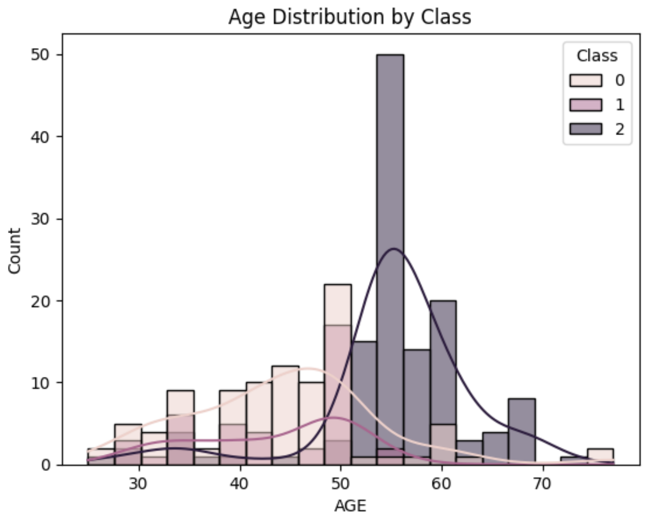
WHAT I LEARNED

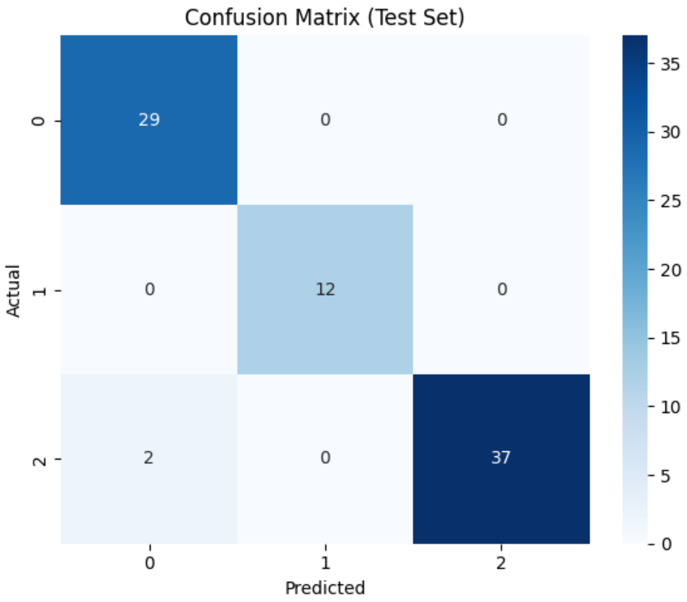
- **How to handle imbalanced multiclass data**
- The **importance** of **feature engineering** and **model tuning**
- Model explainability with **SHAP** or **feature importance**

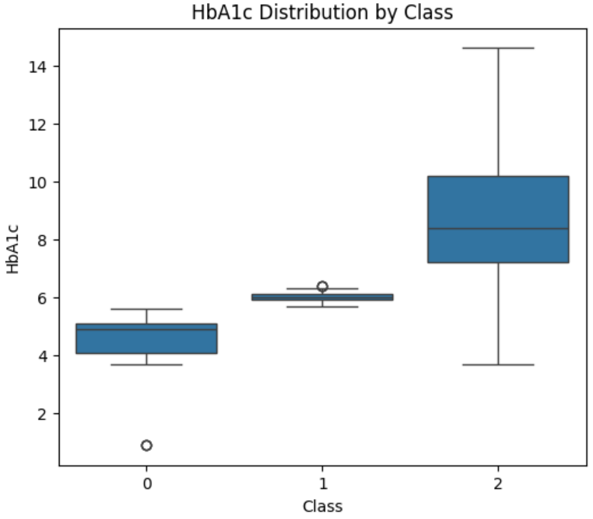
MY PROJECTS – MULTICLASS CLASSIFICATION FOR

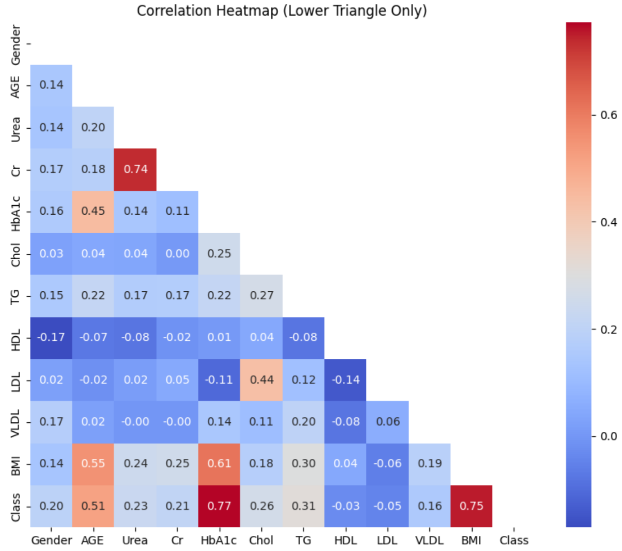
Variable	F-statistic	p-value
AGE	60.368	0.00000
BMI	190.565	0.00000
TG	14.218	0.00000
HbA1c	200.415	0.00000
Chol	9.881	0.00007
Urea	9.115	0.00015
Cr	6.466	0.00182
VLDL	3.712	0.02573
LDL	0.957	0.38526
HDL	0.488	0.61457

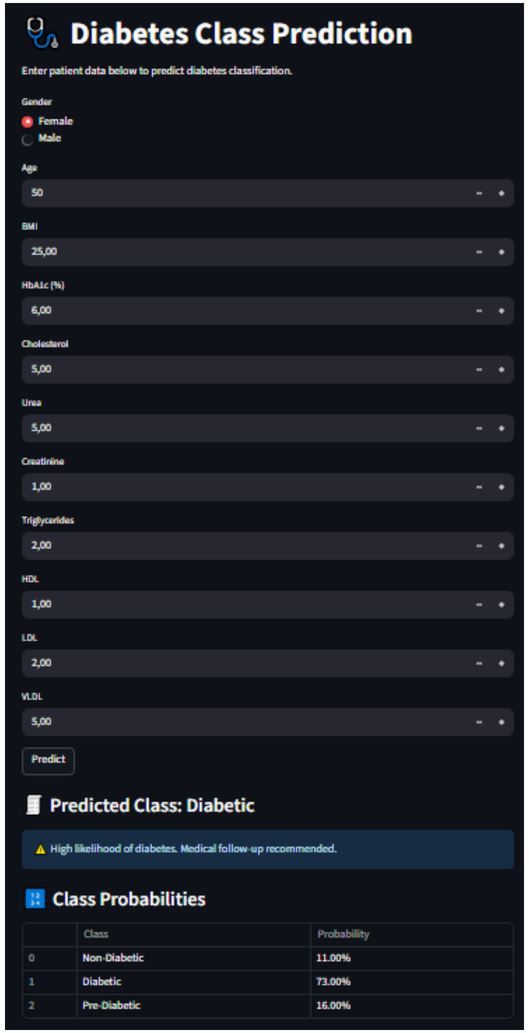











Diabetes Class Prediction

Enter patient data below to predict diabetes classification.

Gender

● Female
○ Male

Age: 50

BMI: 25,00

HbA1c (%): 6,00

Cholesterol: 5,00

Urea: 5,00

Creatinine: 1,00

Triglycerides: 2,00

HDL: 1,00

LDL: 2,00

VLDL: 5,00

Predicted Class: Diabetic

⚠ High likelihood of diabetes. Medical follow-up recommended.

Class Probabilities

	Class	Probability
0	Non-Diabetic	11.00%
1	Diabetic	73.00%
2	Pre-Diabetic	16.00%

MY PROJECTS – THE CKD AND DYALISIS PREDICTION

CONTEXT

- **Dataset**: Kidney Disease Risk Dataset
- **Goal**: Predict CKD status and dialysis need based on clinical and biological data
- **Type**: Supervised, Binary classification (2 targets: CKD_Status, Dialysis_Needed)
- **Size**: 2304 patients, 9 features

BUILT WITH

- Scikit-learn
- Pandas, NumPy
- Seaborn
- Matplotlib
- XGBoost
- Streamlit

Open the repository

Open the interface

APPROACH

- **EDA & Preprocessing**: Explored feature relationships, handled missing values, encoded categorical data, scaled numerical features.
- **Model tested**: Logistic Regression, Random Forest, Gradient Boosting, XGBoost and K-Nearest Neighbour
- **Cross Validation**: Ensured robustness and reduced overfitting risk.
- **Evaluation Classification Report**, **ROC-AUC**, **F1-Score**, **Accuracy**
- Best model was Gradient Boosting, but due to class imbalance, Random Forest gave more reliable results for generalization.

  → **Accuracy**: 100% for CKD_Status — Accuracy 100% but a F1-Score 0.97 (class imbalance) for the Dialysis_Needed

  → **Separate models trained for each target**

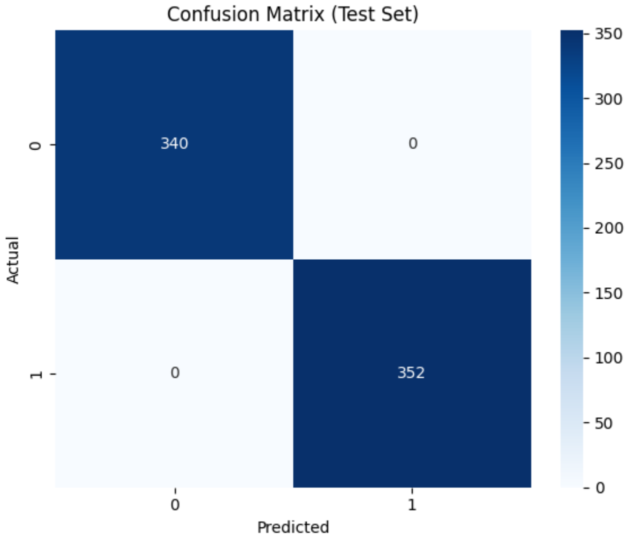
WHAT I LEARNED

- **Managing dual target classification**
- **Handling noisy** and **medical data**
- Improving interpretability with **SHAP values**

MY PROJECTS – THE CKD AND DYALISIS PREDICTION

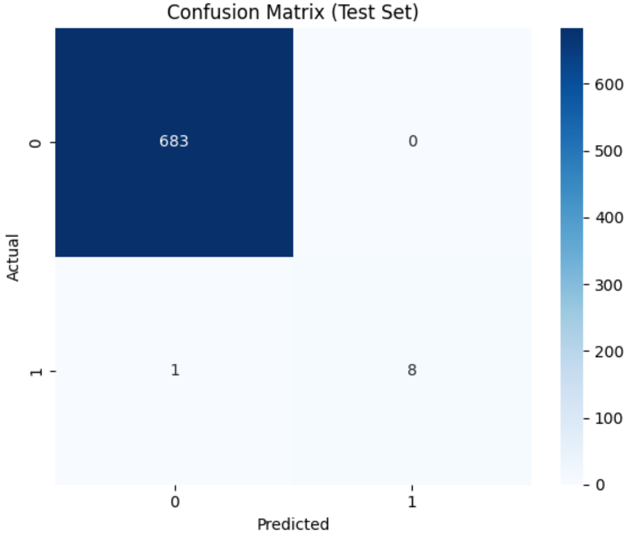
Evaluation on Test Set – CKD_Status

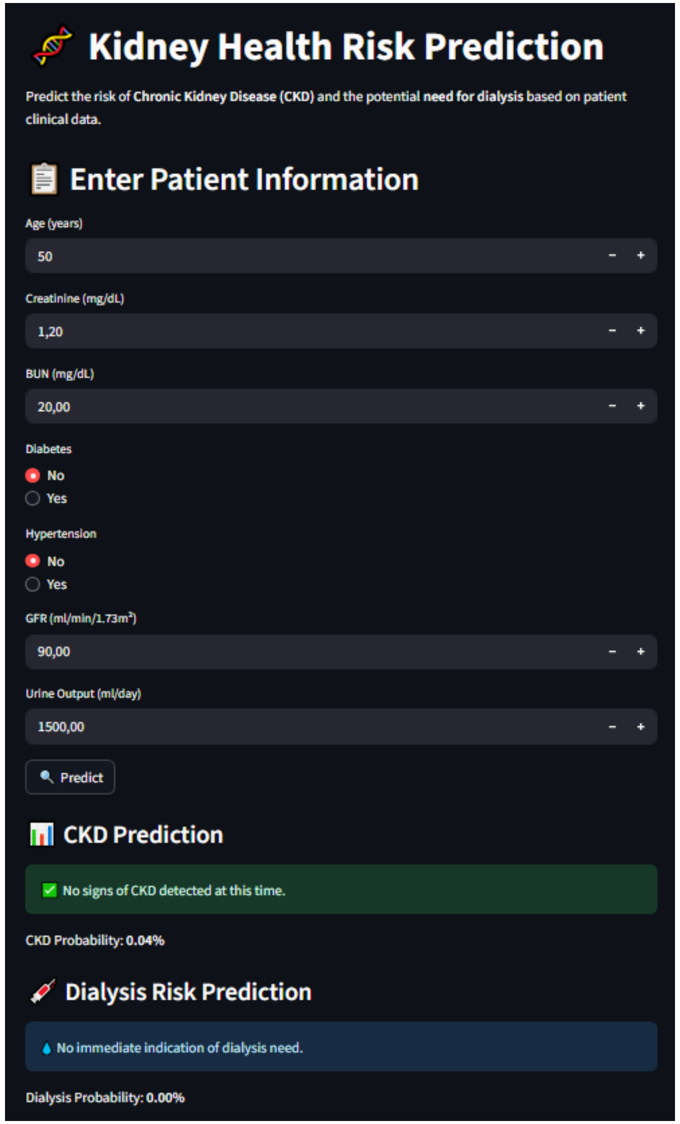
Class	Precision	Recall	F1-Score	Support
0	1.00	1.00	1.00	340
1	1.00	1.00	1.00	352
Accuracy			**1.00**	692
Macro avg	1.00	1.00	1.00	692
Weighted avg	1.00	1.00	1.00	692


Confusion Matrix (Test Set)

Evaluation on Test Set – Dialysis_Needed

Class	Precision	Recall	F1-Score	Support
0	1.00	1.00	1.00	683
1	1.00	0.89	0.94	9
Accuracy			**1.00**	692
Macro avg	1.00	0.94	0.97	692
Weighted avg	1.00	1.00	1.00	692


Confusion Matrix (Test Set)



Kidney Health Risk Prediction

Predict the risk of **Chronic Kidney Disease (CKD)** and the potential **need for dialysis** based on patient clinical data.

Enter Patient Information

Age (years): 50

Creatinine (mg/dL): 1,20

BUN (mg/dL): 20,00

Diabetes: ● No ○ Yes

Hypertension: ● No ○ Yes

GFR (ml/min/1.73m²): 90,00

Urine Output (ml/day): 1500,00

CKD Prediction

✅ No signs of CKD detected at this time.

CKD Probability: 0.04%

Dialysis Risk Prediction

No immediate indication of dialysis need.

Dialysis Probability: 0.00%

DON’T HESITATE TO REACH ME OUT