

EPHEC

Avenue Konrad Adenauer 3  
1200 Bruxelles

DEBONGNIE Nathan  
BDA00350  
3BDAA

# **Analyse et prédition de l'évolution des performances des nageurs Belges**

Travail de fin d'études présenté  
pour l'obtention du diplôme de  
bachelier de spécialisation en Business Data Analysis

Année académique 2024-2025

# I. Avant-propos

## Remerciements

Avant toute chose, je souhaite remercier les personnes qui ont rendu la réalisation de ce travail possible.

Premièrement, je tiens à exprimer ma gratitude à mes proches et amis, dont le soutien moral m'a été précieux tout au long de mes études et de ce travail. Un remerciement particulier à Matthieu Luyckx, Data Scientist chez Jetpack.AI, pour les conseils et son expertise en data mining.

Je remercie également Christian Kaufmann ([results@swimrankings.net](mailto:results@swimrankings.net)) et Swimrankings en général pour les autorisations et l'accessibilité des données utilisées dans le cadre de ce projet.

Déclaration de non plagiat

Année académique 2024-2025

Haute Ecole **EPHEC**

BACHELIER DE  
SPECIALISATION EN  
BUSINESS DATA  
ANALYSIS

Je soussigné(e) *Nathan DEBONGNIE* ,

Étudiant(e) de la 3BDAA (Bachelier spécialisé en Business Data Analysis) , déclare par la présente que le travail ci- joint est exempt de tout plagiat et respecte en tous points le règlement des études en matière d'emprunts, de citations et d'exploitation de sources diverses signé lors de mon inscription à l'EPHEC, ainsi que les instructions et consignes mises à ma disposition sur le Moodle et/ou l'intranet étudiants. Par ma signature, je certifie sur l'honneur avoir pris connaissance des documents précités et que le travail présenté est original et exempt de tout emprunt à un tiers non cité correctement.

Date : 29-05-25

Signature :

DEBONGNIE Nathan

### Déclaration de l'usage de l'IA générative dans le cadre du travail de fin d'études

L'encart suivant doit être complété par l'étudiant et inséré sur la première page de son travail écrit, juste après la page de garde son travail de fin d'études. Attention, il est possible de cocher plusieurs cases.

Tout étudiant qui réalise un travail écrit doit remplir l'encart relatif à l'utilisation de l'IA générative au même titre que la déclaration de plagiat.

Le but n'est pas de dissuader les étudiants de remplir cet encart, c'est-à-dire que si l'étudiant joue le jeu et que son utilisation de l'IA générative est intelligente, ce dernier ne doit pas être sanctionné pour la seule raison d'avoir utilisé une IA générative.

Néanmoins, si un étudiant fait une déclaration mensongère sur son utilisation, par exemple : "j'ai écrit l'intégralité de mon texte sans avoir eu recours à un outil d'IA générative" alors qu'il s'avère qu'il l'a utilisé, celui-ci s'expose à des sanctions. Lors de sa défense orale, l'étudiant pourra être interrogé de manière plus approfondie sur les parties de textes qui semblent poser question. L'étudiant pourrait être sanctionné dans l'hypothèse où les doutes concernant la fausse déclaration seraient confirmés par un examen approfondi.

Dans cette étude, le rôle de l'IA générative a été :

- |  |
|--|
| <input type="checkbox"/> A) J'ai écrit l'intégralité de mon texte sans avoir eu recours à un outil d'IA générative ;   |
| <input type="checkbox"/> B) J'ai rédigé le contenu de mon travail mais j'ai sollicité un outil d'IA générative pour améliorer  |
| <input type="checkbox"/> C) L'orthographe  |
| <input type="checkbox"/> D) La grammaire   |
| <input type="checkbox"/> E) La syntaxe   |
| <input checked="" type="checkbox"/> F) C) J'ai consulté un outil d'IA générative pour m'inspirer et puiser des idées de rédaction au niveau du contenu ou de la structure ;                                    |
| <input checked="" type="checkbox"/> G) D) J'ai construit des idées que j'ai ensuite soumises à un outil d'IA générative qui m'a aidé à développer mon texte sur base de ces idées ;                            |
| <input type="checkbox"/> H) E) J'ai sollicité un outil d'IA générative à des fins de traduction ;  |
| <input type="checkbox"/> I) F) J'ai confronté plusieurs propositions de contenu produit par l'IA générative pour en sélectionner les passages les plus pertinents, et j'ai édité et amélioré pour la plupart ; |
| <input type="checkbox"/> J) G) J'ai édité et amélioré une proposition de contenu produit par l'IA générative ;   |
| <input type="checkbox"/> K) H) Une ou des parties de mon travail ont été intégralement produites au moyen d'un outil d'IA générative sans apport original de ma part.  |

Ci-dessous, il est indiqué ce qui est attendu en matière de référencement, sur base des usages des outils d'IA générative renseignés ci-dessus. Les différentes mentions devront être rassemblées dans une section dédiée (intitulée « Mentions des usages, prompts et productions d'outils d'IA générative dans le cadre de ce travail ») en fin d'étude.

**A) J'ai écrit l'intégralité de mon texte sans avoir eu recours à un outil d'IA générative ;**

⇒ Aucune démarche de référencement de l'outil d'IA générative n'est requise.

**B) J'ai rédigé le contenu de mon travail mais j'ai sollicité un outil d'IA générative pour améliorer :**

- l'orthographe,
- la grammaire,
- Et/ou la syntaxe ;

⇒ Cet usage d'outil d'IA générative ne doit faire l'objet d'aucune démarche de référencement académique.

**C) J'ai consulté un outil d'IA générative pour m'inspirer et puiser des idées de rédaction au niveau du contenu ou de la structure ;**

⇒ Cet usage nécessite de mentionner le ou les outils d'IA générative utilisés en l'indiquant dans un encart dédié à cet effet (voir exemple) après la bibliographie ;  
⇒ En outre, l'étudiant tiendra à disposition des évaluateurs le relevé des échanges avec l'outil d'IA générative et devra les produire en cas de demande explicite.

**D) J'ai construit des idées que j'ai ensuite soumises à un outil d'IA générative qui m'a aidé à développer mon texte sur base de ces idées ;**

⇒ Cet usage nécessite de mentionner le ou les outil(s) d'IA générative utilisé(s) en l'indiquant dans un encart dédié à cet effet (voir exemple) après la bibliographie ;  
⇒ En outre, l'étudiant tiendra à disposition des évaluateurs le relevé des échanges avec l'outil d'IA générative et devra les produire en cas de demande explicite.

**E) J'ai sollicité un outil d'IA générative pour traduire du contenu que j'ai produit dans ma langue maternelle vers une langue-cible ;**

⇒ Cet usage nécessite de mentionner :

- le ou les outil(s) d'IA générative utilisés en l'indiquant dans un encart dédié à cet effet (voir exemple) après la bibliographie ;
- la ou les parties traduites ayant fait l'objet d'une traduction.

**F) J'ai confronté plusieurs propositions de contenu produits par l'IA générative pour en sélectionner les passages les plus pertinents, que j'ai édité et amélioré pour la plupart ;**

⇒ Cet usage nécessite de mentionner :

- le ou les outil(s) d'IA générative utilisés en l'indiquant dans un encart dédié à cet effet (voir exemple) après la bibliographie ;
- les requêtes (*prompts*) utilisés en l'indiquant dans un encart dédié à cet effet (voir exemple) après la bibliographie ;
- l'intégralité des propositions produites par l'outil d'IA générative dans un encart dédié à cet effet (voir exemple) après la bibliographie ;

**G) J'ai édité et amélioré une proposition de contenu produit par l'IA générative ;**

⇒ Cet usage nécessite de mentionner :

- i. le ou les outil(s) d'IA générative utilisés en l'indiquant dans un encart dédié à cet effet (voir exemple) après la bibliographie ;
- ii. les requêtes (*prompts*) utilisés en l'indiquant dans un encart dédié à cet effet (voir exemple) après la bibliographie ;
- iii. l'intégralité des propositions produites par l'outil d'IA générative dans un encart dédié à cet effet (voir exemple) après la bibliographie ;

**H) Une ou des parties de mon travail ont été intégralement produites au moyen d'un outil d'IA générative sans apport original de ma part.**

⇒ Cet usage nécessite de mentionner :

- i. le ou les outil(s) d'IA générative utilisés en l'indiquant dans un encart dédié à cet effet (voir exemple) après la bibliographie ;
- ii. les requêtes (*prompts*) utilisés en l'indiquant dans un encart dédié à cet effet (voir exemple) après la bibliographie ;
- iii. l'intégralité des propositions produites par l'outil d'IA générative dans un encart dédié à cet effet (voir exemple) après la bibliographie ;

## II. Executive summary

Dans le monde du sport, la recherche de performance est un objectif constant. En natation, comprendre la progression des athlètes est essentiel pour optimiser l'évolution d'un ou une sportif.ve.

C'est sur cette logique que ce base l'objectif de ce travail. Les résultats de nombreuses compétitions sportives sont enregistrés, et il est possible d'en déduire un grand nombre d'informations permettant de prédire l'évolution future des performances individuelles.

Le projet est organisé en plusieurs étapes. La collecte automatisée des résultats via du web scraping, la structuration d'une base de données adaptée, l'exploration des données comportant une première phase d'analyse à l'aide de rapports graphiques (Power BI) et finalement le développement d'un modèle permettant de réaliser des prédictions de résultats associé à d'autres dashboards visuels qui complète les analyses.

A l'aide de cet outil, il deviendra plus aisé pour les entraîneurs et les clubs de natation d'adapter leurs programmes d'entraînement et d'identifier les nageurs à haut potentiel rapidement. Les nageurs eux-mêmes pourront profiter également de cet outil pour se projeter vers des objectifs réalisables à court ou moyen terme.

# Table des matières

I.	Avant-propos.....	II
II.	Executive summary.....	VII
III.	Introduction.....	I
IV.	Méthode de travail.....	II
V.	Compréhension métiers .....	III
VI.	Récupération des données.....	IV
a.	Web scraping.....	V
b.	Base de données.....	VII
c.	Serveur.....	VIII
VII.	Compréhension des données.....	IX
a.	Visualisation .....	X
b.	Statistiques descriptives .....	XI
VIII.	Préparation des données .....	XII
a.	ETL.....	XII
b.	Feature engineering .....	XIV
IX.	Modélisation des données.....	XV
a.	Régression polynomiale .....	XVI
b.	Réseau de neurones .....	XVII
c.	Modèle final .....	XIX
d.	Prédictions.....	XXI
X.	Visualisation.....	XXII
a.	Clubs .....	XXII
b.	Nageurs.....	XXIII
XI.	Conclusion.....	XXIV
a.	Evaluation.....	XXIV
b.	Pistes d'amélioration .....	XXIV
c.	Difficultés rencontrées .....	XXV
XII.	Bibliographie.....	XXVI

### III. Introduction

Depuis très jeune, le sport, et plus particulièrement la natation, occupe une place importante dans ma vie. Des entraînements presque quotidiens et des compétitions régulières toujours à la recherche de la meilleure performance possible. En parallèle de cela, je me suis découvert un attrait particulier pour les données. Ce travail de fin d'études est une opportunité idéale pour lier les 2.

La performance sportive est au cœur de toutes les disciplines sportives compétitives. Outre la réflexion sur les méthodologies d'entraînements, la nutrition ou l'équipement, certains sports intègrent une notion d'analyse de données pour améliorer les performances. Ce n'est pas encore vraiment le cas dans la natation, alors que beaucoup de données existent. Le site [Swimrankings.net](http://Swimrankings.net) par exemple, reprends la majorité des résultats de natation dans divers pays. Ces informations pourraient-elles permettre de prévoir l'évolution d'un nageur ?

L'objectif principal de ce projet est donc de proposer un outil capable d'analyser les performances passées et de prédire l'évolution d'un nageur sur une course spécifique. Cet outil pourrait s'avérer utile pour tous les acteurs du sport de nage, que ce soit pour identifier les futurs talents, ou accompagner les nageurs vers des objectifs réalisables.

Pour mener à bien ce projet, il est organisé en plusieurs étapes. Une fois les objectifs clairement définis, la première étape est de comprendre le cadre et les besoins spécifiques au bon déroulement du projet. La partie suivante consiste à récupérer toutes les données utiles. Sur base de ces données et du sujet du projet, une analyse approfondie des données est nécessaire. Cette compréhension des données permet par la suite de réaliser les modifications nécessaires afin qu'elles soient propres et utilisables pour construire un modèle de prédictions.

Un des principaux défis du projet est la mise en pratique d'un modèle de data mining, d'autant plus dans un cadre aussi spécifique que celui-ci. En effet, il s'avère que faire de la prédition sur des résultats sportifs n'est pas une tâche facile.

Ce document détaille les étapes ayant été nécessaires pour parvenir au résultat final du projet. En partant du constat qu'il devait être possible d'apporter une analyse constructive des données de résultats de compétitions, proposer un dashboard de visualisation pouvant aider les nageurs et les clubs pour comprendre et suivre les performances ainsi qu'un modèle prédictif de résultats pour pouvoir fixer des objectifs réalisables.

## IV. Méthode de travail

Réaliser un projet aussi important nécessite une bonne organisation. Pour cela, je me suis basé sur la méthode « Crisp DM », qui permet de structurer l'approche d'un projet data. Il est composé de 6 phases bien définies, qui sont toutes aussi importantes que les autres. Le démarrage d'un projet data passe par une phase d'analyse et de réflexion. L'objectif est de déterminer la structure organisationnelle du projet, de définir le cadre et les objectifs SMART et de comprendre le contexte dans lequel le projet s'inscrit. La seconde partie est une continuité et est en lien rapproché avec la phase de compréhension métiers. La compréhension des données est une phase cruciale et souvent trop peu prise en compte à l'entame d'un projet. Elle permet d'étoffer sa compréhension du projet et d'identifier des informations clés supplémentaires. Récupérer des statistiques descriptives des variables et une première visualisation des valeurs sont les 2 techniques généralement mises en place. Il est parfois nécessaire de revenir à la compréhension métier pour s'assurer des informations extraites des données. Lorsque toutes les informations nécessaires au bon déroulement du projet sont récupérées, la partie technique peut commencer. Pour commencer, une phase de préparation des données doit être réalisée. Cette phase consiste à transformer toutes les données afin de pouvoir les utiliser dans la modélisation. Que ce soit de la discrétisation de variables quantitatives, la normalisation, la standardisation ou la création de nouvelles variables. Il y a souvent une grande quantité de travail, et cette partie est souvent reprise et modifiée en fonction des résultats obtenus lors de la phase d'après, la modélisation. La modélisation consiste à appliquer des techniques statistiques aux données. Le choix du modèle dépend de l'objectif et du type de données utilisées. L'étape implique de nombreuses itérations, ou parfois la phase de préparation de données est reprise également. Ensuite, une fois que le modèle développé est suffisamment performant, vient l'étape d'évaluation. Il s'agit de vérifier que les objectifs définis plus tôt ont été atteints. C'est ici qu'une conclusion est développée. La phase finale consiste à déployer le modèle et à finaliser le projet au moyen d'un rapport complet rendu au client.

Ce projet comporte cependant quelques nuances. Tout d'abord, une étape supplémentaire est nécessaire. Les données ne sont pas accessibles immédiatement et nécessitent une étape de récupération des données. Elles sont accessibles en ligne et doivent être récupérées à l'aide d'un outil de scraping. Ensuite, toute une dimension d'analyse et de visualisation est intégrée au projet. Cette partie est présente à la fois dans les premières phases, pour aider à la compréhension des données, mais aussi dans la phase finale pour présenter des résultats. Pour les phases d'évaluation et de déploiement, il y a également de petites différences. L'évaluation passe aussi par la proposition de visuels dans un dashboard Power BI pour conclure ce projet, cette partie étant l'un des objectifs du projet. Pour le déploiement, il s'agit ici uniquement de remettre le projet pour l'évaluation finale, ce qui ne nécessite pas de mise en place et de communication particulière.

## V. Compréhension métiers

La natation et le monde sportif en général sont des domaines assez complexes où chaque détail compte dans l'objectif d'améliorer les performances. Que ce soit dans le très haut niveau, où les gains marginaux font la différence, ou à d'autres niveaux moins exigeants, car la motivation d'avoir des objectifs réalisables y est précieuse. Beaucoup d'investissements sont faits pour améliorer les équipements et la gestion des entraînements, mais il manque encore d'outils permettant de visualiser et anticiper les performances des nageurs. Pour pallier cela, une piste intéressante est envisageable dans l'analyse des résultats, accessibles en grande quantité en ligne sur le site de [swimrankings](#). Analyser ces résultats pourrait apporter une vision globale sur l'évolution des performances de natation et améliorer le suivi des nageurs.

Les objectifs définis sont les suivants :

1. Créer et modéliser une base de données pour enregistrer les données
  - a. Des nageurs : nom, prénom, club, sexe, année de naissance, nationalité
  - b. Des résultats : nageur, distance, style, piscine (50m ou 25m), temps, points FINA, date, lieu
2. Construire un flux ETL pour enregistrer les propres données dans la base de données
  - a. Retirer les colonnes inutiles
  - b. Transformer les types de données
  - c. Effacer les données incomplètes
3. Proposer un dashboard à l'aide d'un outil BI pour présenter les performances des membres d'un club
  - a. Voir les performances collectives pour identifier des améliorations possibles dans la planification des entraînements
  - b. Voir les performances individuelles des membres pour suivre leur évolution et voir des objectifs réalisables à court terme
4. Proposer un dashboard à l'aide d'un outil BI pour présenter et comparer les performances de nageurs
  - a. Pour comparer son niveau aux résultats d'autres nageurs belges à chaque âge
  - b. Pour identifier des objectifs à court ou moyen terme réalisables
5. Créer un modèle de prédiction de performances d'un nageur sur une course afin d'identifier des objectifs atteignables
  - a. Choisir la technique d'apprentissage
  - b. Appliquer la technique aux données existantes
  - c. Réaliser des prédictions sur des nouvelles données

Les autres parties de la mise en place d'une structure organisationnelle du projet ne sont pas réellement d'application pour ce travail de fin d'études. Aucune autre personne n'est impliquée, que ce soit dans l'organisation ou dans la réalisation. Cela a des avantages, comme la réduction des dépendances et du temps nécessaire à organiser la communication. À l'inverse, cela peut rendre l'analyse incomplète et rendre la compréhension plus difficile si les connaissances dans le domaine sont insuffisantes.

## VI. Récupération des données

Cette étape est une étape intermédiaire qui, en général, n'a pas lieu d'être dans un projet data. La plupart du temps, les données sont déjà enregistrées au sein de l'entreprise du client, et seul l'accès à celles-ci est nécessaire. Dans certains cas, comme ici, les données ne sont pas à disposition immédiate dans une base de données ou un datawarehouse. Elles n'appartiennent pas à une entreprise spécifique, mais sont accessibles publiquement sur le site internet [www.swimrakings.net](http://www.swimrakings.net). Il faut donc passer par un outil de Web Scraping.

Belgium Messieurs, Cat. générale			100m Libre						Grand bassin (50m) Meilleurs à vie		
Dernière mise à jour 11 mai 2025 - 0:01											
Afficher les meilleurs temps Premier <<250 <<50 <<25 Depuis le rang: 1			GO 25>> 50>> 250>> Rangs de 1 à 9035								
Nom, Prénom	A.n.	nat.	Club	Temps	Pts.	Rang dans class.	Date	Ville (Pays)	BEL	Région	Club
TIMMERS, Pieter	1988	BEL	Zwemclub Brabo Antwerpen	47.80	942	1. 1. 1.	10 aoû 2016	Rio (BRA)			
SURGELOOSE, Glenn	1989	BEL	Zwemclub Brabo Antwerpen	48.64	894	2. 2. 2.	19 mai 2016	London (GBR)			
DEKONINCK, Dieter	1991	BEL	Zwemclub Brabo Antwerpen	48.79	885	3. 3. 3.	29 jul 2012	London (GBR)			
GRANDJEAN, Yoris	1989	BEL	Liege Natation	48.82	884	4. 1. 1.	12 aoû 2008	Beijing (CHN)			
HENVEAUX, Lucas	2000	BEL	Liege Natation	48.96	876	5. 2. 2.	NEW 25 avr 2025	Antwerpen			
AERENTS, Jasper	1992	BEL	Koninklijke Brugse Zwem- & Reddingskring	49.08	870	6. 1. 1.	28 fév 2016	Antwerpen			

Figure 1 : Swimstats - Résultats Hommes Belge

La récupération des données est une tâche compliquée qui nécessite une compréhension des langages web (HTML et CSS principalement) et une solution pour récupérer ces données. Pour cela, il est possible de passer par des outils, parfois payants, déjà existants. L'avantage de ces applications est qu'il suffit de les paramétriser légèrement pour automatiser la récupération des données, et tout cela sans avoir besoin de comprendre réellement la logique du développement web. Ils ne permettent cependant pas une aussi grande liberté qu'en utilisant des librairies Python, qui servent de base pour décomposer et identifier les zones de données dans une page internet. L'adaptabilité et la souplesse de Python en font une solution bien plus appropriée pour ce travail. L'automatisation à l'aide de librairies comme 'Requests' et 'BeautifulSoup' permet de récupérer les données sur un grand nombre de pages web et offre une meilleure gestion des erreurs.

Toutes ces données doivent ensuite être enregistrées dans une base de données afin de permettre leur exploitation structurée. Pour cela, une phase de modélisation préalable est nécessaire afin de concevoir un schéma logique et normalisé. L'outil Draw.io a été utilisé pour créer le diagramme du modèle relationnel. Sur base de ce schéma, plusieurs technologies peuvent être utilisées dans la mise en place de la base de données. Microsoft SQL Server est une solution intéressante et cohérente, notamment en raison de sa complémentarité avec SQL Server Integration Services

(SSIS), un outil d'ETL également développé par Microsoft. Ce dernier sera utilisé pour automatiser la transformation et le chargement des données brutes vers la base de données modélisée et normalisée. C'est finalement PostgreSQL qui a été utilisé dans le cadre de ce projet. Cette technologie open source présente une prise en main plus simple et une plus grande flexibilité. Elle est également bien adaptée à un déploiement sur serveur, ce qui facilite son intégration dans un environnement distant, nécessaire pour le projet. En outre, l'exportation des données au format CSV y est particulièrement aisée, ce qui est un atout supplémentaire dans l'optique de vouloir récupérer ces données en local facilement.

Dans le but de pouvoir avancer sur d'autres parties du projet pendant l'exécution des scripts de récupération de données, le scraping a été réalisé sur un serveur distant. Cela a permis, non seulement de libérer du temps et des ressources sur un poste de travail local, mais aussi d'assurer une meilleure stabilité de connexion. Il existe ici aussi un grand nombre de services d'hébergement. Afin de réduire les coûts à zéro, DigitalOcean est la solution idéale pour les étudiants. Ils proposent une formule gratuite largement suffisante pour réaliser ce type de projets. Pour faciliter encore davantage la gestion, le projet est passé par la création de conteneurs Docker. Ils permettent de faire tourner des scripts dans des environnements indépendants sur un même serveur, et de faciliter la gestion des erreurs.

### a. Web scraping

Dans un premier temps, l'objectif est d'enregistrer les informations relatives aux nageurs et nageuses belges. La course la plus pertinente pour cela est le 100m nage libre. Étant la course la plus nagée et le style de nage le plus commun, ce sont ces résultats qui expliquent le mieux l'évolution d'un ou d'une nageur.se. Ces informations sont affichées sur 2 URLs distinctes : L'une pour les hommes et l'autre pour les femmes. Chaque page ne contient que le meilleur temps de 25 sportifs, dont la première place est identifiée dans le lien de recherche. Il faut donc créer un programme itératif, qui incrémente la valeur de départ dans l'URL afin de récupérer les données page par page de tous les nageurs et nageuses belges. Une première réflexion sur la pertinence des données est nécessaire. Il y a dans le script une fonction « sleep », qui ajoute un temps d'attente, permettant d'éviter que le site bloque les requêtes réalisées depuis une même source trop rapidement. C'est une protection standard pour éviter des attaques de type DDOS, qui peut faire dysfonctionner le site. Le problème est que cela augmente considérablement la durée d'exécution et pousse à réfléchir à une limitation du nombre de personnes pertinentes pour l'analyse. Sur base des résultats disponibles sur Swimrankings, il y a plus de 9000 hommes et 9000 femmes ayant réalisé au moins un 100m nage libre en compétition. Les performances les plus lentes parmi elles n'ont pas de valeur ajoutée pour la suite du projet. Ces résultats sont soit réalisés par des jeunes, ce qui ne permet donc pas de suivre une évolution sur plusieurs années, soit par des sportifs moins réguliers, qui ne sont pas représentatifs d'une évolution standard. La décision a donc été prise de limiter les résultats sur base des points FINA (cf. 0 compréhension des données). Seuls les

nageurs ayant un résultat atteignant au moins 400 points FINA sont gardés, ce qui réduit le nombre total de nageurs à un peu plus de 6500.

Avant de passer à une nouvelle partie de script, pour récupérer l'entièreté des résultats, il faut encore choisir quels résultats sont pertinents à enregistrer. Il y a en Belgique un total de 17 disciplines différentes, réparties dans 4 (5 en comptant le 4 nages) styles (nage libre (crawl), papillon, dos et brasse) et des distances allant de 50m à 1500m. Parmi celles-ci, quelques-unes sont plus souvent courues. Principalement les 100m des 4 nages principales, 200m et 400m nage libre et le 200m

Meilleurs temps:	Sélectionner...	Classements personnels:	100m Libre	Tous les résultats:	Sélectionner...
<b>Personal rankings for 100m Libre</b>					
Grand bassin (50m)					X
53.62	896	9 fév 2025	Antwerp		
54.18	869	16 mar 2025	Edinburgh (GBR)		
54.51	853	10 fév 2023	Antwerpen		
54.52	853	15 avr 2025	Stockholm (SWE)		
54.56	851	23 avr 2023	Antwerpen		
54.72	843	10 fév 2023	Antwerpen		
54.78	841	23 avr 2023	Antwerpen		
54.82	839	24 jan 2025	Genève (SUI)		
54.99	831	16 mar 2025	Edinburgh (GBR)		
55.06	828	9 fév 2025	Antwerp		
55.19	822	3 déc 2022	Rotterdam (NED)		
55.24	820	3 déc 2022	Rotterdam (NED)		
55.34	815	9 jul 2022	Otopeni (ROU)		
55.36	814	24 jan 2025	Genève (SUI)		
55.42	812	15 avr 2025	Stockholm (SWE)		
55.48	809	8 jul 2022	Otopeni (ROU)		
55.51	808	5 jan 2025	Wezenberg		
55.68	800	11 aoû 2022	Rome (ITA)		
55.82	794	8 jul 2022	Otopeni (ROU)		
55.89	791	27 jul 2023	Fukuoka (JPN)		
55.90	791	29 mai 2024	Barcelona (ESP)		
55.95	789	11 aoû 2022	Rome (ITA)		
56.00	787	5 déc 2021	Castellon (ESP)		
56.01	786	5 déc 2021	Castellon (ESP)		
56.01	786	24 avr 2022	Antwerpen		
Petit bassin (25m)					X
52.61	871	15 déc 2024	Copenhagen (DEN)		
52.78	862	10 nov 2024	Gent		
53.12	846	10 nov 2024	Gent		
53.15	845	27 oct 2024	Aachen (GER)		
53.50	828	27 oct 2024	Aachen (GER)		
53.78	815	15 déc 2024	Copenhagen (DEN)		
54.71	774	16 oct 2022	Aachen (GER)		
54.84	769	16 oct 2022	Aachen (GER)		
54.89	767	18 déc 2022	De Sprint		
54.95	764	8 mai 2022	Hasselt		
55.01	762	20 oct 2024	Drachten (NED)		
55.53	741	10 nov 2019	Sint Amandsberg		
55.90	726	10 nov 2019	Sint Amandsberg		
57.55	665	18 nov 2018	Nijlen		
58.82	629	18 nov 2018	Nijlen		
59.13	613	11 mar 2018	Bree		
59.48	602	19 nov 2017	Nijlen		
1:00.08	585	19 nov 2017	Nijlen		
1:00.20	581	15 oct 2017	Overpeelt		
1:02.26	525	4 jun 2017	Bree		
1:04.50	472	18 déc 2016	Bree		
1:07.71	408	5 jun 2016	Bree		
1:09.09	384	6 mar 2016	Maasmechelen		
1:10.24	366	27 déc 2015	Mol		
1:10.89	356	21 fév 2016	Bree		

Figure 2 : Swimstats - Résultats Individuels 100m nage libre

4 nages. Chaque nageur a une page reprenant tous les résultats de chaque course séparément, dans un tableau de 2 colonnes. Une pour les résultats en grand bain (piscine de 50m) et une pour les résultats en petit bain (25m). Cette nuance est importante à prendre en compte aussi lors de la récupération de ces données de résultats. Il faut donc créer un script qui itère sur les 7 différentes courses sélectionnées pour chaque nageur et qui récupère les résultats de toutes les compétitions en gardant bien la nuance entre le bassin de 25m et de 50m. Une information supplémentaire qui est parfois disponible est le ou les temps intermédiaire(s) de la course. Cette information se retrouve dans un autre tableau qui ne s'affiche que lorsqu'un nageur survole un résultat. Comme cette information est facilement accessible, le script fera en sorte d'enregistrer ces informations également.

## b. Base de données

Avoir des données est une chose, s'assurer de l'intégrité, de la normalisation et de la consistance de celles-ci en est une autre. Enregistrer ces données dans une base de données relationnelle permet de solutionner ces problèmes potentiels. Pour cela, il faut en premier lieu un schéma cohérent, respectant les principes de la normalisation, ce qui réduit les risques de redondance et d'incohérences.

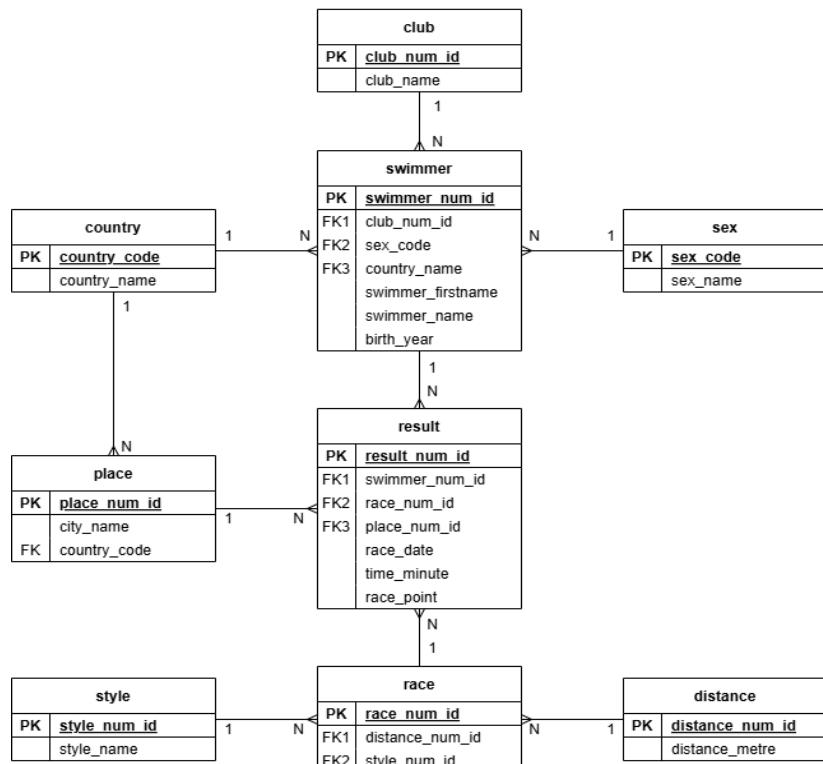


Figure 3 : Schéma relationnel

Le schéma se rapproche d'un modèle en flocon. La table centrale « result » regroupe les faits, les résultats de compétition, observés. Elle est liée à plusieurs tables de dimension, dont la principale est la table « swimmer ». Cette dernière référence également d'autres tables de dimension qui sont principalement des tables de code, qui répondent aux problèmes de redondance. La table « race » est une table dite d'intersection. Elle est nécessaire pour normaliser et éviter les doublons dans le cas d'une relation « Many to Many ». Cela permet aussi une meilleure flexibilité dans le cas où de nouvelles épreuves devaient être ajoutées.

Ce schéma est ensuite implémenté dans PostgreSQL. Pour cela, le script de création de tables et de clés primaires et étrangères est développé de manière à respecter les contraintes du diagramme, ainsi que les contraintes liées aux données. Chaque colonne est liée à un domaine qui limite les valeurs dans un certain type de données afin de s'assurer qu'il n'y ait pas de valeurs incorrectes enregistrées.

## c. Serveur

La dernière étape de la mise en place de la récupération automatisée des données consiste à encapsuler les scripts dans des images Docker. Un fichier Dockerfile permet de définir les ressources nécessaires, les bibliothèques à installer, les variables d'environnement et le comportement du conteneur lors de l'exécution. Pour faciliter la communication entre les conteneurs avec le script python et celui de la base de données PostgreSQL, un réseau Docker est d'abord créé via la commande : `docker network create swimstats_nw`

Les deux images déployées sur le serveur sont d'une part pour la base de données (swimstats:db) et d'autre part pour le script de scraping python (swimstats:py). Une fois créées à l'aide de la commande « `docker build -t <nom_image>` » et ensuite mises à disposition sur Docker même avec « `docker push <nom_image>` ». Elles sont dès lors récupérables depuis le serveur à l'aide de la commande « `docker pull <nom_image>` », et exécutables en utilisant cette commande : « `docker run -network swimstats_nw -name <nom_conteneur> -d -t <nom_image>` ». Les paramètres de cette dernière commande permettent respectivement d'identifier le réseau dans lequel il doit être exécuté, le nom associé au conteneur, le lancement en mode 'daemon' (pour que le conteneur s'exécute en tâche de fond), et dernièrement le nom de l'image à exécuter. « `docker logs <nom_conteneur>` » est la commande à lancer pour vérifier le statut du script et voir si des logs peuvent expliquer une potentielle erreur. Dernièrement, 2 commandes similaires ont été utilisées pour extraire les données brutes du conteneur PostgreSQL vers des fichiers CSV :

- `docker exec -it -u <user> <nom_conteneur> psql -d <nom_database> -c "COPY (SELECT * FROM races) TO STDOUT WITH CSV HEADER" > races.csv`
- `docker exec -it -u <user> <nom_conteneur> psql -d <nom_database> -c "COPY (SELECT * FROM swimmers) TO STDOUT WITH CSV HEADER" > swimmers.csv`

Une subtilité importante dans cette configuration est l'utilisation de 2 serveurs distincts. Chaque serveur est dédié à la récupération des résultats pour un sexe. Cette approche permet de paralléliser l'exécution des scripts, ce qui réduit considérablement le temps nécessaire pour récupérer toutes les données. De plus, cette séparation sur 2 serveurs avec des adresses IP différentes permet de répartir la charge des requêtes et de limiter les risques de blocage de la part du site.

root@ubuntu-s-1vcpu-2gb-ams3-01:~# docker ps -a						
CONTAINER ID	IMAGE	COMMAND	CREATED	STATUS	PORTS	NAMES
485085a50d0c	swithan/swimstats:py	"python s.py"	3 seconds ago	Up 1 second		py
ef6f71971230	swithan/swimstats:db	"docker-entrypoint.s..."	5 months ago	Up 5 months	5432/tcp	db

Figure 4 : Liste des conteneurs actifs sur un serveur

Pour récupérer les données exportées dans les fichiers CSV, une dernière commande doit être exécutée sur l'ordinateur local, en utilisant `scp`, un protocole de transfert de fichiers.

```
scp <user>@<server_ip>:<path_to_csv> <path_to_output>
```

## VII. Compréhension des données

Explorer et analyser les données permet de dégager des observations générales sur les données. Tout d'abord, il est important de comprendre chaque colonne des fichiers de données récupérés lors de la phase précédente. La plupart des informations sont assez explicites, comme le nom, le prénom, la date de naissance, le sexe, la nationalité et le dernier club d'un nageur, associé à son ID spécifique. Pour les données relatives aux résultats, il y a davantage de subtilités. Premièrement, il y a la colonne avec l'identifiant du nageur, permettant de relier les 2 tables entre elles. Ensuite, toutes les informations relatives au résultat : la distance et le style, la date, le lieu (ville et pays), la longueur de la piscine, le temps réalisé (en secondes) et le nombre de points FINA auquel le résultat correspond. Cette dernière est un nombre de points (P) calculé sur la base du temps réalisé (T) et d'un temps de référence (B) à l'aide de la formule suivante :  $P = 1000 * (B / T)^3$ . Les temps de référence sont redéfinis chaque année sur la base des derniers records du monde ratifiés. Il y a des temps de référence pour chaque course, différents pour les hommes et les femmes, et également séparés pour les courses en grand bassin (piscine de 50m) et en petit bassin (piscine de 25m). La phase de récupération des données a permis de récupérer un total de 6651 nageurs qui, ensemble, ont réalisé 1480181 résultats. Comment ces résultats sont-ils distribués selon l'âge ou le sexe du nageur ? Quel est le plus grand club du pays en termes de nombre de nageurs ? Est-ce que ce club est aussi celui avec le plus grand nombre de courses par nageur ? Au niveau des courses, quelles sont les disciplines et distances les plus populaires ? Comment évoluent les résultats en fonction de l'âge du nageur ?

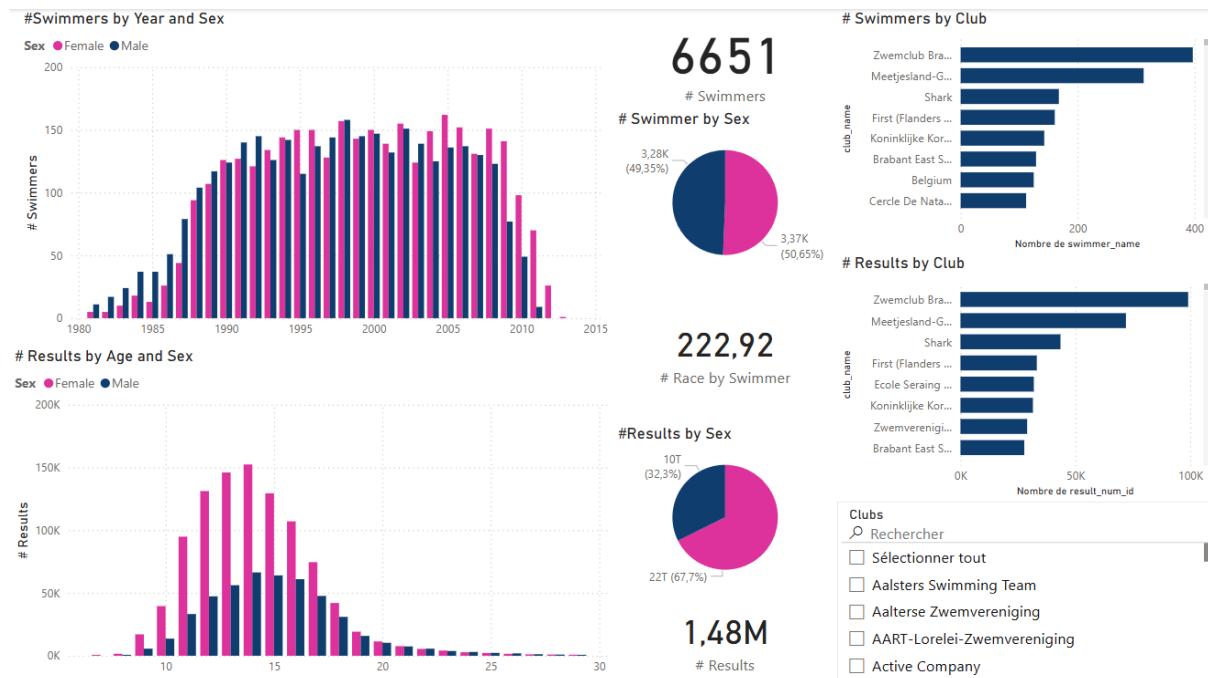


Figure 5 : Dashboard nageurs & clubs

## a. Visualisation

Un premier dashboard réalisé sur Power BI permet de distinguer la distribution du nombre de nageurs par année de naissance et par sexe, le nombre de nageurs par club et le nombre de courses réalisées par âge et sexe. Des KPI reprennent le nombre de résultats et de nageurs, ainsi qu'une mesure calculée permettant d'identifier le nombre de courses par nageur. On peut premièrement constater que, malgré un nombre de nageurs assez similaire dans les 2 sexes, le nombre de courses réalisées par les femmes est 2 fois plus important que pour les hommes. Cette information est visible dans les pie charts, mais aussi déductible dans les bar charts. La distribution de ces mêmes KPI par club suit une distribution globalement similaire pour le nombre de nageurs et de résultats, expliquée par le troisième KPI indiquant le nombre de courses par nageur.

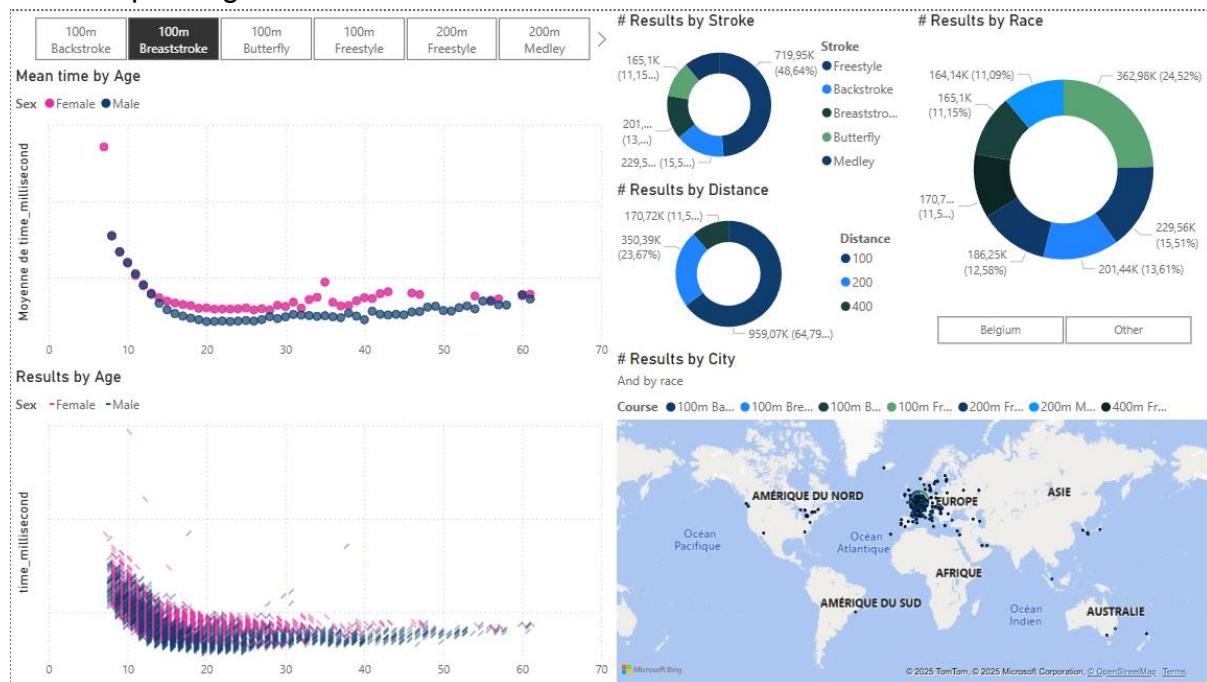


Figure 6 : Dashboard résultats

Un second dashboard présente d'une part la distribution et l'évolution des résultats avec l'âge pour chaque course séparément. Ces graphes présentent tous une évolution rapide à un jeune âge, avant de se stabiliser et de finir par augmenter légèrement. Des donut-charts permettent de voir la distribution des courses par style et par distance afin d'indiquer la ou les courses les plus communes. Dans ce dashboard, une carte présente finalement les lieux où les compétitions ont eu lieu, avec un filtre permettant de se concentrer sur les compétitions ayant eu lieu en Belgique. On aperçoit ici quelques manquements liés à la qualité des données. Certains lieux comme Anvers indiquent plusieurs points au même endroit, sans assembler les valeurs. Cela est dû à un manque de normalisation des valeurs. Les compétitions à Louvain ne sont pas associées avec celles réalisées à Leuven.

## b. Statistiques descriptives

Afin d'améliorer encore la compréhension des données, une analyse statistique descriptive est réalisée. Les statistiques calculées varient en fonction du type de données. Pour les données quantitatives, un graphe de distribution des temps est réalisé. Cela permet d'identifier des valeurs aberrantes, de voir la moyenne, les minimums et maximums ainsi que les écarts interquartiles. Une information surprenante visible ici est que, bien que le meilleur temps de chaque course soit toujours réalisé en piscine de 25m, la distribution moyenne des courses réalisées dans des longueurs de 50m est plus rapide et l'écart interquartile est plus petit.

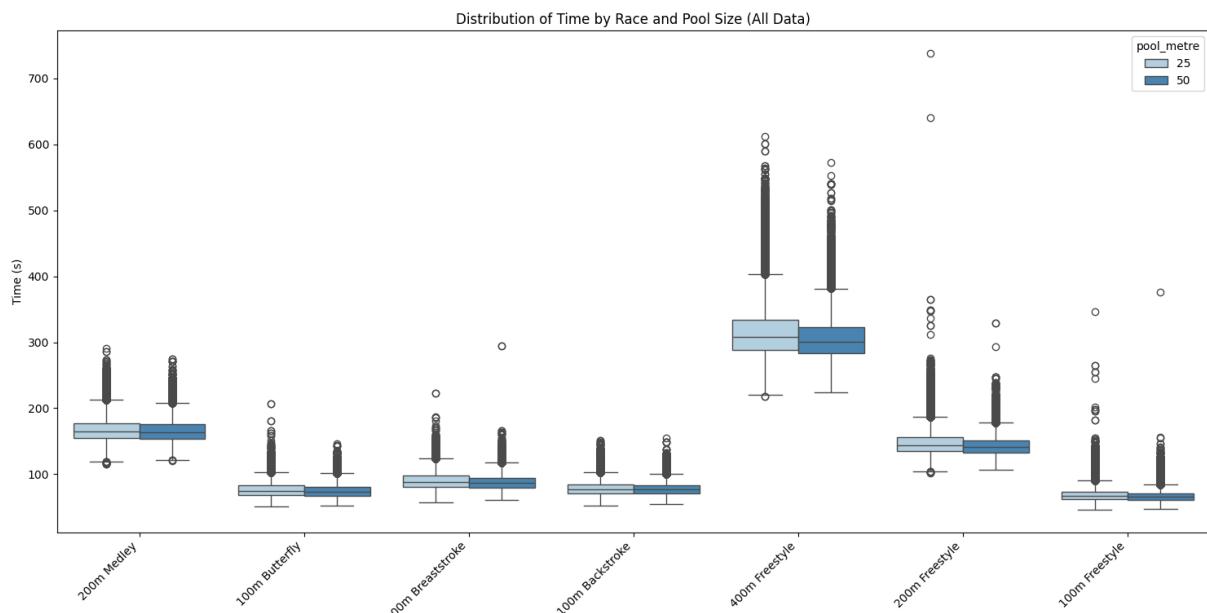


Figure 7 : Distribution des temps par course

Pour les données temporelles comme la date de la course ou l'année de naissance, les statistiques suivantes sont identifiées :

Colonne	# valeurs	Min	Max	Mode
<b>result_date</b>	3801	1988-09-23	2025-03-30	2005-07-28 (3678)
<b>birth_year</b>	51	1956	2013	1998

On peut dès lors affirmer que les données traitent de résultats réalisés entre 1988 et 2025, pour des nageurs nés entre 1956 et 2013. Les données qualitatives comme le style de nage ou le nom du club ne nécessitent pas de statistiques approfondies. Elles sont déjà expliquées à l'aide des dashboards Power BI présentés plus tôt.

## VIII. Préparation des données

La qualité des données est une condition indispensable pour pouvoir créer un modèle prédictif performant. Avant cela, il est nécessaire de passer par une phase de préparation des données. Lors de cette étape, la première tâche importante est de nettoyer les données non normalisées, de gérer les données manquantes et de supprimer les données redondantes. Cela est réalisé à l'aide d'un outil d'ETL. SSIS, l'outil d'ETL développé par Microsoft, en plus d'être étudié dans le cadre du bachelier de spécialisation, propose une interface visuelle très intuitive. Chaque étape est clairement identifiable, ce qui facilite la gestion d'erreurs et l'adaptabilité. Sur base de ces données propres et exploitables, la seconde partie implique la création de nouvelles variables. Ces nouvelles informations permettent d'augmenter la précision des prédictions d'un modèle. Pour effectuer cela, Python et sa librairie Pandas, sont la solution privilégiée. La principale raison est que la modélisation passe également par l'utilisation de ce langage de programmation et que revenir à la phase de « feature engineering » (création de variables) est très souvent nécessaire. Cela facilite donc l'itération de ces 2 étapes.

### a. ETL

Une approche structurée et réfléchie est nécessaire pour mettre en place un bon flux ETL. Sur base des données brutes enregistrées dans divers fichiers CSV et du schéma de base de données, chaque étape de transformation peut être mise en place. En premier lieu, toutes les données en CSV sont transférées vers des tables de base de données temporaires. La seule modification faite lors de cette étape est la séparation des informations du lieu de la compétition. Les informations de localisation varient en fonction du pays dans lequel la course a eu lieu. Dans le cas où la compétition s'est déroulée en Belgique, seul le nom de la ville est partagé. Pour les courses parcourues en dehors de la Belgique, il y a généralement un code de pays noté dans cette même colonne. Une séparation est donc faite entre ces deux informations, pour créer une colonne ville et une colonne pays. Pour compléter les données, une valeur n'ayant pas de pays spécifié originellement se voit attribuer le pays belge automatiquement.

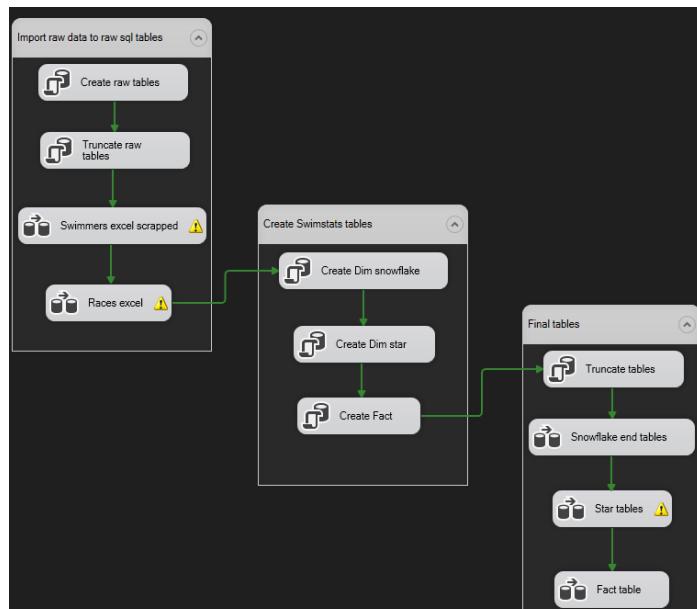


Figure 8 : Flux de données global

Le second groupe rassemble toutes les tâches d'exécution des scripts de création des tables modélisées et normalisées. Comme illustré dans le chapitre Base de données, un modèle en flocon a été préféré dans le cadre du projet. Ceci implique 3 niveaux de tables. La table centrale reprend les faits, dans ce cas les résultats. Directement connectées à la table de fait, sont les tables de dimension principales. Celles-ci réfèrent généralement encore à des tables de dimension secondaires, ou périphériques. C'est par ces dernières que l'exécution des scripts de création de tables démarre car elles ne comportent aucune clé étrangère, donc aucune dépendance à d'autres tables. Les tables de dimension principales viennent ensuite en référençant les tables déjà créées. Pour finir, le script de création de la table des résultats (table de faits) est exécuté, avec la création des clés étrangères associées.

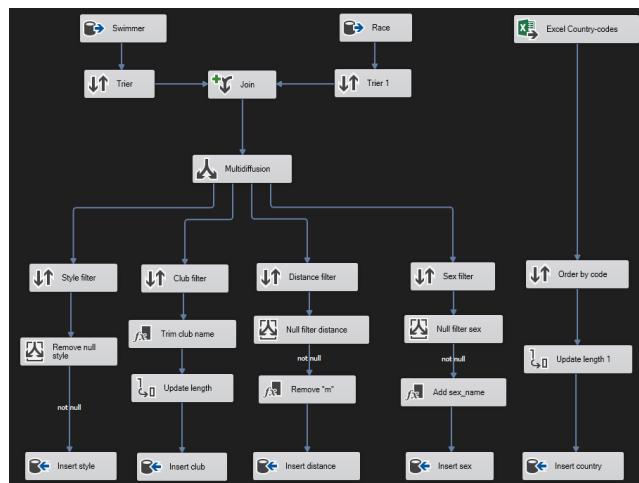


Figure 9 : Flux des tables de dimensions périphériques

Insérer les données dans les nouvelles tables créées demande davantage de travail. En reprenant l'ordre d'exécution des scripts de création de tables, chaque table peut être remplie sur base des données brutes enregistrées lors de la première étape du flux ETL. Afin de respecter les contraintes des tables normalisées qui permettent d'assurer une qualité des données, plusieurs transformations sont nécessaires. Les tables de dimension périphériques sont toutes des tables de code. Elles reprennent toutes les valeurs uniques possibles pour les styles et les distances de course, les noms des clubs et le sexe des nageurs extraits depuis les tables de données brutes. La table « country » est un peu particulière. Celle-ci ne récupère pas les données extraites du site web, mais depuis un fichier Excel supplémentaire. Ce fichier comporte une liste des pays avec leur code correspondant (ex. BEL pour Belgique). Le flux illustré à la Figure 9 : Flux des tables de dimensions périphériques montre le processus d'extraction, de filtrage et de transformation appliqué à chaque concept. Les données brutes sont récupérées dans la base de données et assemblées. Pour chaque dimension, un filtre récupère uniquement les valeurs distinctes pour chacune des colonnes pertinentes. Des transformations spécifiques sont appliquées afin de gérer correctement le type de données (Ex. : les distances sont enregistrées avec les unités dans la structure de données non transformée, information qui doit être retirée). Chaque flux se termine par une tâche de destination pour remplir la table correspondante.

Pour les tables de dimension principales et la table de fait, le flux de données SSIS nécessite une approche particulière. Les lignes de données sont récupérées à l'aide de requêtes SQL spécifiques. Ces requêtes permettent de sélectionner uniquement les colonnes utiles, de joindre les tables sur base des clés étrangères facilement et d'appliquer des filtres précis. C'est le cas du code SQL ci-dessous, qui permet de récupérer toutes les courses distinctes présentes dans les données brutes et relie les informations de distance et de style de course aux tables de dimension secondaires. La commande SQL précise que les seules colonnes finalement récupérées sont les identifiants de style et de distance, qui sont les informations à récupérer et insérer dans la table « race ».

```

SELECT DISTINCT
    d.distance_num_id,
    s.style_num_id
FROM public.race AS r
JOIN swimstats.distance AS d
    ON CAST(REPLACE(r.distance, 'm', '') AS INTEGER) =
d.distance_metre
JOIN swimstats.style AS s
    ON r.stroke = s.style_name;

```

## b. Feature engineering

Dans le but d'optimiser les performances du modèle, créer des « features » ou colonnes supplémentaires est indispensable. Ces nouvelles informations permettent de mieux représenter les tendances et les comportements identifiables dans les données d'origine. Un modèle de réseau de neurones fonctionne beaucoup mieux lorsque beaucoup de variables non corrélées sont ajoutées. Des premières variables calculées directement sont facilement identifiées. L'âge auquel le résultat est réalisé est dépendant de l'année de naissance et de la date de la course. Avec une réflexion plus approfondie, en natation, l'âge est basé sur une saison sportive allant de septembre à aout, ce qui implique une modification du calcul de l'âge : l'année de la course est calculée sur la saison. Une course en mars 2025 est donc réalisée dans la saison 2024, tout comme une course en octobre de cette dernière année. Pour d'autres colonnes, une normalisation des informations est nécessaire. Par exemple, le sexe (M ou F) ou la longueur de la piscine (25m ou 50m) sont transformés en unités binaires (0 ou 1). Les données catégoriques sont adaptées pour qu'un modèle puisse en extraire de l'information pertinente. En cherchant à optimiser encore les performances du modèle, les résultats précédents, sur la même course, sont également ajoutés dans une nouvelle colonne, tout comme le temps écoulé depuis la dernière course effectuée. L'information de l'âge lors de la première course peut également être un facteur pertinent pour l'entraînement du modèle. La dernière

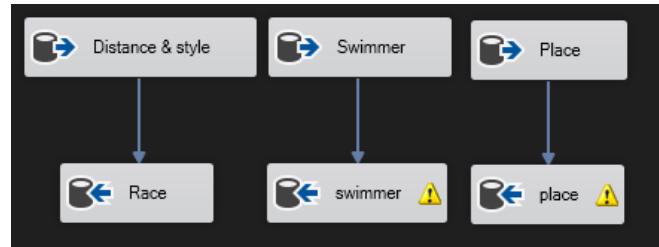


Figure 10 : Flux des tables de dimensions principales

colonne ajoutée est une colonne avec un calcul complexe. À l'aide d'une régression linéaire, la tendance d'évolution sur les 3 dernières courses est calculée. C'est une information cruciale qui permet d'expliquer l'évolution du temps d'un nageur sur un très court terme. Si cette valeur est négative, cela indique une progression. À l'inverse, un coefficient de pente positif indique un manque de forme récent.

```
def compute_trend(group, window=3):
    trends = []
    for i in range(len(group)):
        if i < window - 1:
            trends.append(np.nan)
        else:
            y = group['time_millisecond'].iloc[i - window
+ 1:i + 1].values.reshape(-1, 1)
            x = np.arange(window).reshape(-1, 1)
            model = LinearRegression().fit(x, y)
            slope = model.coef_[0][0]
            trends.append(slope)
    return pd.Series(trends, index=group.index)
```

## IX. Modélisation des données

Avant de passer à la mise en place de techniques statistiques sur les données, il est important de s'assurer de comprendre le cadre dans lequel il s'inscrit. L'objectif du projet est de prédire les résultats de compétition, indiquant qu'une technique de prédiction d'estimation doit être appliquée. Cette technique s'inscrit dans le cadre d'un apprentissage supervisé, avec des prédicteurs (colonnes X) et une valeur correspondante Y, le résultat de la course. De plus, il y a une relation claire entre les variables explicatives et les résultats. L'âge en est le principal exemple. Dans ce contexte, les modèles de régression linéaire ou polynomiale et les réseaux de neurones sont les techniques les plus adaptées.

Pour confirmer ces hypothèses, à l'aide des visualisations réalisées en amont, il est simple d'identifier une tendance d'évolution des résultats de natation en fonction de l'âge. Avec une évolution rapide au début, et une tendance atteignant une limite minimale avant d'augmenter légèrement à nouveau. Ceci indique qu'une régression polynomiale pourrait expliquer la relation et généraliser l'évolution du temps d'un sportif. Pour s'assurer de trouver le meilleur modèle, configurer un réseau de neurones et l'optimiser est la seconde technique appliquée aux données.

Comme expliqué lors de la phase de préparation des données et du « Feature engineering », la création d'un modèle performant passe par beaucoup d'adaptations et de changements au niveau de la création de variables. Il y a ensuite de nombreuses itérations nécessaires pour trouver la configuration du modèle la plus optimale.

## a. Régression polynomiale

L'implémentation d'un modèle de régression polynomiale se fait en plusieurs étapes. Sur base des données importées dans le script python et des nouvelles colonnes créées lors de la phase de feature engineering, une sélection de variable est réalisée. Les colonnes avec les variables prédictives sont assemblées, et la colonne cible est gardée séparément. Ensuite, les données sont séparées en groupe de données d'entraînement, qui sera utilisé pour entraîner le modèle, et le groupe de test, qui sert de vérification des résultats de la phase d'entraînement. L'entraînement d'un modèle de régression est un processus itératif. Une valeur de pente ( $w$ ) et d'ordonnée à l'origine ( $b$ ) est attribuée aléatoirement. À l'aide de l'algorithme du gradient descent, des itérations sont réalisées sur les données pour faire varier les valeurs  $w$  et  $b$ , jusqu'à trouver un optimum local. Cet optimum local est l'endroit où l'erreur d'estimation est la moins importante en moyenne. Sur base des valeurs optimales trouvées lors de l'entraînement avec les données d'entraînement, une vérification est réalisée sur les données de test. Lors de cette étape, il est possible d'identifier des problèmes dans le modèle. Si les résultats de ce modèle ne sont pas suffisants pour les données d'entraînement et les données de test, le modèle pourrait se trouver en situation d'underfitting. Cela signifie que le modèle ne représente pas suffisamment les données et ne peut donc pas être utilisé pour généraliser l'estimation. À l'inverse, si le modèle est très optimisé vis-à-vis des données d'entraînement mais que l'erreur est importante sur les données de test, le risque d'overfitting est grand. Le modèle n'est alors pas généralisable non plus, car il est uniquement applicable aux données sur lesquelles il a été construit.

Pour paramétrier un entraînement de régression linéaire, plusieurs valeurs sont modifiables. Premièrement, les colonnes de variables explicatives. C'est sur ces données que le modèle va se baser pour faire une prédiction de temps. En plus des colonnes de la base de données, il y a l'ajout des colonnes lors de la Préparation des données. Les autres paramètres sont :

1. Les colonnes explicatives pertinentes
2. Les valeurs initiales de  $w$  et  $b$  (attribution aléatoire)
3. Le coefficient de régularisation (lambda : éviter le surapprentissage)
4. Le nombre d'itérations (iterations : nombre de recalculs de  $w$  et  $b$ )
5. Le taux d'apprentissage (alpha : taille des sauts pour le gradient descent)

La phase d'apprentissage du modèle de régression polynomiale est très fastidieuse. Le temps requis pour atteindre le nombre d'itérations (ou un arrêt anticipé si le modèle ne s'améliore pas suffisamment entre 2 itérations) est important. Pour quantifier les performances du modèle, un calcul de l'erreur quadratique moyenne (RMSE) permet d'identifier que l'erreur absolue de prédiction d'un temps est de l'ordre de 3,3 secondes. Cela représente une erreur de 5,11% du temps d'une course.

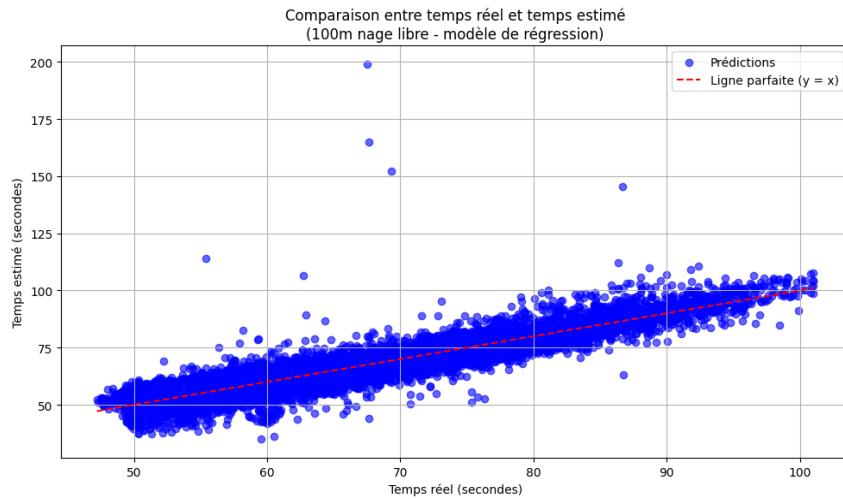


Figure 11 : Résultats modèle de Régression Polynomiale (100m nage libre)

Dans l'idée de comparer ce modèle à un réseau de neurones se basant sur les mêmes informations, les données utilisées pour l'entraînement du modèle se limitent aux résultats obtenus sur le 100m nage libre. Cette course est la plus représentative, tant en termes du nombre total de résultats récupérés qu'en termes de représentation de l'évolution des performances globales d'un nageur. Sur base du meilleur modèle pour cette course, une analyse et un entraînement plus approfondi sont appliqués sur les autres distances et styles de nage.

## b. Réseau de neurones

Un modèle de réseau de neurones se compose de plusieurs couches de neurones, reliés à l'aide de fonctions d'activation. La couche d'entrée représente les variables explicatives du modèle. Il y a ensuite une ou plusieurs couches cachées, dont les résultats ne sont pas interprétables par l'humain. Ces couches consécutives utilisant des fonctions d'activation finissent par atteindre la couche de sortie, généralement à l'aide d'une fonction d'activation linéaire dans le cas d'un modèle de prédiction d'estimation. Cette dernière couche est composée d'un seul neurone, qui donne le résultat de la prédiction. L'entraînement du réseau de neurones s'effectue en « epochs », qui représente un cycle de passage des données d'entraînement à travers le réseau de neurones. Lors du cycle, les données suivent en premier lieu une propagation avant : toutes les données passent par les couches du modèle pour donner un résultat. Ensuite, sur base de ce premier passage, une fonction de coût calcule l'erreur, autrement dit l'écart entre le résultat obtenu et le résultat attendu. Vient ensuite un processus itératif de propagation arrière permettant de corriger les poids associés à chaque couche du réseau.

Il y a donc plusieurs paramètres à prendre en compte lors de la mise en place du modèle :

1. Les colonnes explicatives pertinentes
2. Le nombre d'Epochs
3. Le nombre de couches de neurones du modèle
4. Le nombre de neurones par couche du réseau
5. Les fonctions d'activation pour chacune des couches
6. Le batch size (nombre de données par échantillon)

```
model_2 = Sequential([
    Dense(16, activation='relu',
          input_shape=(x_train.shape[1],)),
    Dense(8, activation='relu'),
    Dense(4, activation='relu'),
    Dense(1, activation='linear')
], name='model')
model.fit(
    x_train, y_train,
    epochs=500,
    batch_size=64,
    validation_split=0.1,
    verbose=0
)
```

Ce type de modèle est beaucoup plus paramétrable que le modèle de régression polynomiale créé plus tôt. La structure globale du réseau de neurones (nombre de couches et de nœuds par couche, type de fonction d'activation) est le paramètre le plus manipulé dans l'optique d'optimiser le résultat.

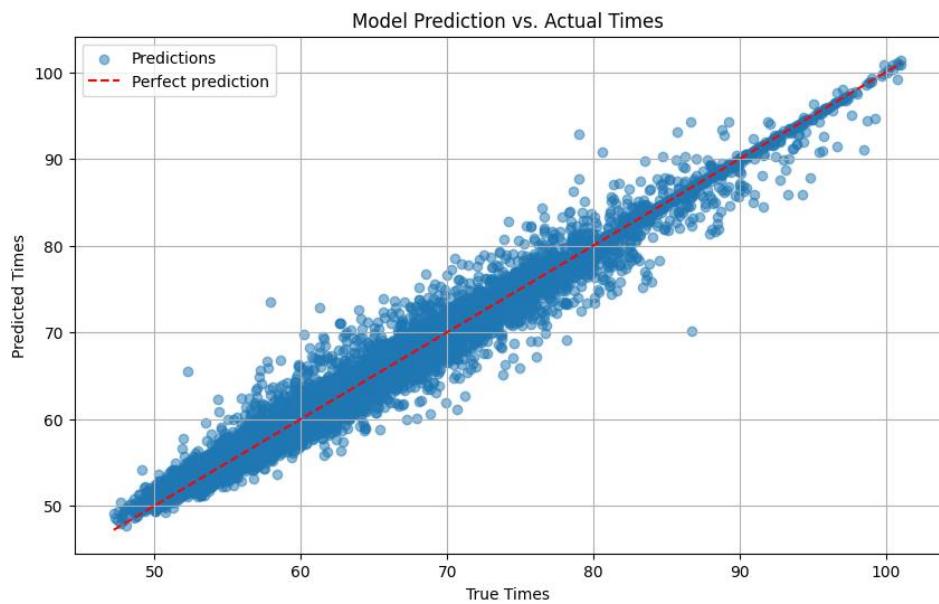


Figure 12 : Résultats modèle de Réseau de Neurones (100m nage libre)

Comme pour le modèle de régression polynomiale, la première course analysée est le 100m nage libre. Les résultats de l'entraînement de ce modèle sont extrêmement performants. Pour juger la qualité du modèle, la RMSE (Root Mean Square Error) est calculée sur les données d'entraînement, ce qui donne la qualité de l'entraînement à proprement dit, et cette même erreur est calculée sur les données de test. La comparaison des deux valeurs permet de s'assurer que le modèle n'est pas en sous- ou sur-apprentissage. Pour le résultat de l'entraînement, le modèle prédit des résultats en moyenne 1,22 seconde à côté du résultat attendu. Pour les données de test, cette erreur est même inférieure, avec 1,20 seconde. Ceci montre que le modèle n'est pas en surapprentissage. Cette petite différence est simplement due au hasard de la distribution des données d'entraînement et de test. Elle correspond à un écart de 1,88% du temps moyen.

### c. Modèle final

Sur base des deux modèles appliqués et configurés, le réseau de neurones sort largement du lot en termes de performances. Le temps nécessaire à son entraînement dans la configuration actuelle est de moins de 9 minutes contre près de 27 minutes pour le modèle de régression. Ce temps est précieux, surtout dans le cas où d'autres courses doivent également être entraînées et modélisées à l'aide de la même configuration. Pour ce qui est des performances intrinsèques des modèles de prédiction, le réseau de neurones est capable de prédire des résultats 2,5 fois plus précis en moyenne. En comparant les graphiques de prédictions, la distance entre la valeur prédite et la valeur réelle du second modèle n'est jamais aussi élevée que celle du modèle de régression. Ce dernier prédit dans certains cas des temps 2, voire 3 fois supérieurs aux temps réels.

En gardant donc le modèle de réseau de neurones, l'objectif est d'appliquer cette même configuration sur les résultats des autres courses. À ce niveau, les résultats indiquent des performances de précisions assez similaires. L'écart de prédictions en pourcentage entre les prédictions obtenues et les résultats attendus oscille entre 1,74% et 2,02% du temps moyen pour chaque combinaison de distance et de style. Pour rappel, les résultats du modèle pour les 100m crawl se situent en moyenne à 1,88% du temps total. Ceci indique donc que les paramètres optimaux du réseau de neurones initialement développé sont applicables aux autres courses.

Pour approfondir l'analyse des différentes courses, certaines architectures de réseau de neurones considérées sous-optimales pour le 100m nage libre sont réévaluées ici. Voici le détail des analyses :

	Couches	Nœuds	Fonctions d'activation
<b>Model 1</b>	4	16 – 8 – 4 – 1	ELU
<b>Model 2 (opti. 100m nl)</b>	4	16 – 8 – 4 – 1	RELU
<b>Model 3</b>	4	4 – 8 – 4 – 1	RELU
<b>Model 4</b>	3	8 – 4 – 1	ELU

Courses	Modèles		model 1		model 2		model 3		model 4		Total rmse moyen		Total rmse % moyen	
			rmse moyen	rmse % moyen	rmse moyen	rmse % moyen	rmse moyen	rmse % moyen	rmse moyen	rmse % moyen	rmse moyen	rmse % moyen	rmse moyen	rmse % moyen
100m br			1,47	1,80%	1,48	1,81%	1,68	2,05%	1,49	1,82%	1,53	1,87%		
100m dos			1,31	1,79%	1,32	1,81%	1,44	1,97%	1,36	1,86%	1,3575	1,86%		
100m nl			1,22	1,89%	1,22	1,88%	1,34	2,07%	1,22	1,89%	1,25	1,93%		
100m pap			1,36	1,93%	1,42	2,02%	1,61	2,29%	1,42	2,03%	1,4525	2,07%		
200m 4n			2,57	1,64%	2,71	1,74%	2,96	1,90%	2,88	1,84%	2,78	1,78%		
200m nl			2,48	1,82%	2,47	1,82%	2,65	1,95%	2,62	1,93%	2,555	1,88%		
400 nl			5,46	1,89%	5,73	1,98%	6,7	2,32%	5,81	2,01%	5,925	2,05%		
<b>Total général</b>	<b>2,267142857</b>	<b>1,82%</b>	<b>2,335714286</b>	<b>1,87%</b>	<b>2,625714286</b>	<b>2,08%</b>	<b>2,4</b>	<b>1,91%</b>	<b>2,407142857</b>	<b>1,92%</b>				

Figure 13 : Performances de chaque Modèle par Course

On peut identifier 2 configurations plus précises que les autres. La seule différence entre les 2 modèles réside en l'utilisation de deux fonctions d'activation distinctes. La fonction 'ELU', bien que moins utilisée, permet une approche plus évolutive et douce lors de l'approche du seuil de la fonction, tandis que 'ReLU' réalise un changement net au niveau du seuil. La première approche semble très légèrement se démarquer en termes de performances (RMSE % moyen 1,82% contre 1,87%). Ces valeurs étant assez insignifiantes, il est intéressant de comparer la vitesse d'exécution de chacun de ces modèles. Cette information est également très importante à prendre en compte dans le cas où de nombreux nouveaux calculs sont nécessaires.

Moyenne de temps	Modèles	Course	model 1	model 2	Total général
		100m br	261,93	261,11	261,52
		100m dos	323,99	326,76	325,375
		100m nl	549,54	545,28	547,41
		100m pap	246,68	246,02	246,35
		200m 4n	268,38	262,4	265,39
		200m nl	257,48	256,46	256,97
		400 nl	216,87	240,04	228,455
		<b>Total général</b>	<b>303,55</b>	<b>305,44</b>	<b>304,50</b>

Figure 14 : Vitesse d'exécution de chaque Modèle par Course (en secondes)

À nouveau, les performances en termes de temps nécessaire à l'entraînement des modèles sont légèrement en faveur du premier modèle, pour une différence inférieure à 2 secondes. En conclusion, il est préférable de garder le modèle numéro 1, utilisant une fonction d'activation 'ELU'.

## d. Prédictions

Pour réaliser les prédictions, une dernière partie de script est développée. Cette partie implique la création des informations d'une compétition que l'on souhaite prédire. Ceci implique le nom du nageur, la course réalisée et la date de cette compétition ainsi que la taille de la piscine. Sur base de ces informations, les colonnes calculées utilisées comme « features » du modèle doivent être calculées. Il faut ensuite récupérer le modèle associé à la course et utiliser la méthode `model.predict(input)`. Cette fonction donne le résultat final de la prédiction en secondes, qui est donc une estimation avec une marge d'erreur propre au modèle.

Cependant, lors de tentatives de prédictions sur différents nageurs et nageuses, une erreur surprenante s'est présentée. Régulièrement, le nageur ou la nageuse ne comportait aucun résultat, rendant la prédiction impossible. À la suite de cette constatation et d'une analyse du code et de l'évolution des données au fil de celui-ci, il est possible d'identifier que le filtre de données est trop sélectif. La distribution du nombre de nageurs utilisés lors de l'entraînement des données est extrêmement limitée. Seuls 550 à 1300 d'entre eux étaient repris dans chaque groupe de données par course. Il faut donc un retour en arrière et une modification de ce code de sélection des variables, pour augmenter le nombre de données utilisées lors de l'entraînement.

```
# Filter swimmers with > 5 races and presence at age 16
group_df = group_df.groupby('swimmer_name').filter(
    lambda x: len(x) > 5 and 16 in x['age'].values # Anciennes valeurs : min 10 courses et au moins une course à 14 et 18 ans
)
```

Figure 15 : Code de tri modifié

Grâce à ces changements, le nombre de nageurs pris en compte pour l'entraînement de chaque course a plus que triplé : entre 2500 et 4200 nageurs par course. Sur ces nouvelles données, l'application du modèle de régression polynomiale et du réseau de neurones donne des résultats différents et plus performants. Les estimations du premier modèle pour le 100m nage libre atteignent une erreur de 2,8 secondes en moyenne, contre 3,3 avant. Une prédiction du réseau de neurones ne comporte une erreur plus faible également avec 1,15 seconde (1,22 seconde avant modifications). Un nombre de données plus important implique en revanche un temps d'entraînement plus long. Le modèle de régression nécessite près de 2 heures pour atteindre une forme optimale, soit plus de 4 fois plus long. Il faut un peu moins de 23 minutes pour l'entraînement du réseau de neurones, contre 9 minutes avant la modification. Ceci justifie encore davantage l'utilisation de cette dernière approche pour créer des modèles adaptés à toutes les autres courses.

```
Using model for 100_Freestyle: 100_Freestyle
1/1 ━━━━━━━━━━━━━━━━ 0s 34ms/step
 Predicted time for Gjon VATA in 100m Freestyle on 2025-05-27:
53.96 seconds
```

## X. Visualisation

Une partie des objectifs du projet est de proposer des dashboards à l'aide d'un outil BI. L'un doit permettre aux clubs de pouvoir voir les tendances d'évolution de tous les nageurs. Le second offre aux nageurs une possibilité de se comparer avec les résultats d'autres nageurs. Pour réaliser ces dashboards, l'application Power BI a été utilisée pour la gratuité d'une partie de son utilisation et la personnalisation qu'elle offre. Sa connectivité avec des bases de données afin d'avoir accès à des données dynamiques est un atout supplémentaire.

### a. Clubs

L'outil proposé aux clubs permet d'avoir une vue d'ensemble sur les résultats des nageurs et nageuses du club. Il montre principalement la distribution des résultats en fonction de divers facteurs comme l'âge, la course ou l'année de naissance. Ces distributions permettent de voir l'évolution de la taille du club et de gérer le flux de nageurs. D'autres graphes permettent d'identifier l'évolution des résultats moyens et individuels par âge. L'interactivité de ce dashboard est importante. Beaucoup de listes et de 'segments' sont présentés pour permettre de comparer des données plus précises, comme identifier la courbe d'évolution d'un nageur spécifique ou identifier la course la plus nagée au cours d'une année précise. De plus, l'interactivité entre les graphes est gérée de telle manière à pouvoir garder un aperçu global de certaines informations tout en étant capable de se focaliser sur un élément spécifique. Beaucoup de couleurs et d'informations sont présentes, mais les informations clés restent facilement identifiables et compréhensibles.



Figure 16 : Dashboard dédié aux Clubs de natation

Grâce à ce dashboard, un entraîneur peut déjà identifier le potentiel de certains nageurs. La direction d'un club peut également utiliser cet outil pour avoir un aperçu de l'évolution et de l'état actuel du club. L'exemple ci-dessus démontre une croissance forte du nombre de courses depuis 2016. La diminution en 2020 et 2021 est liée au Covid et n'est donc pas représentative du statut du club.

## b. Nageurs

Le second rapport visuel Power BI destiné aux nageurs, mais également pertinent pour les entraîneurs, offre la possibilité de voir l'évolution d'un nageur et de comparer celle-ci avec un ou plusieurs autres nageurs. Il permet de présenter et comparer les meilleurs temps et le nombre de participation à chaque course. De plus, une carte est ajoutée pour identifier les endroits où les résultats ont été réalisés. Pour finir, les courbes d'évolution du record personnel à chaque âge sont présentées. Ce visuel comporte également des segments de tri afin de pouvoir sélectionner le ou les nageurs à comparer, ainsi que la course dont la courbe d'évolution est présentée.



Figure 17 : Dashboard dédié aux Nageurs

L'exemple affiché ci-dessus est une comparaison entre 2 nageurs, a priori frères. Le premier graphe montre que 'Ruben' a des résultats plus rapides pour toutes les courses sauf les 100m Backstroke (dos) et 100m Breaststroke (brasse). Les meilleurs temps sur le 100m Freestyle (nage libre) sont très proches. De plus, cette dernière course est celle dont le nombre de participations est le plus élevé comme l'indique le graphe du nombre de courses par nageur. La carte des lieux de compétitions indique que la plupart des résultats pour le 100m nage libre ont été réalisés à Anvers. Grâce au dernier visuel, présentant l'évolution des temps avec l'âge, l'identification du nageur le plus rapide sur la course sélectionnée est réalisable facilement.

## XI. Conclusion

Ce travail fut l'occasion d'appliquer un grand nombre d'acquis des cours du Bachelier de Spécialisation. Chaque objectif fixé au début du projet a été abordé, analysé et implémenté. La réflexion et l'analyse des résultats sur base de ces objectifs est encore nécessaire. Pour finaliser un projet, il faut également expliquer comment le projet doit être délivré et utilisé. Tout projet comporte également une réflexion globale sur les pistes d'amélioration possibles et une analyse sur les difficultés rencontrées.

### a. Evaluation

Le premier objectif du projet était de réfléchir et de modéliser une base de données. Cette étape cruciale pour le reste du projet offre un résultat satisfaisant. Toutes les données pertinentes sont reprises au sein d'une structure normalisée. Cette partie n'a pas été possible sans la mise en place d'un flux ETL. Celui-ci a permis d'avoir des données complètes et structurées indispensables pour la suite du projet.

Les dashboards Power BI réalisés respectent également les objectifs initialement prévus. Chaque cible est capable d'identifier les informations clés afin de comprendre et d'analyser les performances. La seule information non reprise sur ces visuels est le résultat des prédictions.

Ces résultats, obtenus à l'aide des modèles de data mining, offrent des résultats satisfaisants. Les prédictions estiment une valeur seulement en moyenne 1,71% supérieure ou inférieure au temps total de la course. Cette précision permet de prédire des résultats à très court terme (<6 mois – 1 an). Étant donné que le modèle repose sur des variables temporelles, il est plus compliqué de réaliser ses prédictions sur le long terme. Pour estimer un résultat longtemps à l'avance, il faudrait estimer des résultats intermédiaires avec pour chacun d'entre eux une marge d'erreur, qui s'accumule et mène à une marge trop importante pour que le résultat soit représentatif.

### b. Pistes d'amélioration

Le projet dans sa globalité est complet et fonctionnel. Il y a cependant toujours des points qui pourraient mériter une attention plus particulière, ou des fonctionnalités supplémentaires qui permettraient d'améliorer le projet final.

Une des améliorations qui pourrait améliorer le projet final est de travailler sur une gestion de la cohérence des informations de lieux de compétition. Ces données ne peuvent être utilisées en l'état pour tirer des conclusions correctes. Pour cela, il serait possible d'appliquer des règles générales ou d'appliquer des techniques de correspondances automatiques à l'aide de librairies Python pour réaliser du « fuzzy matching ».

Au niveau des données, il pourrait également être pertinent de récupérer davantage de résultats. Certaines courses n'ont pas été analysées et n'ont pas de modèle de prédiction adapté. Si un nageur est particulièrement fort sur une seule distance, comme le 1500m nage libre, il est impossible d'estimer des résultats, ni de comparer ces performances avec d'autres sportifs.

Une dernière amélioration, qui est nécessaire pour pouvoir faire évoluer le modèle et avoir des estimations réalistes sur le long terme, est d'ajouter un script capable de récupérer les résultats des nouvelles compétitions, et ensuite de réentraîner les modèles. Comme expliqué plus tôt, ce modèle est très dépendant de données temporelles, et le manque de données récentes est le facteur principal de la diminution de la pertinence des prédictions. Cela pourrait être réalisé avec davantage de temps à l'aide des ressources et des résultats publiés chaque semaine sur swimrankings.net.

### c. Difficultés rencontrées

Un travail aussi large implique également des problèmes et des difficultés. Dès le début, le défi pour récupérer les données est le temps nécessaire. La méthode de collecte des résultats repose sur le Web Scraping. C'est un processus long qui a nécessité une réflexion approfondie pour réduire le nombre de données finalement récupérées. Malgré la réduction du nombre de nageurs total à 6651 nageurs sur les presque 20.000 athlètes avec un résultat officiel, et la décision de ne garder que les 7 courses les plus pertinentes, les 2 serveurs étaient réellement nécessaires pour optimiser le temps de collecte.

La seconde difficulté fut rencontrée lors de la création des visuels sur Power BI. En plus du souci lié aux villes et pays, qui rend l'analyse des données impossible, les outils de BI ne sont pas bien adaptés pour afficher des données de type durée. La solution est d'utiliser deux colonnes distinctes pour la même valeur. L'une au format nombre décimal, qui donne le temps en secondes (valeur récupérée dans la base de données), et l'autre au format texte, qui extrait les minutes de la valeur numérique et modifie la mise en forme du texte au format 'mm:ss.ff'. Cette valeur peut ensuite être affichée comme étiquette ou dans la bulle d'information d'une valeur sur un graphe.

Pour finir, optimiser et entraîner plusieurs modèles nécessite beaucoup de temps. Malgré les connaissances théoriques acquises où cours de la formation, il n'est pas évident d'implémenter cela. Mes compétences en informatique et en développement Python m'ont beaucoup aidé lors de cette phase. Finalement, le résultat obtenu est extrêmement satisfaisant, et avoir pu appliquer ces techniques moi-même est une expérience supplémentaire importante pour le développement futur.

## XII. Bibliographie

1. **Belgium results**, Swimrankings, <https://www.swimrankings.net/index.php?page=rankingDetail&club=BEL>, Consulté dernièrement le 20 mai 2025
2. **Swimmer results**, Swimrankings, <https://www.swimrankings.net/index.php?page=athleteDetail&athleteId=XXXXXX>, Consulté dernièrement le 20 mai 2025
3. **FINA Swimming points**, World Aquatics, <https://www.worldaquatics.com/swimming/points>, Consulté le 23 mai 2025
4. **Réseau de neurones en Python**, eaQbe, [https://docs.eaqbe.com/fr/machine\\_learning/neural\\_network](https://docs.eaqbe.com/fr/machine_learning/neural_network), Consulté dernièrement le 10 avril 2025
5. **Documentation SSIS**, Microsoft, <https://learn.microsoft.com/fr-fr/sql/integration-services/sql-server-integration-services?view=sql-server-ver17>, Consulté dernièrement en février 2025
6. **Stackoverflow**, Stackoverflow, <https://stackoverflow.com/>, Consulté dernièrement en avril 2025

## Usage de l'IA générative

Chat GPT : <https://chatgpt.com/share/6838db78-1cf8-8006-97f4-0219c6c0adf9>